

EAI/Springer Innovations in Communication and Computing

Anandakumar Haldorai  
Arulmurugan Ramu  
Sudha Mohanram  
Chow Chee Onn *Editors*

# EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing

BDCC 2018

 **EAI**  
RESEARCH MEETS INNOVATION

 Springer

# **EAI/Springer Innovations in Communication and Computing**

## **Series editor**

Imrich Chlamtac, European Alliance for Innovation, Gent, Belgium

## **Editor's Note**

The impact of information technologies is creating a new world yet not fully understood. The extent and speed of economic, life style and social changes already perceived in everyday life is hard to estimate without understanding the technological driving forces behind it. This series presents contributed volumes featuring the latest research and development in the various information engineering technologies that play a key role in this process.

The range of topics, focusing primarily on communications and computing engineering include, but are not limited to, wireless networks; mobile communication; design and learning; gaming; interaction; e-health and pervasive healthcare; energy management; smart grids; internet of things; cognitive radio networks; computation; cloud computing; ubiquitous connectivity, and in mode general smart living, smart cities, Internet of Things and more. The series publishes a combination of expanded papers selected from hosted and sponsored European Alliance for Innovation (EAI) conferences that present cutting edge, global research as well as provide new perspectives on traditional related engineering fields. This content, complemented with open calls for contribution of book titles and individual chapters, together maintain Springer's and EAI's high standards of academic excellence. The audience for the books consists of researchers, industry professionals, advanced level students as well as practitioners in related fields of activity include information and communication specialists, security experts, economists, urban planners, doctors, and in general representatives in all those walks of life affected ad contributing to the information revolution.

## **About EAI**

EAI is a grassroots member organization initiated through cooperation between businesses, public, private and government organizations to address the global challenges of Europe's future competitiveness and link the European Research community with its counterparts around the globe. EAI reaches out to hundreds of thousands of individual subscribers on all continents and collaborates with an institutional member base including Fortune 500 companies, government organizations, and educational institutions, provide a free research and innovation platform.

Through its open free membership model EAI promotes a new research and innovation culture based on collaboration, connectivity and recognition of excellence by community.

More information about this series at <http://www.springer.com/series/15427>

Anandakumar Haldorai • Arulmurugan Ramu  
Sudha Mohanram • Chow Chee Onn  
Editors

# EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing


BDCC 2018


 Springer

 **EAI**  
RESEARCH MEETS INNOVATION



*Editors*

Anandakumar Haldorai   
Department of Computer Science &  
Engineering  
Sri Eshwar College of Engineering  
Coimbatore  
Tamil Nadu, India

Arulmurugan Ramu   
Department of Computer Science &  
Engineering  
Presidency University  
Bengaluru, India

Sudha Mohanram  
Sri Eshwar College of Engineering  
Coimbatore  
Tamil Nadu, India

Chow Chee Onn  
Department of Electrical Engineering,  
Faculty of Engineering  
University of Malaysia  
Kuala Lumpur  
Kuala Lumpur, Malaysia

ISSN 2522-8595                      ISSN 2522-8609 (electronic)  
EAI/Springer Innovations in Communication and Computing  
ISBN 978-3-030-19561-8              ISBN 978-3-030-19562-5 (eBook)  
<https://doi.org/10.1007/978-3-030-19562-5>

© Springer Nature Switzerland AG 2020  
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.  
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.  
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

We are delighted to introduce the proceedings of the first edition of the 2018 European Alliance for Innovation (EAI) International Conference on Big Data Innovation for Sustainable Cognitive Computing (BDCC 2018). This conference has brought researchers, developers, and practitioners around the world who are leveraging and developing Big Data technology for a smarter and more resilient data. The theme of BDCC 2018 was “Big Data Innovation for Sustainable Cognitive Computing.”

The technical program of BDCC 2018 consisted of 53 full papers in oral presentation sessions at the main conference tracks. The conference tracks were Track 1, Big Data in Cognitive Computing, and Track 2, Big Data in Sustainable Computing. Aside from the high-quality technical paper presentations, the technical program also featured two keynote speeches, two invited talks, and two technical workshops. The keynote speakers were Dr. Sri Devi Ravana, Professor and Head from the Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, and Dr. P. Arun Raj Kumar from the Department of Computer Science and Engineering, National Institute of Technology (NIT), Calicut, India.

The invited talks were presented by Dr. Umamaheswari K., Professor and Head from the Department of Information Technology, PSG College of Technology, Tamil Nadu, and Mr. Mohamed Azharuddin M., Business Technology Analyst, Deloitte, Hyderabad, India. The two workshops organized were the Analysis of Big Data and Big Data and Society. The workshop was aimed to gain insights into key challenges, understanding, and design criteria of employing wireless technologies to develop and implement future Big Data-related services and applications.

Coordination with the Steering Chairs, Imrich Chlamtac and Dr. Anandakumar Haldorai, was essential for the success of the conference. We sincerely appreciate their constant support and guidance. It was also a great pleasure to work with such an excellent organizing committee for their hard work in organizing and supporting the conference. In particular, the Technical Program Committee, led by our TPC Chair, Dr. Arulmurugan Ramu, and Publication Chairs, Dr. Chow Chee Onn and Prof. Suriya Murugan, has completed the peer-reviewed process of technical papers

and made a high-quality technical program. We are also grateful to Conference Managers, Ms. Radka Pincakova and Ms. Karolina Marcinova, for their support and all the authors who submitted their papers to the BDCC 2018 conference and workshops.

We strongly believe that BDCC 2018 conference provides a good forum for all researcher, developers, and practitioners to discuss all science and technology aspects that are relevant to smart cities. We also expect that the future BDCC 2018 conference will be as successful and stimulating, as indicated by the contributions presented in this volume.

## Conference Organization

<b>Steering Committee</b>	
Imrich Chlamtac	Bruno Kessler Professor, University of Trento, Italy
Dr. Sudha Mohanram	Sri Eshwar College of Engineering, Coimbatore, India
<b>Organizing Committee</b>	
<i>General Chair</i>	
Dr. Anandakumar Haldorai	Sri Eshwar College of Engineering, Coimbatore, India
<i>TPC Chair</i>	
Dr. Arulmurugan Ramu	Presidency University, Bangalore, India
<i>Sponsorship and Exhibit Chair</i>	
Dr. V.S. Akshaya	Sri Eshwar College of Engineering, Coimbatore, India
<i>Local Chair</i>	
Prof. K. Karthikeyan	SNS College of Engineering, Coimbatore, India
<i>Workshops Chair</i>	
Prof. K. Sivakumar	National Institute of Technology, Karnataka, India
<i>Publicity &amp; Social Media Chair</i>	
Dr. Chow Chee Onn	University of Malaya, Malaysia
<i>Publications Chair</i>	
Dr. S. Gokuldev	Amrita University, Mysore, India
Prof. Suriya Murugan	Bannari Amman Institute of Technology, Erode, India
<i>Web Chair</i>	
Mr. P. Raja	Tata Consultancy Services, USA
<i>Posters and Demo Track Chair</i>	
Prof. K. Aravindhan	SNS College of Engineering, Coimbatore, India
<b>Technical Program Committee</b>	
Dr. Chan Yun Yang	National Taipei University, Taiwan
Dr. Shahram Rahimi	Southern Illinois University, Illinois, USA
Dr. Marie Nathalie Jauffret	Director of BECOM Program, Principality of Monaco

(continued)

Dr. Mohan Sellappa Gounder	Al Yamamah University, Saudi Arabia
Dr. Vani Vasudevan Iyer	Al Yamamah University, Saudi Arabia
Prof. Rojesh Dahal	Kathmandu University, Kathmandu, Nepal
Dr. Ram Kaji Budhathoki	Nepal Engineering College, Nepal
Dr. Hooman Samani	National Taipei University, Taiwan
Prof. M.D. Arif Anwary	United International University, Bangladesh
Prof. Fachrul Kurniawan	Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia
Dr. Deepak B. Dhami	Nepal Engineering College, Nepal
Dr. K.R. Baskaran	Kumaraguru College of Technology, Coimbatore, India
Dr. C. Malathy	SRM Institute of Science and Technology, Tamil Nadu, India
Dr. G.K.D. Prasanna Venkatesan	Karpagam University, Coimbatore, India
Dr. M.G. Sumithra, Professor	Bannari Amman Institute of Technology, Tamil Nadu, India
Dr. U. Dinesh Acharya	Manipal Institute of Technology, Manipal, India
Dr. Latha Parameswaran	Amrita Vishwa Vidyapeetham, Coimbatore, India
Dr. Bhabesh Nath	Tezpur University Assam, India
Dr. K.V. Prema, Professor	Manipal Institute of Technology, Manipal, India
Dr. E. Poovammal, Professor	SRM Institute of Science and Technology, Tamil Nadu, India
Dr. Ashalatha Nayak	Manipal Institute of Technology, Manipal, India

Coimbatore, India  
 Bengaluru, India  
 Coimbatore, India  
 Kuala Lumpur, Malaysia

Anandakumar Haldorai  
 Arulmurugan Ramu  
 Sudha Mohanram  
 Chow Chee Onn

# Contents

## Part I Main Track

<b>1</b>	<b>Data Security in the Cloud via Artificial Intelligence with Vector Quantization for Image Compression</b> .....	<b>3</b>
	Srinivasa Kiran Gottapu and Pranav Vallabhaneni	
<b>2</b>	<b>A Hybrid Ant–Fuzzy Approach for Data Clustering in a Distributed Environment</b> .....	<b>9</b>
	K. Sumangala and S. Sathappan	
<b>3</b>	<b>S-Transform-Based Efficient Copy-Move Forgery Detection Technique in Digital Images</b> .....	<b>17</b>
	Rajeev Rajkumar, Sudipta Roy, and Kh. Manglem Singh	
<b>4</b>	<b>Neuro-Fuzzy Ant Bee Colony Based Feature Selection for Cancer Classification</b> .....	<b>31</b>
	S. Gilbert Nancy, K. Saranya, and S. Rajasekar	
<b>5</b>	<b>Entity Resolution for Maintaining Electronic Medical Record Using OYSTER</b> .....	<b>41</b>
	Tanya Gupta and Varad Deshpande	
<b>6</b>	<b>Lifetime Improvement of Wireless Sensor Networks Using Tree-Based Routing Protocol</b> .....	<b>51</b>
	Sushaptha Rajagopal, R. Vani, J. C. Kavitha, and R. Saravanan	
<b>7</b>	<b>An Energy-Efficient Distributed Unequal Clustering Approach for Lifetime Maximization in Wireless Sensor Network</b> .....	<b>63</b>
	S. Manikandan and M. Jeyakarthic	
<b>8</b>	<b>An Effective Big Data and Blockchain (BD-BC) Based Decision Support Model for Sustainable Agriculture System</b> .....	<b>77</b>
	M. Dakshayini and B. V. Balaji Prabhu	

<b>9</b>	<b>An SDN-Based Strategy for Reliable Data Transmission in Mobile Wireless Sensor Networks</b> .....	<b>87</b>
	V. Shubha Rao and M. Dakshayini	
<b>10</b>	<b>Different Aspects of 5G Wireless Network: An Overview</b> .....	<b>97</b>
	Akash R. Kathavate, Bhanu Priya, Rajeshwari Hegde, and Sharath Kumar	
<b>11</b>	<b>Intelligent Systems for Volumetric Feature Recognition from CAD Mesh Models</b> .....	<b>109</b>
	Vaibhav Hase, Yogesh Bhalariao, Saurabh Verma, and G. J. Vikhe	
<b>12</b>	<b>Factors Affecting a Mobile Learning System: A Case Study</b> .....	<b>121</b>
	Sudhindra B. Deshpande, Shrinivas R. Mngalwede, and Padma Dandannavar	
<b>13</b>	<b>Document Similarity Approach Using Grammatical Linkages with Graph Databases</b> .....	<b>131</b>
	V. Priya and K. Umamaheswari	
<b>14</b>	<b>Missing Data Handling by Mean Imputation Method and Statistical Analysis of Classification Algorithm</b> .....	<b>137</b>
	K. Maheswari, P. Packia Amutha Priya, S. Ramkumar, and M. Arun	
<b>15</b>	<b>Task Identification System for Elderly Paralyzed Patients Using Electrooculography and Neural Networks</b> .....	<b>151</b>
	S. Ramkumar, G. Emayavaramban, K. Sathesh Kumar, J. Macklin Abraham Navamani, K. Maheswari, and P. Packia Amutha Priya	
<b>16</b>	<b>A Software-Defined Networking (SDN) Architecture for Smart Trash Can Using IoT</b> .....	<b>163</b>
	T. Vairam, S. Sarathambekai, and D. Vigneshwaran	
<b>17</b>	<b>Modified K-Nearest Neighbor Fuzzy Classifier Using Group Prototypes and Its Application to Skin Segmentation</b> .....	<b>173</b>
	Priyadarshan Dhabe, Mukesh P. Chugwani, and Vaibhav B. Kahalekar	
<b>18</b>	<b>Enhancing Cooperative Spectrum Sensing in Flying Cell Towers for Disaster Management Using Convolutional Neural Networks</b> .....	<b>181</b>
	M. Suriya and M. G. Sumithra	
<b>19</b>	<b>Emoticons and Their Effects on Sentiment Analysis of Twitter Data</b> .....	<b>191</b>
	P. S. Dandannavar, S. R. Mangalwede, and S. B. Deshpande	
<b>20</b>	<b>Prediction of Customer Churn Using Machine Learning</b> .....	<b>203</b>
	Saifil Momin, Tanuj Bohra, and Purva Raut	

**21 Prediction of Crop Yield Using Fuzzy-Neural System**..... 213  
 Bindu Garg and Tanya Sah

**22 Speed Estimation and Detection of Moving Vehicles Based on Probabilistic Principal Component Analysis and New Digital Image Processing Approach** ..... 221  
 T. V. Mini and V. Vijayakumar

**23 A Posture Recognition System for Assisted Self-Learning of Yoga by Cognitive Impaired Older People for the Prevention of Falls** ..... 231  
 K. Ponmozhi and P. Deepalakshmi

**24 Improved UFHLSNN (IUFHLSNN) for Generalized Representation of Knowledge and Its CPU Parallel Implementation Using OpenMP** ..... 239  
 Priyadarshan S. Dhabe and Sanman D. Sabane

**25 Performance Evaluation of Multihop Multibranch DF Relaying Cooperative Wireless Network**..... 249  
 M. Dayanidhy and V. Jawahar Senthil Kumar

**26 Predicting Property Prices: A Universal Model**..... 259  
 E. Poovammal, Mayank Kumar Nagda, and K. Annapoorani

**27 Facial Based Human Age Estimation Using Deep Belief Network** .... 269  
 Anjali A. Shejul, Kishor S. Kinage, and B. Eswara Reddy

**28 Randomized Agent-Based Model for Mobile Customer Retention Behaviour Prediction**..... 279  
 N. Sandhya, Philip Samuel, and Mariamma Chacko

**29 Keyword-Based Approach for Detecting Civil Unrest Events from Social Media** ..... 287  
 J. Joslin Iyda and P. Geetha

**30 Socioeconomic Status Classification of Geographic Regions in Sri Lanka Through Anonymized Call Detail Records**..... 299  
 W. O. K. I. S. Wijesinghe, C. U. Kumarasinghe, J. Mannapperuma, and K. L. D. U. Liyanage

**Part II Workshop on the Analysis of Big Data**

**31 Hand Gesture Based Human-Computer Interaction Using Arduino** ..... 315  
 S. Shreevidya, N. Namratha, V. M. Nisha, and M. Dakshayini

**32 An Automatic Diabetes Risk Assessment System Using IoT Cloud Platform** ..... 323  
 M. Sujaritha, R. Sujatha, R. Anitha Nithya, A. Sunitha Nandhini, and N. Harsha

**33 Message and Image Encryption Embedding Data to  $GF(2^m)$  Elliptic Curve Point for Nodes in Wireless Sensor Networks** ..... 329  
 G. Leelavathi, K. Shaila, and K. R. Venugopal

**34 Crack Detection in Welded Images: A Comprehensive Survey** ..... 339  
 L. Mohanasundari and P. Sivakumar

**35 An Effective Hybridized Classifier Integrated with Homomorphic Encryption to Enhance Big Data Security** ..... 353  
 R. Udendhran and M. Balamurgan

**36 AI Powered Analytics App for Visualizing Accident-Prone Areas** .... 361  
 Preethi Harris, Rajesh Nambiar, Anand Rajasekharan, and Bhavesh Gupta

**Part III Workshop on Big Data and Society**

**37 IOT Based Autonomous Inventory Management for Warehouses** .... 371  
 A. Madhu Vamsi, P. Deepalakshmi, P. Nagaraj, Akash Awasthi, and Anup Raj

**38 Internal Repeats of Human Organs** ..... 377  
 B. Ramya and E. S. Samundeeswari

**39 Bitcoin Prediction and Time Series Analysis** ..... 381  
 Krishna Chakravarty, Manjusha Pandey, and Siddharth Routaray

**40 Smart Active Helmet** ..... 393  
 W. Gracy Theresa and A. Gayathri

**Index** ..... 401



# **Part I**

## **Main Track**

# Chapter 1

## Data Security in the Cloud via Artificial Intelligence with Vector Quantization for Image Compression



Srinivasa Kiran Gottapu and Pranav Vallabhaneni

### 1.1 Introduction

Images play a vital role in today's digital world; they are used as a representation object. They are widely used in gaming, television, satellites, mobile phones and medical field. Images are the latest internet sensations where they are used to showcase about a person in social media websites. When an image is captured, a huge amount of data is also produced which makes it infeasible for storage as well as transmission. A solution for such problem is image compression, where the original data is reduced by fewer bits without compromising on the image quality, by removing the redundant information and restoring the useful and important information.

There are basically two types of compression techniques:

1. Lossless
2. Lossy

Lossless compression technique is a form in which compression takes place without loss of any data or without any quality loss. It is an exact copy of the original image. Such type of compression has applications in the field of medicine where loss of any data can result in an improper and poor diagnosis, in business documents, text documents, source code, etc.

On the contrary, lossy compression technique is a form in which compression takes place with loss of some redundant and unwanted data, where some com-

---

S. K. Gottapu (✉)

Department of Electrical Engineering, University of North Texas, Denton, TX, USA

P. Vallabhaneni

Department of Computer Science and Engineering, Sir C. R. Reddy College of Engineering, Eluru, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_1](https://doi.org/10.1007/978-3-030-19562-5_1)

3

promise on quality is acceptable. Such compression techniques are used where the requirement of compression is high and where some loss of information is acceptable. They are usually used for storage purposes or for transmission through web.

*Objective:* The main objective of this chapter is to introduce an algorithm which combines an artificial intelligence technique with a standard compression technique to achieve desirable compression ratios. The flow of the chapter is as follows: Sect. 1.1 gives an overall introduction about the image compression. Section 1.2 is about the related work done in image compression, a survey on related standard papers and their methodologies and results are discussed. Section 1.3 gives a detailed explanation of the proposed algorithm with flow charts and stepwise explanations. Section 1.4 includes the results and observations obtained through the proposed algorithm. Section 1.5 concludes the report with scope for the future work using this algorithm.

## 1.2 Literature Survey

In [1] a single hidden layer neural network with four neurons in the hidden layer is used for image compression. A vector quantizer with codebook of 256 code vectors is used in the hidden layer for digital transmission of 0.5 bpp. In an input that is a sub-image of size  $4 \times 4$  pixels, 16 pixels is given as input to the network. The output vector from the hidden layer is smaller than the size of the input vector because the input contains 16 neurons whereas the hidden layer consists of only 4 neurons which gives the compressed form of the data. The sub-image is reconstructed at output layer which consists of 16 neurons like the input layer. The analysis of results is carried out by comparing the proposed technique with various other compression techniques that include VQ as residual technique in it. The proposed technique is also compared with 8, 12, 16 hidden neural network. The results show that a good level of PSNR of about 30 dB is obtained with the proposed technique with different number of neurons in hidden layer than the other compression techniques used in comparison.

In [2], the work is on 2-Dimensional Discrete Wavelet Transform (2D-DWT) with Multistage Vector Quantization (MSVQ). The code-book is generated using LBG algorithm for vector quantization (VQ) in different stages. The Radial Basis Function (RBF) neural network is used for training the indices in the MSVQ stages. The method is then applied for different techniques for comparison such as the DCT and (2D-DCT). This method is applied on multiple images of resolution  $128 \times 128$  each. The applied method gives better results in terms of image quality like the PSNR and compression ratio as compared to other transforms. The evaluation of proposed scheme is based on the compression efficiency and distortion measures. The result shows that the output obtained from the above method generates a high-quality compressed image along with better PSNR value and low MSE.

In [3] two levels of VQ are applied. One is applied on the transformed image obtained by hybrid wavelet transform and then it is applied on the error image. At both the levels of VQ generation same size of codebook is obtained. At the first the original image is compressed using transform and an acceptable compression ratio of about 42.6 is obtained, but it produces some distortion. So therefore the VQ is then applied on transformed image for better compression and quality of the image. The combination of these two techniques increases the compression ratio. The obtained distortion in transform technique is eliminated by applying VQ on the error image and then both these compressed images are added which reduces the distortion by 10%.

In [4] compression technique is proposed for gray scale medical images using feed forward neural network along with the back propagation algorithm. The MRI image is applied on the network that consists of three hidden layers. Training is first performed on sufficient sample images to store the node weight and activation values and then it is applied on the targeted image. A compression is achieved since hidden neurons are less in number than the input image pixels. The algorithms are tested for different number of compressor nodes and for different sub-image block size and the performance is evaluated for the compression ratio and PSNR. The algorithm has compression ratio of 1:30 to JPEG2000 with PSNR of 39.56 dB. Therefore the chapter concludes that FNN can achieve good compression performance to the existing techniques for medical images.

### 1.3 Methodology

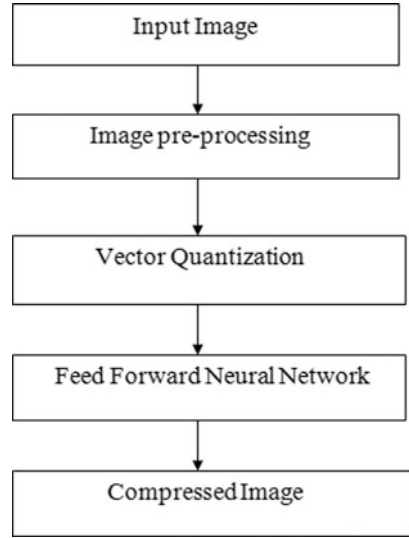
The methodology mainly consists of the five steps included in the flowchart as shown in Fig. 1.1.

*Input Image:* here an image is read from the files which acts as the original image in the process. This image is fed as an input to the next step and on which the compression takes place. The input image can be of any format such as jpg, png, and tiff. The file size of the input image is calculated so as to compare it with the final compressed image [5].

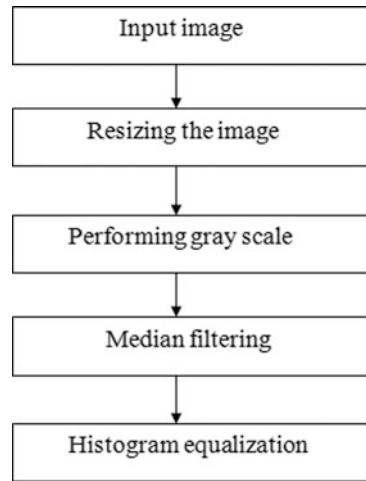
*Image pre-processing:* image pre-processing is performed on the input image; it performs some necessary and application-specific changes in the input image that makes it ready for the next step of compression. The image pre-processing consists of the steps shown in Fig. 1.2.

*Vector Quantization:* A lossy data compression technique is a widely used technology for data storage and transfer. VQ makes use of the rounding off technique or it optimally approximates from an input data to an output data. The compression in VQ is obtained using the 'Codebook', which contains the approximated [6] values using some sort of clustering technique like K-mean clustering. The codebook is used to map the original data or input data to some approximated values which gives the compressed output [7].

**Fig. 1.1** Flowchart for Image compression



**Fig. 1.2** Image pre-processing flowchart



### 1.4 Results

The compression is performed on the standard Lena image as shown in the picture above. The input image is of the resolution of  $512 \times 512$  pixel and is of the size 32,637 bytes. The compression is performed for different values of  $K$  (number of centroids). The range of  $K$  depends upon the input image resolution and the tile size. The tile size chosen here is 8; therefore each tile will contain  $8 \times 8 = 64$  pixels, and the input image resolution is  $512 \times 512 = 262,144$  pixels. No. of tiles =  $262,144/64 = 4096$  tiles. Therefore the value of  $K$  can range from 0 to 4096, for

**Table 1.1** Compression parameters

	Compression ratio	SNR	PSNR
$K = 50$	2.09	13.67	19.32
$K = 100$	1.95	13.76	19.42
$K = 150$	1.85	13.81	19.47
$K = 200$	1.80	13.85	19.50
$K = 250$	1.70	13.9	19.60
$K = 500$	1.70	13.95	19.61
$K = 1000$	1.63	13.9	19.60

**Table 1.2** Compressed file sizes

	Original image size (in kb)	VQ image size (in kb)	Final image size (in kb)
$K = 50$	32.637	16.722	15.598
$K = 100$	32.637	18.243	16.661
$K = 150$	32.637	19.701	17.566
$K = 200$	32.637	20.150	17.852
$K = 250$	32.637	21.103	18.534
$K = 500$	32.637	23.492	19.148
$K = 1000$	32.637	25.415	19.975

$512 \times 512$  input image and tile size 8. But we prefer to study the observations for the ideal values of  $K$  that range between 50 and 250, since greyscale image has colour intensities between 0 and 255. But additionally we can even compare it with higher values than 255. Therefore the values  $K$  chosen are 50, 100, 150, 200, 250, 500 and 1000.

Tables 1.1 and 1.2 give the compression parameters. We observe that as the value of  $K$  increases from 50 to 1000 the compression ratio decreases and the PSNR increases. We observe that a compression of about half the size of the original image is obtained with an average PSNR of 20 dB.

## 1.5 Conclusion

Images are an important part of the digital world today. They are used as representation objects in various fields like medicine, satellites, televisions, and internet. So therefore storing and transmitting of these images needs an efficient solution to reduce their cost of storage and transmission. Hence we make use of the various compression techniques. In this project, a compression algorithm using both Vector Quantization (VQ) and Feed Forward Neural Network (FFNN) is introduced. On the input image (standard image Lena) the VQ is applied first using the K-Mean Clustering with a tile size of 8, and some compression is achieved. The VQ compressed image acts as an input to FFNN and an additional compression is achieved. The results and observations indicate that an acceptable amount of

compression ratio of around 2 which is half of the size of the original image and PSNR of about 20 dB is achieved. It is observed that as the value of  $K$  (number of centroids) increases from 50 to 1000 for the set of observations, it is seen that the compression ratio decreases and PSNR increases.

## References

1. E.M. Saad, A.A. Abdelwahab, M.A. Deyab, Using feed forward multilayer neural network and vector quantization as an image data compression technique, in *Proceedings of the Third IEEE Symposium on Computers and Communications, 1998, ISCC'98*, Athens, 1998, pp. 554–558. <https://doi.org/10.1109/ISCC.1998.702592>
2. V.D. Raut, S. Dholay, Analyzing image compression with efficient transforms & multistage vector quantization using radial basis function neural network, in *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, 2015, pp. 1–6. <https://doi.org/10.1109/ICETECH.2015.7275009>
3. P. Natu, S. Natu, T. Sarode, Hybrid image compression using VQ on error image, in *2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)*, Jaipur, 2017, pp. 173–176. <https://doi.org/10.1109/INTELCCT.2017.8324040>
4. W.K. Yeo et al., Grayscale medical image compression using feedforward neural networks, in *2011 IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, Penang, 2011, pp. 633–638. <https://doi.org/10.1109/ICCAIE.2011.6162211>
5. P.K. Shah, R.P. Pandey, R. Kumar, Vector quantization with codebook and index compression, in *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, Moradabad, 2016, pp. 49–52. <https://doi.org/10.1109/SYSMART.2016.7894488>
6. W. Zhang, H. Li, X. Long, An improved classified vector quantization for medical image, in *2015 IEEE Tenth Conference on Industrial Electronics and Applications (ICIEA)*, Auckland, 2015, pp. 238–241. <https://doi.org/10.1109/ICIEA.2015.7334118>
7. A.J. Hussain, A. Al-Fayadh, N. Radi, Image compression techniques: a survey in lossless and lossy algorithms. *Neurocomputing* **300**, 44–69 (2018)

# Chapter 2

## A Hybrid Ant–Fuzzy Approach for Data Clustering in a Distributed Environment



K. Sumangala and S. Sathappan

### 2.1 Introduction

In the era of big data analytics, more and more databases are becoming available on the Internet and there is a rapidly growing number of online transactions. Data mining and distributed systems are leading the charge of big data analytics.

Data mining is the computer-assisted technique that sequentially seeks out and analyzes enormous sets of data to extract the appropriate information to meet certain requirements. In this web age, numerous online sites are available, which basically fall into two categories: providing services to people and doing smart business through that. The greater parts of the community use those sites based on their requirements. Some sites display the queried results, along with some other information. The burden of this kind of challenge was reduced by the algorithm proposed in this paper.

A distributed database is one in which storage devices are not close to a common processor. Data may be stored in numerous locations, like secondary storage devices, other computers, cloud storage, etc. They are mostly dispersed over a network of organized computers and so on. The proposed approach has combined both mining and distributed environments with a fuzzy-based bio-inspired approach.

Ant-based clustering [1, 2] and categorization use two types of natural ant behavior, i.e., while clustering, ants gather items to form heaps and, while sorting, ants separate different kinds of items and spatially arrange them according to their properties. In prior research work, the enhanced ant clustering algorithm (EACA)

---

K. Sumangala (✉)  
Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India

S. Sathappan  
Erode Arts and Science College, Erode, Tamil Nadu, India



[3], K-means [4], and probabilistic ant-based clustering (PACE) [5] algorithms were used for interzone (global) clustering. Using ant-based clustering always gave better results than the other methods [6]. As per Ref. [7], in previous works, various approaches were used for distributed clustering. In this work, the ant-fuzzy clustering (AFC) algorithm is proposed, with the use of fuzzy clustering for the interzone clustering process. The ant colony building (ACB) [3] and agglomeration [5] approaches are modified by the fuzzy clustering algorithm in AFC, acting to reduce the effects of the issues discussed in Refs. [3, 5] and enrich the performance of the process.

## 2.2 Related Work—Interzonal and Intrazonal Clustering Algorithms

*PACE* [1]: The PACE algorithm [5] is based on the popular particle swarm algorithm of distributed data clusters [5, 8, 9]. Normally, each and every family of ants possess their own unique odor [9, 10]. By using this unique odor, they can be distinguished from other families of ants. This behavior of ants is adopted in this algorithm to identify and form a group of ants carrying related data objects. In a distributed database, the search keywords (data objects) are uniquely treated and identified from the databases of various data sites. The number of occurrences of keywords is computed using the hit ratio and probability values based on the hit ratio are assigned to sites. The high order of probability sites are considered and divided into larger zones. The ants are moved to various locations without any restrictions in their own area and are used to collect the various data objects (food) as freely as they wish. The data object to which they cluster around uniquely identifies the ant group and forms a group (family). Then, each ant family begins to build their colony with the collected data objects inside the zones based on the ant odor identification model (AOIM) [5]. The ants carry the data objects to a specific colony based on the picking and dropping probabilities (ACB) [8] after forming the family and zones. The colony is built similarly to a heap tree. Finally, the heap trees formed by the ants are reordered or sorted to enable agglomeration. In the PACE algorithm, the local and primary cluster was constructed by ACB and agglomeration comprised applied clustering of the interzone.

*EACA* [5]: The EACA includes a special feature to cluster the distributed databases. This algorithm uses the methodology of PACE with a modification, that is, applying the ant clustering algorithm in intrazone and interzone clustering of data items. In EACA, ACB [1, 2] was used to cluster the local data items within the zone and also outside the zone for global clustering. This is an important feature noted in EACA that gave better results than other algorithms in terms of accuracy and error rates.

### 2.3 Proposed Work

The proposed AFC algorithm combines the features of bio-inspired metaheuristics algorithms and reality-based fuzzy algorithms. Fuzzy is intimately connected with the concept of uncertainty. Here, the most fundamental aspect of combining a nature-based algorithm with fuzzy C-means, especially ant clustering with fuzzy C-means for distributed databases, is to find a solution for the uncertainty involved among the cluster zones. Fuzzy C-means allows a single datum to belong to two or more clusters, based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - C_j\|^2 \quad 1 \leq m < \infty \quad (2.1)$$

where  $m$  is any real number greater than 1,  $u_{ij}$  is the degree of membership of  $x_i$  in cluster  $j$ ,  $x_i$  is the  $i$ th piece of  $d$ -dimensional data, and  $C_j$  is the cluster center of  $d$ -dimensional data. Fuzzy partitioning is computed through an iterative optimization of the objective function with the membership updating  $u_{ij}$  and a cluster center  $C_j$  using the following formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left[ \frac{\|x_i - C_k\|}{\|x_i - C_j\|} \right]^{\frac{2}{m-1}}} \quad (2.2)$$

$$C_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (2.3)$$

This iteration will stop when  $(k+1) - u_{ij}(k) < \varepsilon$ , where  $\varepsilon$  is a termination criterion between 0 and 1 and  $k$  is an iteration step.

The algorithm steps of the proposed work for a distributed database are as follows:

1. Inquiring the databases
  - (a) Initialize the counter value for keywords
  - (b) Calculate the thump ratio  $T_r(r_m)$  of the keywords in various database sites using the following formula:

$$T_r(r_m) = \sum_{d_m=1}^{d_m=m} \sum_{i=1}^{i=n} \frac{1}{k_i} \quad (2.4)$$

where  $k$  denotes keywords and  $n$  is the number of keywords

2. Compute the possibility of assorted database sites using the following formula:

$$P_r(d_m) = 1 - T_r(r_m) \quad (2.5)$$

Assign the ceiling of zone formation for the database sites having higher probability.

3. Apply the AOIM to construct primary clusters and implement fuzzy ACB for intrazone clustering

- (a) Apply the fuzzy clustering algorithm for interzone clustering to group the clusters to a single cluster that has a high count of similar keywords
- (b) Clusters are validated using cluster validation measures

After the completion of a finite number of steps, the cluster concludes with the most relevant documents retrieved from the distributed database.

## 2.4 Results and Discussion

This research work presents the experimental results on synthetic and real-world datasets to investigate the properties of the proposed algorithm and compare its effectiveness and scalability with related methods. The experimental setup is carried out with real-time data as well as the benchmark datasets ‘Iris’ and ‘Wine’ from the UCI Machine Learning Repository in order to assess the performance of the proposed algorithm. To evaluate the performance of the algorithm, the datasets are permuted and randomly spread in the sites with a certain number of overlapping data.

The AFC algorithm has the enrichment of the EACA [3, 5], as the attention of ants is directed to more than just the highly apparent data objects. Also, sorting of the heaps [3, 5] of data is done to promulgate the grouping together of highly similar and most probable data. The evaluation methodology was inspired by Refs. [2, 4].

The proposed algorithm is compared with the popular K-means algorithm, the PACE algorithm, and the EACA. Figure 2.1 describes the ant-based clustering of the dataset and the different symbols show the groups of similar data objects.

A confusion matrix is used to evaluate the performance of the algorithm numerically. The  $F$ -measure is computed from the confusion matrix, which is adopted for comparing the clustering results. Table 2.1 shows the results of the  $F$ -measure and error rates of this work, as well as other existing works.

Table 2.1 depicts the results of the  $F$ -measure and error rates for the K-means, PACE, EACA, and AFC algorithms using the Iris and Wine datasets. It is found that the AFC algorithm performs better clustering than the existing K-means and PACE algorithms, while the results of 93% and 90% clustering accuracy for the Iris and Wine datasets, respectively, prove its better performance compared to the other methods.

Figures 2.2 and 2.3 represent the clustering accuracy of the proposed and existing algorithms. It is observed that the proposed AFC algorithm clustered well for the Iris dataset and performs better clustering than the other algorithms for the Wine dataset.

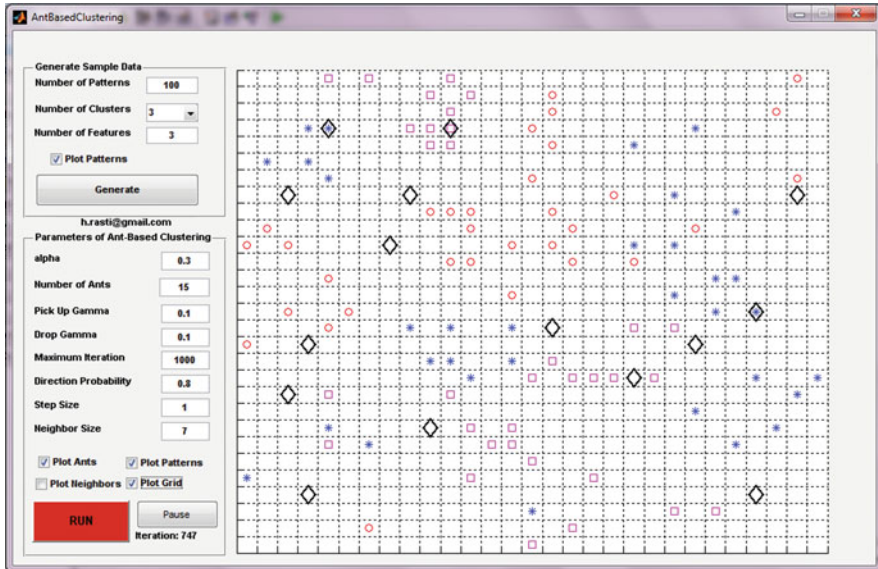


Fig. 2.1 Plotted predicted data

Table 2.1 Comparison of the  $F$ -measure and error rates for the Iris and Wine datasets

Datasets	Algorithms	$F$ -measure value	Minimum errors	Maximum errors	Avg. errors
Iris	K-means	0.8110	0.3	0.8	0.33
	PACE	0.8224	0.2	0.4	0.28
	EACA	0.8334	0	0.32	0.21
	<i>AFC</i>	<i>0.9340</i>	0.12	0.31	<i>0.21</i>
Wine	K-means	0.8217	0.55	0.83	0.57
	PACE	0.8771	0.3	0.45	0.31
	EACA	0.8995	0	0.36	0.29
	<i>AFC</i>	<i>0.9015</i>	0.2	0.33	<i>0.24</i>

On average, it is found that the proposed clustering method yields better results for retrieving relevant data from a large distributed dataset.

Table 2.1 shows that the proposed clustering algorithm is very successful in clustering distributed databases, with a minimum error rate of 2% and the average error rate of the proposed algorithm of 0.21 for the Iris dataset is much smaller compared to the other existing algorithms, namely, K-means, PACE, and EACA. The same is true for the Wine dataset also.

Figures 2.4 and 2.5 show the error rates of the proposed and existing algorithms for the Iris and Wine datasets, respectively. It is clearly demonstrated that the proposed AFC algorithm gives a lower percentage of error than the other existing algorithms.

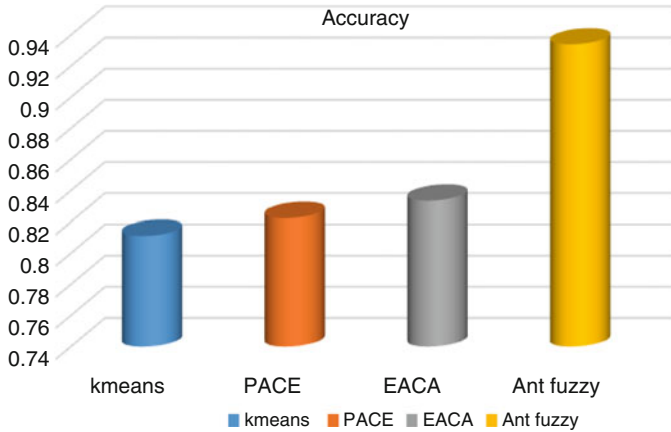


Fig. 2.2 Accuracy of clustering with the Iris dataset

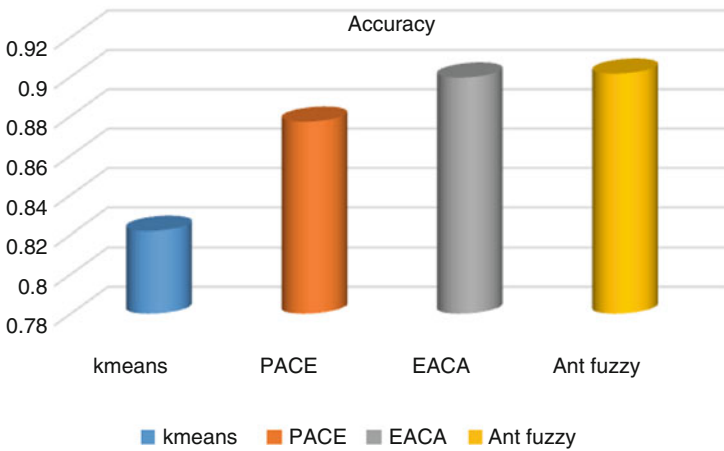


Fig. 2.3 Accuracy of clustering with the Wine dataset

Thus, the above results confirm that the proposed algorithm produces better results than the currently existing algorithms.

## 2.5 Conclusion

This research work presented an ant-based fuzzy clustering of distributed databases. Here, ant clustering is combined with a fuzzy approach in the disseminated databases. The outcomes indicated that the AFC algorithm performs well. The error rate is reduced in each case of the AFC algorithm. The proposed algorithm

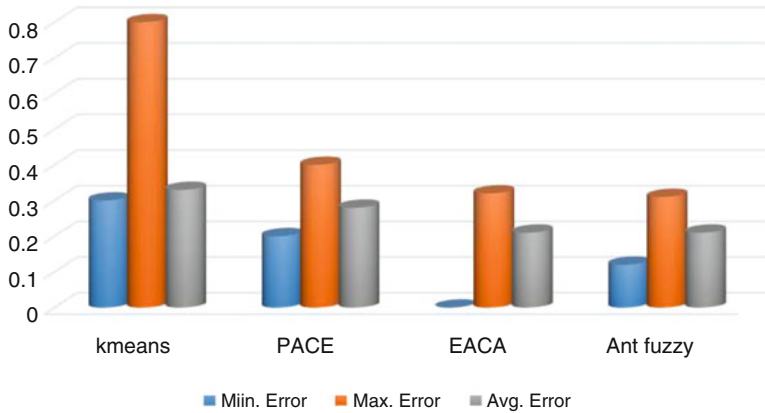


Fig. 2.4 Comparison of the error rates with the Iris dataset

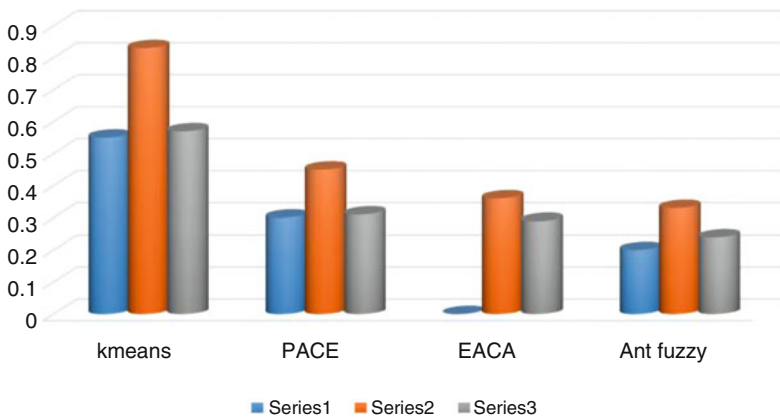


Fig. 2.5 Comparison of the error rates with the Wine dataset

will assist the fresh and budding online sites to improve their performance when searching the data items from distributed sites, as well as local sites. This research addresses a fuzzy-based ant clustering algorithm and gives an overview of web usage mining applications’ attempts to discover useful knowledge from secondary data obtained from the interactions of web users. This method can handle a huge volume of heterogeneous datasets. The performance of this algorithm can also be tested for real commercial problems. In the future, the vector quantization technique may be applied for zone formation. In addition to AFC, further study can include combination and analysis using genetic algorithm machine learning and artificial intelligence (AI) techniques.

## References

1. M. Dorigo, E. Bonabeau, G. Theraulaz, Ant algorithms and stigmergy. *Fut. Gener. Comput. Syst.* **16**(8), 851–871 (2000)
2. J. Handl, J. Knowles, M. Dorigo, Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1D-som. Technical Report TR/IRIDIA/2003-24 (IRIDIA, Universite Libre de Bruxelles, Bruxelles, 2003)
3. K. Sumangala, Enhanced ant clustering algorithm, in *Proceedings of the IEEE Fourth International Conference on Computing, Communication and Network Technologies*, India, Jul 2013
4. E. Bonabeau, M. Dorigo, G. Theraulaz, *Swarm Intelligence—From Natural to Artificial Systems* (Oxford University Press, New York, 1999)
5. R. Chandrasekar, V. Vijayakumar, T. Srinivasan, Probabilistic ant based clustering for distributed databases, in *Proceedings of the IEEE International Conference on Intelligent Systems 2006*, London, UK, Sept 2006
6. N. Dhivya, K. Sumangala, A brief survey on ant based clustering for distributed databases. *Int. J. Comput. Sci. Eng.* **6**(9), 540–544 (2018)
7. D. Singh, A. Gosain, A comparative analysis of distributed clustering algorithms: a survey, IEEE Digital Library, 2013
8. J.L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chretien, The dynamics of collective sorting: robot-like ants and ant-like robots, in *Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, ed. by J. A. Meyer, S. W. Wilson, vol. 1, (MIT Press, Cambridge, MA, 1991), pp. 356–363
9. J. Handl, B. Meyer, Improved ant-based clustering and sorting in a document retrieval interface, in *Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature (PPSN VII)*, LNCS, vol. 2439, (Springer, Berlin, 2002), pp. 913–923
10. M. Dorigo, G. Di Caro, Ant colony optimization: a new metaheuristic, in *New Ideas in Optimization*, ed. by D. Corne, M. Dorigo, F. Glover, (McGraw-Hill, London, 1999), pp. 11–32

# Chapter 3

## S-Transform-Based Efficient Copy-Move Forgery Detection Technique in Digital Images



Rajeev Rajkumar, Sudipta Roy, and Kh. Manglem Singh

### 3.1 Introduction

Digital images are the most influential and broadly utilized standard for communication which has a major influence on our culture and can play more and more significant part in our everyday life [1]. The rise in technological advancement enables us to feature or take away vital capabilities from an image, such that it is problematic to detect hints of tampering [2]. The use of some of the worldly experienced and knowledgeable editing software such as Photoshop, 3D Max, and CorelDraw makes manipulating and altering digital images easy and causes digital forgeries which is shown in Fig. 3.1 [3]. Digital image forgery detection has currently established significant attention because of the increasing number of crime activities and forgeries, especially during the past few years [4] (Fig. 3.1).

### 3.2 Related Work

#### 3.2.1 Block-Based Detection Methods

A nonintrusive blind CMF detection method using undecimated Dyadic wavelet transform (DyWT) was introduced in [5]. In [6] a singular technique used for

---

R. Rajkumar (✉) · S. Roy  
Department of Computer Science and Engineering, Assam University, Silchar, Assam, India

Kh. Manglem Singh  
Department of Computer Science and Engineering, National Institute of Technology, Manipur, Imphal, Manipur, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_3](https://doi.org/10.1007/978-3-030-19562-5_3)

17





**Fig. 3.1** CMF

detection and localization based at the J-Linkage set of rules that can carry out a sturdy clustering in the geometric transformation was introduced. A new detection scheme that uses the Local Binary Pattern (LBP) [7] which was based on filtering and overlapping of the images was proposed to efficiently recognize the forged regions. In advance [8] a local binary pattern variance (LBPV) over the low approximation components of the stationary wavelets is applied over the circular regions. In addition to the filtering techniques [9] uses the rotation and the scaling invariant features, namely Polar Sine Transform (PST) and Polar Harmonic Transform (PHT), to enhance the detection accuracy. Furthermore, seven invariant moments of the maximum circle area in each overlapping block are calculated as moment features to detect the copy-move regions in the image block efficiently [10].

### 3.2.2 *Feature-Based Detection Methods*

Function-based techniques attempt to keep away from obstacles via choosing equal characteristics in photo, in its area of blocks, relying on neighborhood visible functions like SIFT [11]. Simple Linear Iterative Clustering (SLIC) algorithm [12] can also be used to extract SURF key points from the picture block. Histogram of Orientated Gradients (HOG), a blind forensics approach that makes the similarity estimation less complicated for the detection of CMF, becomes added in [13]. Some hybrid feature-based CMF detection introduced in [14] utilizes a robust interest point detector KAZE and combined with SIFT to extract more feature points. Similarly, the method proposed in [15] extracts stationary wavelet transform (SWT)-based features for exposing the forgeries in digital images. A combinational effect of feature extraction, feature matching, and duplicate block identification is used in [16] for detecting copy-move forgery in images under various JPEG compression and Gaussian noise and blurring attacks.

### 3.3 CMF Detection by Stockwell Transform

The proposed CMF detection system [17] is made robust and efficient by exploiting S-transform which utilizes the advantages of both STFT and Wavelet Transform (WT) which is shown in Fig. 3.2. Furthermore, the classifier utilizes the knowledge gathered from (1) the features extracted from the S-transformed image blocks and (2) Euclidean distance (ED) measurement between the overlapped image blocks, to identify the copied regions (Fig. 3.2).

#### 3.3.1 Preliminaries

Let us assume that the dataset  $D$  comprises a collection of  $n$  number of authentic images [18]  $X = \{x_1, x_2, \dots, x_n\}$  and  $m$  number of tampered images  $Y = \{y_1, y_2, \dots, y_m\}$  arranged in an array fashion which can be represented as  $D = \{d_1, d_2, \dots, d_p\}$ , where  $d_i = \{x_i, y_i\}$ . The objective of the proposed forgery detection system is to identify the authentic image  $X$  from  $D$  and to detect the fake regions present in the tampered image  $Y$ . This eliminates the use of initial preprocessing steps in the proposed work. But it is important to divide the forged image  $d_i = \{x_i, y_i\}$  with size  $r \times s$  into  $(r - b + 1) \times (s - b - 1)$  number of image blocks before S-transform which can be expressed as,

$$\{d_i(x, y)\}_{r \times s} = \sum_{i=1}^N \{d'_i(x, y)\}_{(r-b+1) \times (s-b-1)} \tag{3.1}$$

where  $d'_i$  represents the subdivided image blocks of  $d_i$  and  $b$  is the size of the image blocks.

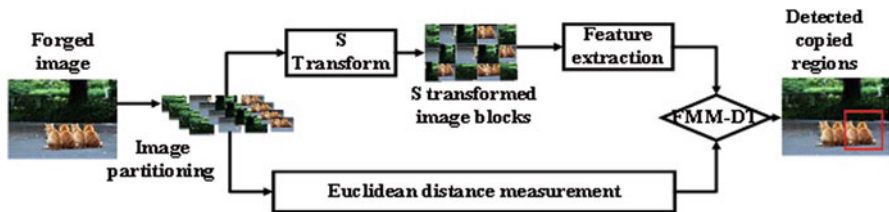


Fig. 3.2 Process of the proposed CMF detection system

### 3.3.2 S Transform

S transform can be regarded as a frequency established STFT or a phase corrected WT. S transform is introduced in this chapter to extract the feature points from all the subdivided image blocks  $d'_i$ . In S transform the amplitude and phase representation of an image block  $d'_i$  can be written as [19]:

$$S[d'_i(\tau, f)] = \mathfrak{I}_l(\tau, f) e^{i\phi_l(\tau, f)} \quad (3.2)$$

where

$$\mathfrak{I}_l(\tau, f) = \text{abs}\{S[d'_i(\tau, f)]\} \quad (3.3)$$

$$\phi_l(\tau, f) = a \tan\{R(S[d'_i(\tau, f)]), I(S[d'_i(\tau, f)])\} \quad (3.4)$$

Here  $\mathfrak{I}_l(\tau, f)$  signifies the amplitude and  $\phi_l(\tau, f)$  represents the phase of image block  $d'$  at time step  $\tau$  for the frequency  $f$  correspondingly. If the intensity level of two image blocks is varied by a constant  $\kappa$ , then Eq. (3.4) can be rewritten as follows:

$$S[d'_i(\tau, f)] = \kappa \mathfrak{I}_l(\tau, f) e^{i\phi_l(\tau, f)} \quad (3.5)$$

From Eq. (3.5) it is known that phase of an image does not vary with the variation in intensity level. Then the features such as mean, standard deviation, and average residual are selectively extracted from the S-transformed image blocks since these features provide more information about the image with reduced effort.

$$\mu = \frac{1}{N} \sum_{i=1}^N \eta(d'_i) \quad (3.6)$$

$$\sigma = \frac{1}{(N-1)^2} \sum_{i=1}^N \sqrt{(\eta(d'_i) - \mu)^2} \quad (3.7)$$

$$\delta_{\text{avg}} = \sum_{i=1}^N |\eta(d'_i) - \mu| \quad (3.8)$$

where  $\mu$ ,  $\sigma$ , and  $\delta_{\text{avg}}$  are the mean, standard deviation, and average residuals, respectively,  $\eta(d'_i)$  is the S-transformed coefficients of image  $d'_i$ . These extracted features are used to create the feature vector for each subdivided image blocks. The feature vector formed for the image block  $d'_i$  is given by:

$$\vec{F}(d'_i) = \{\mu(d'_i), \sigma(d'_i), \delta_{\text{avg}}(d'_i)\} \quad (3.9)$$

By lexicographical sorting, these feature vectors with similar values are grouped collectively. Every function vector is compared with its following vector until a great distinction is determined. Each function vector corresponds to a subdivided image block. In addition to these feature vectors, the knowledge directly gathered by measuring the Euclidean distance between the overlapped image blocks is also used to detect the similarities in the image. The Euclidean distance between two overlapped image blocks is given by:

$$\text{ED}(d'_{ij}) = \frac{1}{MN} \left( \sqrt{\sum_{i=1}^N \sum_{j=1}^M (d'_i - d'_j)^2} \right) \quad (3.10)$$

where  $\text{ED}(d'_{ij})$  is the Euclidean distance between image block  $d'_i$  and  $d'_j$ . These features reveal the nature of duplicated image blocks present in the input image using FMMNN-DT classifier.

### 3.3.3 FMMNN-DT for Forgery Detection

The FMMNN classifier [20] creates lessons through becoming a member of several smaller fuzzy sets into a single set of instructions. Every entered sample is classed primarily based on the degree of club to the corresponding hyperboxes.

A pattern that's contained inside the hyperbox has the club function of 1. The definition of each hyper container fuzzy set  $H_j$  is given by:

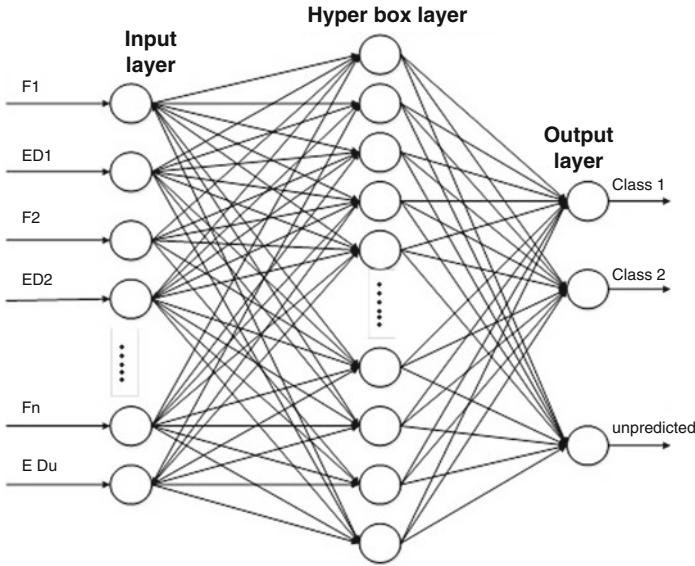
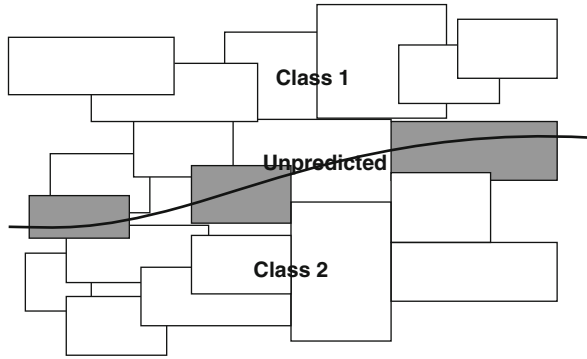
$$H_j = \{X, V_j, W_j, f(X, V_j, W_j)\}, \quad \forall X \in K^n \quad (3.11)$$

where the input pattern is  $X = \{x_1, x_2, \dots, x_n\}$ , the minimum and maximum points of  $H_j$  are  $V_j = \{v_{j1}, v_{j2}, \dots, v_{jn}\}$  and  $W_j = \{w_{j1}, w_{j2}, \dots, w_{jn}\}$ , respectively. Applying the definition of the hyperboxed fuzzy set, the combined fuzzy set that classifies the  $k$ th pattern class  $C_k$  is defined as:

$$C_k = \bigcup_{j \in k} H_j \quad (3.12)$$

The learning algorithm of FMM allows overlapping of hyperboxes of the same class while eliminating overlapping among different classes. The membership function for the  $j$ th hyperbox  $h_j(A_h)$ ,  $0 \leq h_j(A_j) \leq 1$  measures the degree to which the  $h$ th input pattern  $A_h$  falls outside hyperbox  $H_j$  (Fig. 3.3). As  $h_j(A_h)$  approaches 1, the pattern is said to be more contained by the hyperbox. Hyperbox creation is shown in Fig. 3.3. The resulting membership function is given by:

**Fig. 3.3** Hyperbox creation in FMM



**Fig. 3.4** Structure of neural network

$$\begin{aligned}
 h_j(A_h) = \frac{1}{2n} \sum_{i=1}^n & \left[ \max(0, 1 - \max(0, \gamma \min(1, a_{hi} - w_{ji}))) \right. \\
 & \left. + \max(0, 1 - \max(0, \gamma \min(1, v_{ji} - a_{hi}))) \right] \tag{3.13}
 \end{aligned}$$

where  $A_h = \{a_{h1}, a_{h2}, \dots, a_{hn}\} \in K^n$  is the  $h$ th input pattern,  $V_j = \{v_{j1}, v_{j2}, \dots, v_{jn}\}$  is the minimum point for  $H_j$ ,  $W_j = \{w_{j1}, w_{j2}, \dots, w_{jn}\}$  is the maximum point for  $H_j$ , and  $\gamma$  is the sensitivity parameter that regulates how fast the membership values decrease as the distance between  $A_h$  and  $H_j$  increases. Figure 3.4 demonstrates the overall structure of the neural network.

FMMNN-DT is trained offline [21] to measure the similarity between any two image blocks based on the extracted feature vectors and Euclidean distance measurement. The decision will be taken by FMMNN if there is similarity between two image blocks  $d'_i$  and  $d'_j$  as follows:

$$\text{Decision} = \begin{cases} 1, & \text{if } \vec{F}(d'_i) = \vec{F}(d'_j) \text{ and } \text{ED}(d'_{ij}) = \text{ED}(d'_{ij}) \\ \text{unpredictable,} & \text{if } \vec{F}(d'_i) = \vec{F}(d'_j) \text{ or } \text{ED}(d'_{ij}) = \text{ED}(d'_{jk}) \\ 0, & \text{otherwise} \end{cases} \quad (3.14)$$

From Eq. (3.14) the similarity between image blocks will be identified and the unpredictable image blocks are again classified by exploiting DT with the decision of FMMNN. The DT classifier takes the set of unpredictable data  $\text{OD}(d'_i) = \{\text{od}_{x1}, \text{od}_{x2}, \dots, \text{od}_{xn}\}$  resulted because of hyperbox overlapping as the input. Then the distance between the overlapped data and the maximum upper boundary and maximum lower boundary are calculated. This is mathematically expressed as:

$$\text{UD}_i = \max \text{UB} - \text{OD}(d'_i) \quad (3.15)$$

$$\text{LD}_i = \text{OD}(d'_i) - \max \text{LB} \quad (3.16)$$

where  $\text{UD}_i$  is the upper distance of the  $i$ th overlapped data which is the difference between the maximum upper boundary of the created hyperbox and the overlapped data,  $\text{LD}_i$  is the lower distance of the  $i$ th overlapped data which is the difference between the overlapped data and the maximum lower boundary of the created hyperbox.

Now the unpredictable data are classified by using Eq. (3.17) as follows:

$$\text{OD}(d'_i) = \begin{cases} 1, & \text{if } \text{UD}_i > \text{LD}_i \\ 0, & \text{otherwise} \end{cases} \quad (3.17)$$

For ease of understanding the overall process flow of the proposed work is explained in Algorithm 3.1.

---

**Algorithm 3.1: Forgery detection by FMMNN-DT**


---

**Input:** Forged Image  $d_i$  with size  $rxs$

**Output:** Image with detected forgery regions.

```

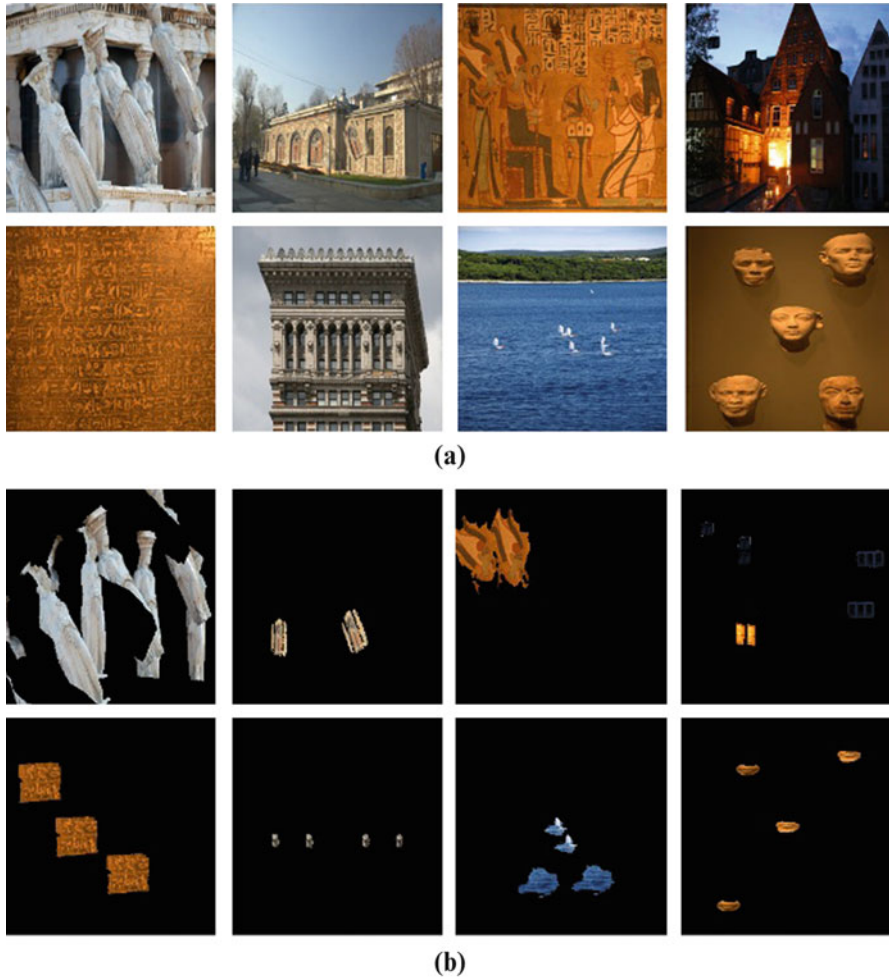
1: Divide  $d_i$  into  $(r - b + 1)(s - b - 1)$  number of image blocks
2: for  $i := 1$  to  $N$  do
3:   Calculate  $ED(d_i)$  for each blocks of  $d_i$ 
4:   Convert  $d_i \rightarrow S$  - transform  $[d_i]$ 
5:    $\bar{F}(d_i) \leftarrow$  Extract  $\mu(d'_i)$ ,  $\sigma(d'_i)$  and  $\delta_{avg}(d'_i)$  from  $S$  - transformed  $d'_i$ 
6:   Input  $\bar{F}(d_i)$  &  $EDd_i$  to FMMNN
7:   while True do
8:     Create hyper box for each input feature vector  $\bar{F}(d_i)$  and  $EDd_i$ 
9:     if  $\bar{F}(d_i) = \bar{F}(d_j)$  &  $ED(d_{ij}) = ED(d_{jk})$  then
10:      Output  $\rightarrow$  class1
11:    else
12:      if  $ED(d_{ij}) \neq ED(d_{jk})$  then
13:        Output  $\rightarrow$  class2
14:      else
15:        Output  $\rightarrow$  unpredictable  $OD(d'_i)$ 
16:      end if
17:    end if
18:  end while
19:  Input unpredictable  $OD(d'_i)$ 
20:  for  $i := 1$  to  $M$  do
21:    calculate  $UD_i$  &  $LD_i$ 
22:    if  $UD_i > LD_i$  then
23:      Output  $\rightarrow$  Class1
24:    else
25:      Output  $\rightarrow$  Class2
26:    end if
27:  end for
28:  Output class 1 as identified forged regions
29: end for

```

---

### 3.4 Simulation Results and Performance Analysis

The proposed forgery detection method is implemented in MATLAB and validated using the dataset named as MIFCC\_600. Figure 3.5a shows some of the eight sample images taken from the MIFCC\_600 dataset and the corresponding results obtained by the proposed forgery detection system is shown in Fig. 3.5b.



**Fig. 3.5** Results of proposed CMF detection system. **(a)** Input images taken from MIFCC\_dataset. **(b)** Output images with identified copied region

Figure 3.6 shows the structure of FMMNN created in this work and the hyperbox created as membership function based on the input image.

### 3.4.1 Performance Analysis and Comparison

In this section we evaluate the performance of our proposed forgery detection framework in terms of precision, recall, false positive rate, and detection accuracy.



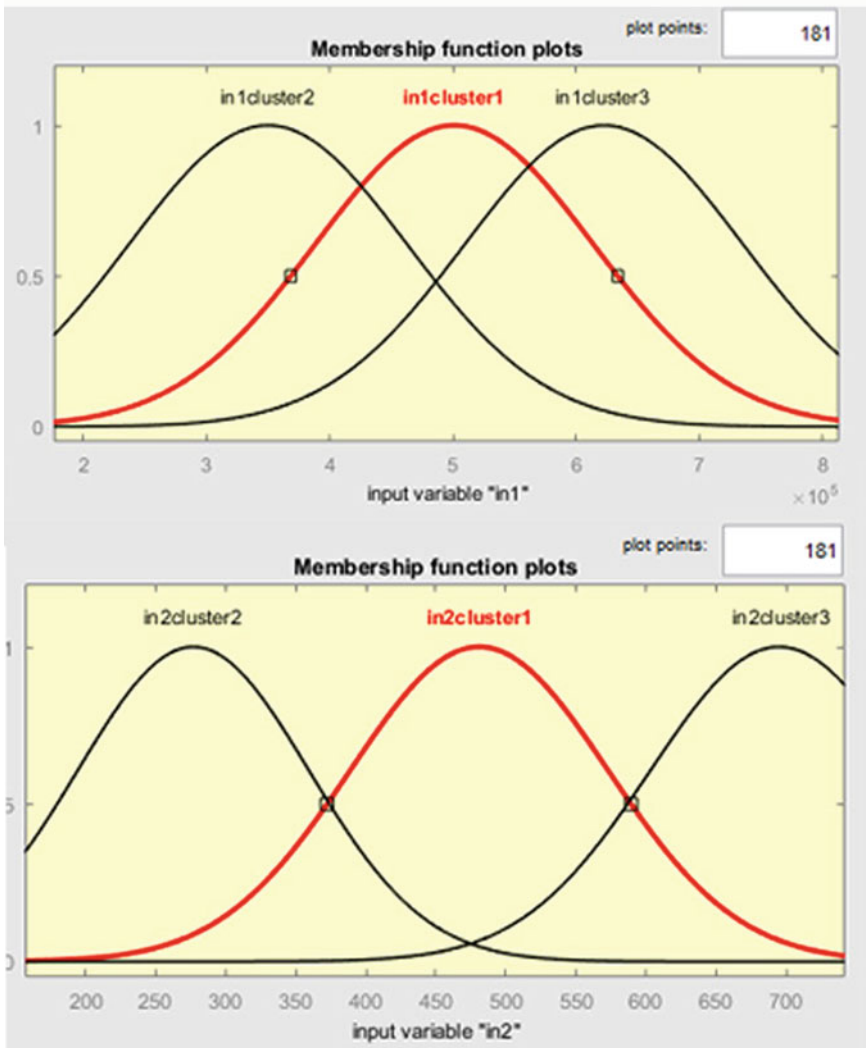
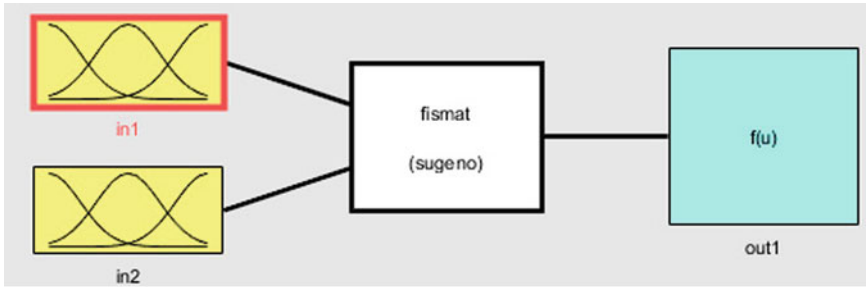


Fig. 3.6 Structure of FMMNN and its membership function

**Table 3.1** Performance of proposed CMF system analyzed with five individual images

	Recall	Precision	Acc	FPR
Img 1	0.9682	1	0.9835	0
Img 2	0.9754	1	0.9874	0
Img 3	0.9805	1	0.9900	0
Img 4	0.9804	1	0.9900	0
Img 5	1	0.974	0.9873	0.02

**Table 3.2** Average contingency values obtained for MIFCC\_600 dataset

	TP	TN	FP	FN
FMM-DT	157	393	7	3
SVM	157	389	11	3
FMM	154	389	11	6
KNN	154	383	17	6
NB	151	380	20	9

**Table 3.3** Performance of proposed CMF system analyzed with MIFCC-600 dataset

	Recall	Precision	Acc	FPR
FMM-DT	0.98125	0.957317	0.98214	0.0175
SVM	0.98125	0.934524	0.975	0.0275
FMM	0.9625	0.933333	0.96964	0.0275
KNN	0.9625	0.900585	0.95892	0.0425
NB	0.94375	0.883041	0.94821	0.05

The performance of our proposed system is compared with some existing classifiers like FMM, SVM, NB, and KNN.

Table 3.1 shows the performance evaluated for the five individual images using their contingency value (shown in Table 3.1) and it is visible that the proposed CMF system results in high values for precision, recall, and accuracy with reduced FPR. Significantly the FPR value for most of the cases (four images out of five) is zero. Table 3.2 shows the average contingency values obtained for MIFCC\_600 dataset in terms of TP, TN, FP, and FN and presents a comparison among classifiers such as SVM, FMM, KNN, and NB with the proposed classifier.

From Table 3.3 it is observed that the accuracy of the proposed system is 98.21% which is significantly greater than SVM classifier which acquires 97.5% accuracy with the same dataset. Similarly, FMM depicts 96.96%, KNN attains 95.89%, and NB acquires 94.82% accuracies, respectively.

From Fig. 3.7 it is visible that there is a linear decrease in precision value with increase in recall. This shows the stability of the proposed classifiers over other classifiers.

## Precision & Recall Comparison

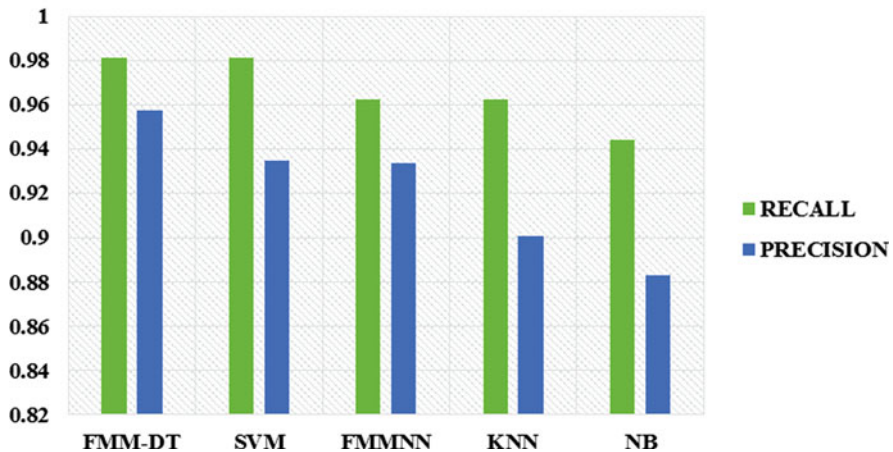


Fig. 3.7 Precision vs. recall for complete dataset

### 3.5 Conclusion

In this chapter, a novel CMF detection system utilizing the advantages of both S transform and a FMM-DT classifier is presented. The proposed CMF system is compared with some existing classifiers such as SVM, FMM, KNN, and NB. The proposed system attains 98.21% detection accuracy which is approximately 0.71% higher than the existing CMF detection systems with SVM classifier.

### References

1. B. Mahdian, S. Saic, Detection of copy-move forgery using a method based on blur moment invariants. *Forensic Sci. Int.* **171**(2), 180–189 (2007)
2. Y. Cao et al., A robust detection algorithm for copy-move forgery in digital images. *Forensic Sci. Int.* **214**(1), 33–43 (2012)
3. A.J. Fridrich, B.D. Soukal, A.J. Lukáš, Detection of copy-move forgery in digital images, in *Proceedings of Digital Forensic Research Workshop*, 2003
4. P. Mukherjee, S. Mitra, A review on copy-move forgery detection techniques based on DCT and DWT. *Int. J. Comput. Sci. Mob. Comput.* **4**(3), 702–708 (2015)
5. G. Muhammad, M. Hussain, G. Bebis, Passive copy move image forgery detection using undecimated dyadic wavelet transform. *Digit. Investig.* **9**(1), 49–57 (2012)
6. I. Amerini et al., Copy-move forgery detection and localization by means of robust clustering with J-Linkage. *Signal Process. Image Commun.* **28**(6), 659–669 (2013)
7. L. Li et al., An efficient scheme for detecting copy-move forged images by local binary patterns. *J. Inf. Hiding Multimed. Signal Process.* **4**(1), 46–56 (2013)

8. T. Mahmood, A. Irtaza, Z. Mehmood, M. Tariq Mahmood, Copy-move forgery detection through stationary wavelets and local binary pattern variance for forensic analysis in digital images. *Forensic Sci. Int.* **279**, 8–21 (2017)
9. S. Li et al., Detecting copy-move forgery under affine transforms for image forensics. *Comput. Electr. Eng.* **40**(6), 1951–1962 (2014)
10. C.-C. Chen, H. Wang, C.-S. Lin, An efficiency enhanced cluster expanding block algorithm for copy-move forgery detection. *Multimed. Tools Appl.* **76**(24), 26503–26522 (2017)
11. H. Gou, A. Swaminathan, M. Wu, Noise features for image tampering detection and steganalysis, in *Proceedings of the IEEE International Conference on Image Processing*, San Antonio, TX, vol. 6, 2007, pp. 97–100
12. J.-C. Lee, C.-P. Chang, W.-K. Chen, Detection of copy-move image forgery using histogram of orientated gradients. *Inf. Sci.* **321**, 250–262 (2015)
13. I. Amerini et al., A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Trans. Inf. Forensics Security* **6**(3), 1099–1110 (2011)
14. Q.-C. Yang, C.-L. Huang, Copy-move forgery detection in digital image, in *Pacific-Rim Conference on Multimedia*, (Springer, Berlin, 2009)
15. T. Mahmood, Z. Mehmood, M. Shah, T. Saba, A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform. *J. Vis. Commun. Image Represent.* **53**, 202–214 (2018)
16. D.-Y. Huang, C.-N. Huang, W.-C. Hu, C.-H. Chou, Robustness of copy-move forgery detection under high JPEG compression artifacts. *Multimed. Tools Appl.* **76**(1), 1509–1530 (2017)
17. J. Li et al., Segmentation-based image copy-move forgery detection scheme. *IEEE Trans. Inf. Forensics Security* **10**(3), 507–518 (2015)
18. R. Davarzani, K. Yaghmaie, S. Mozaffari, M. Tapak, Copy-move forgery detection using multi resolution local binary patterns. *Forensic Sci. Int.* **231**(1), 61–72 (2013)
19. M.V. Chilukuri, P.K. Dash, Multi-resolution S-transform-based fuzzy recognition system for power quality events. *IEEE Trans. Power Deliv.* **19**(1), 323–330 (2004)
20. M. Seera, K. Randhawa, C.P. Lim, Improving the Fuzzy Min–Max neural network performance with an ensemble of clustering trees. *Neurocomputing* **275**, 1744–1751 (2018)
21. A. Quteishat, C.P. Lim, K.S. Tan, A modified fuzzy min–max neural network with a genetic-algorithm-based rule extractor for pattern classification. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **40**(3), 641–650 (2010)

# Chapter 4

## Neuro-Fuzzy Ant Bee Colony Based Feature Selection for Cancer Classification



S. Gilbert Nancy, K. Saranya, and S. Rajasekar

### 4.1 Introduction

Data mining is an outstanding tool to speedily growing field that's involved with developing the techniques to support decision-makers to form intelligent use of those repositories. The main attainment is to get substantive new patterns, correlations, and trends by separation through giant volume of data hold in the repositories using techniques developed in machine learning, pattern recognition, artificial intelligence, arithmetic, and statistics.

Feature selection plays a vital role in data mining for eliminating inappropriate, redundant, and complexity of dimensionality issues in the data warehouse and to provide the wieldy size of data for an effective analyzing process. Feature selection process can be applied for analyzing the critical data, since the datasets always have a far more information than it's needed to construct the model.

### 4.2 Motivation

A swarm of technological advances have resulted in generating a large quantity of microarray data, and have enabled the data to be captured, processed, analyzed, and stored rather inexpensively. The requirement to grasp huge, complex, information-rich data sets is very important to virtually all fields in business, science, and

---

S. G. Nancy (✉) · K. Saranya · S. Rajasekar

Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_4](https://doi.org/10.1007/978-3-030-19562-5_4)

31

engineering. The flexibility to extract helpful information hidden in these knowledge and to act on it information is turning into very important in today's progressively competitive world. Such data (typically terabytes in size) is usually stored in data warehouses and data marts.

### 4.3 Literature Survey

Hong-yan Sang et al. [1] proposed discrete artificial bee colony (DABC) algorithm which solves the minimization problem in  $n$ -job  $m$ -machine flow shop scheduling problem. Initially population is generated based on the quality level to perform search to produce best neighborhood new solutions.

Chabaa et al. [2] presented a model for analyzing a non-Gaussian process using adaptive neuro-fuzzy inference system. This developed model is used for predicting real data, which is also compared with third order moment method for better prediction accuracy.

Suhail M. Odeh [3] presented an automatic diagnosis system for skin cancer using G-flip and ANFIS with the back propagation gradient descent method combined with least square method for improving the classifying accuracy.

Yannis Marinakis et al. [4] presented a new hybrid algorithm for clustering  $N$  objects into  $K$  clusters using Artificial bee colony and greedy randomized adaptive search procedure. This algorithm increases the percentage of the corrected clustered samples.

### 4.4 Proposed System for Feature Selection

The proposed work combined the functionality of ABC and ANFIS. First, the Ant colony optimization (ACO) has to be performed, then the output of the ABC is taken into ANFIS for fast automation of feature selection process. In ACO algorithm, the group of ants and certain ranges are utilized, in which the ants are allowed to find a certain path. Once the path is constructed by the ants, the discrete points are accompanied with that path as candidate points to all the ants in the colony. Then these candidate points are utilized for evaluating the objective function. Then, the bee colony optimization algorithm initialized the location of the food source, which is for optimization process. Then the objective function is assessed by the following stages: employed bee stage, onlooker bee stage, and scout bee stage. Employed bees

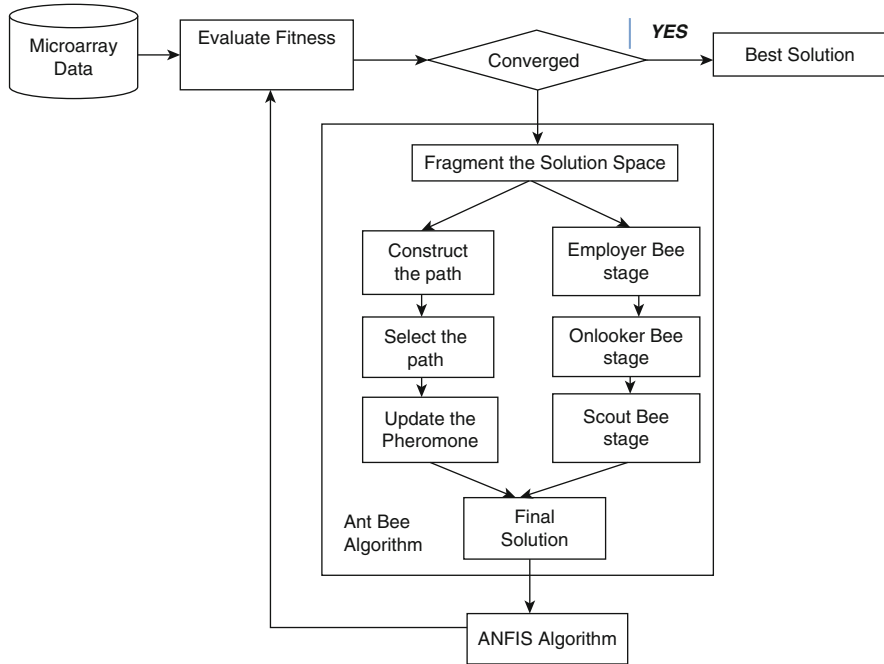


Fig. 4.1 System architecture

are finding the food source and convey to the onlooker bee, the onlooker bees choose the best food source, here the evaluation process is done two times, by which delay convergence will be avoided. Finally the random search is done by the scout bees. After this phase of ABC, ANFIS model is exploited for the automation. The overall performance of the NF-ABC is explained in Fig. 4.1.

## 4.5 Feature Selection Methods

### 4.5.1 Ant Bee Algorithm

Ant Bee algorithm [5, 6] is an optimization algorithm based on the behavior of the ants efforts in search of natural foods. It has two steps: to manage the Bee activity and to calculate the vectors. To calculate the vector, it finds the optimal point either minima or maxima or the fitness function. To manage the bee activity it analyzes the bee movement in a multidirectional in search of different food sources; sometimes it may depend on past experiences, else it finds a new route to find the food sources.

### 4.5.2 ANFIS (Adaptive Network Fuzzy Inference System)

ANFIS model proposed by Jang [7–11], which is a fuzzy inference system with two techniques premise features defines membership function and gradient descent for fine tuning purposes. Supervised learning has the capability of reducing the dimensionality when the input variables are increased to minimize the error measures. This model is broadly used on prediction and nonlinear mapping function. ANFIS encompassed five layers. In the first layer, calculate the degrees of membership for real number values. Every node's output  $e$  is calculated as:

$$T_e^1 = \mu_{A_e}(x), \quad e = 1, 2, \dots, n$$

where  $T_e^1$  signifies the degrees of membership input  $x$ , which is given as the input to membership function  $A_e$ . The piecewise and continuous differentiable function's solutions are among the closed interval  $[0, 1]$ , which is allowable in this first layer as function of membership. Thus, the proposed work used double sigmoid function (dsigf), which calculates from the difference among the two sigmoidal functions.

$$\begin{aligned} \mu(x) &= \text{dsigf}(x, r_1, s_1, r_2, s_2) \\ &= \mu_1(x, r_1, s_1) - \mu_1(x, r_2, s_2) \\ &= \frac{1}{(1 + \exp(-r_1 * (x - s_1)))} - \frac{1}{(1 + \exp(-r_2 * (x - s_2)))} \end{aligned}$$

where  $\{r_1, s_1, r_2, s_2\}$  are premise parameters. In the proposed work, two membership functions are allocated for each input.

In the second layer, each input of nodes has been multiplied, otherwise the nodes gathered the input values from the first layer and returned the product results as firing strength.

$$T_e^2 = w_e = \prod_{e=1}^p \mu_{A_e}(x); \quad e = 1, 2, \dots, n$$

where  $w_e$  is firing strength for  $p$  input in this second layer. In this equation,  $t$ -Norm is used as multiple operators.

In the third layer, normalization played an important role. Normalized weight calculated from the nodes are in the second layer, and is explained here:

$$T_e^3 = \overline{w_e} = \frac{w_e}{\sum_{e=1}^n w_e} \quad e = 1, 2, \dots, n$$



In the fourth layer, every node has function  $g_e$ , which is represented by resultant parameters. Normalized weights are derived from the third layer, which are multiplied by the equivalent parametric functions. For linear function, the output is calculated by the given equation:

$$T_e^4 = \overline{w}_e g_e = \overline{w}_e (a_e x + b_e y + c_e)$$

In the fifth layer, the final output is calculated by the sum of overall outputs of the fourth layer, and is explained below:

$$T_e^5 = \sum_e \overline{w}_e g_e = \frac{\sum w_e g_e}{\sum w_e}$$

Many learning methods are available for training ANFIS, though in this proposed work, hybrid learning rule for training the ANFIS is used due its efficient speed and performance.

## 4.6 Classification Methods

### 4.6.1 *k*-Nearest Neighbor Algorithm

In machine learning, the k-NN [12] is a nonparametric algorithm, which is used for regression and classification. The inputs are taken from k closest features in the sample space. The output of the k-NN is based on the need of the algorithm whether it is for classification or regression. All points correspond to the instances in the multidimensional euclidean space, and is compared with feature vectors to attain the target.

### 4.6.2 SVM Algorithm

Support vector machine [13, 14] is one of the modest and popular approaches in the process of classification. This classifier is used to segregate the data into different classes based on the attributes in the given dataset. In this algorithm, the important level is to find the largest margin hyperplane. Based on the hyperplane, the features in the feature space can be segregated, here the kernel functions are used. During this segregation process, many hyperplanes will be generated, in which the supreme margin hyperplane is needed to find this challenging task which has less generalization errors. The SVM can capture the maximal margin hyperplane and hence it is called the best margin classifier. Classification of data in the feature space using kernel functions is called kernel trick. Different kernel functions are

existing, based on the need, the kernel functions are used. The support vector machine approach uses the small amount of labeled features and provides better accuracy.

### 4.7 Results

The experiments are generated to assess the recital of NF-ABC model by comparing the solutions among the kNN and SVM. The processes of the feature selection and classification have been shown in the subsequent Figs. 4.2, 4.3, 4.4, 4.5, and 4.6 and the performance of the proposed work has been explained in Figs. 4.7 and 4.8. Figures 4.7 and 4.8 show measure of number of features has been selected among NF-ABC and existing feature selection methods and the accuracy of the NF-ABC method. In our work, we find the significant features to attain our goal to predict the cancer type [15, 16] by which can predict the cancer easily and give the proper treatment to the cancer patients.

After selection of qualified feature subsets, they are being applied into data mining algorithms, namely, k-nearest neighbor algorithm and support vector machine for classification, which is shown by the following Figs. 4.5 and 4.6.

The following Figs. 4.7 and 4.8 show accuracy of SVM and kNN classifiers.

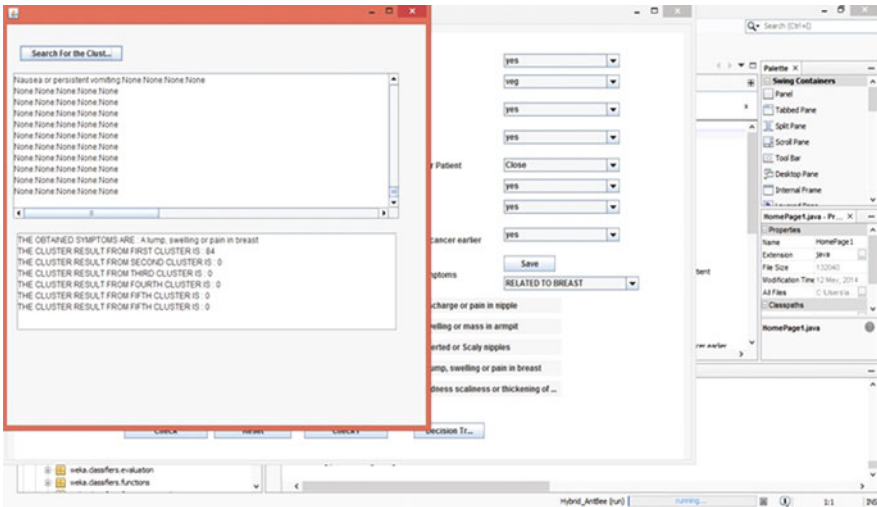


Fig. 4.2 Data collection and clustering



Fig. 4.3 Selection of feature subset

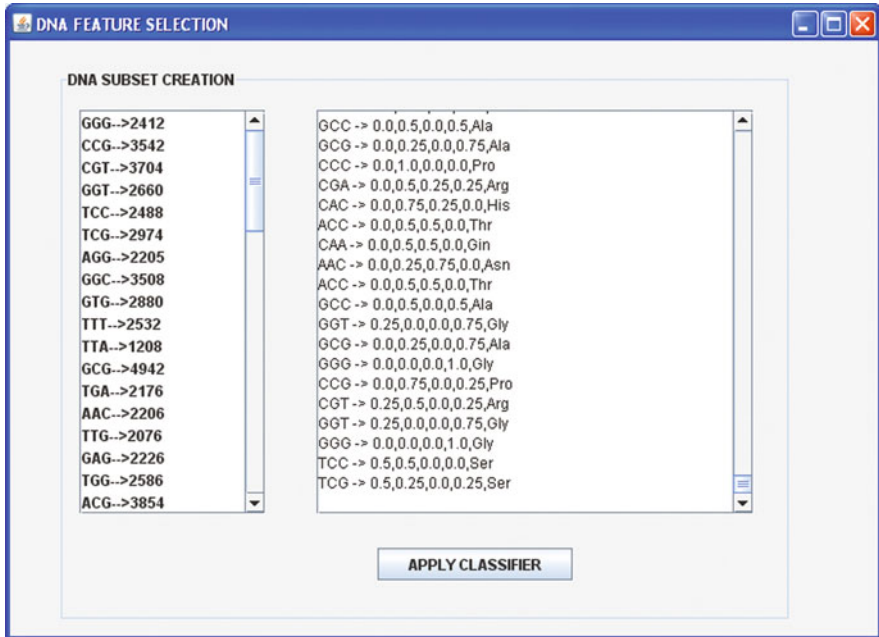


Fig. 4.4 Selected features

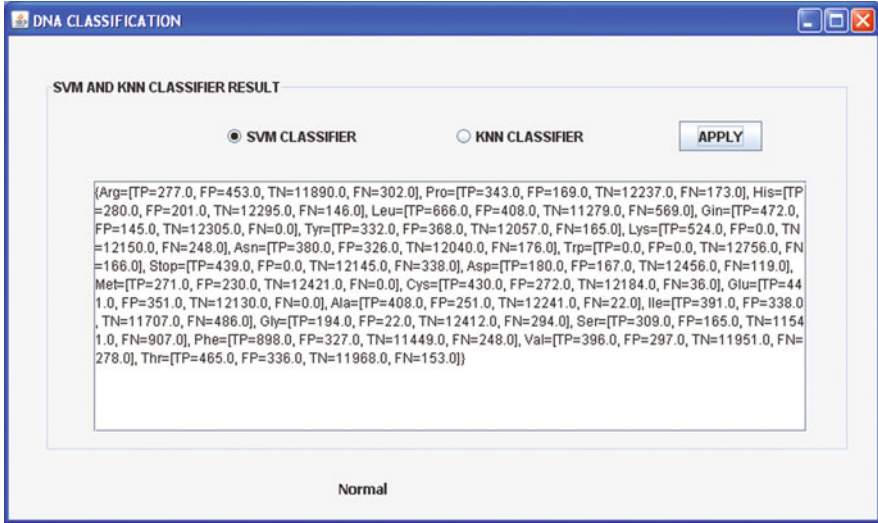


Fig. 4.5 Selected features are fed into SVM Classifier for Classification

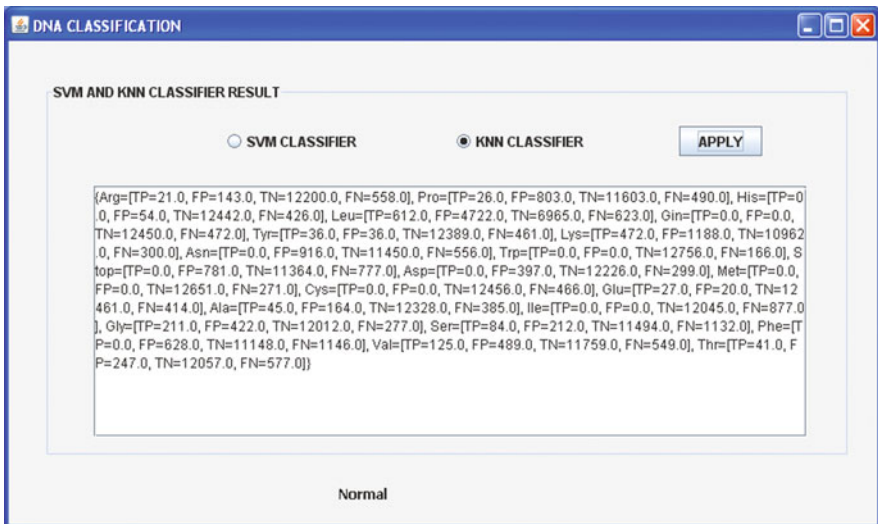
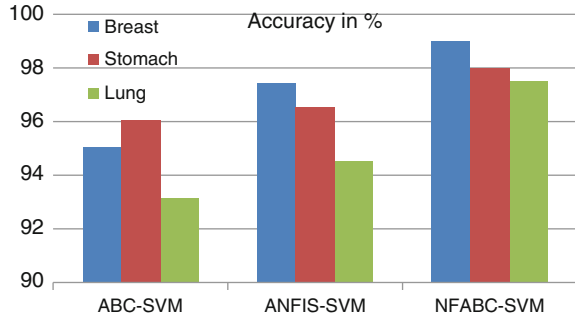
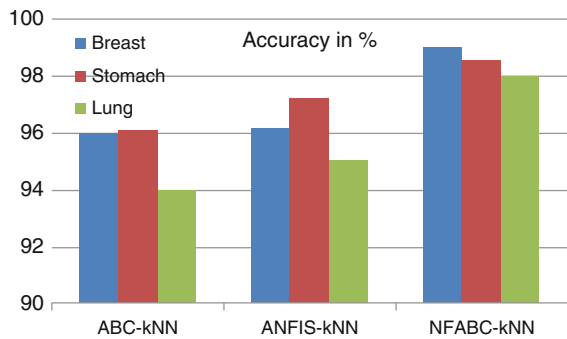


Fig. 4.6 Selected features are fed into kNN Classifier for Classification

**Fig. 4.7** Accuracy comparison for SVM Classifier



**Fig. 4.8** Accuracy comparison for kNN Classifier



## 4.8 Conclusion

In this proposed work the modest endeavor has been made. The issues of the microarray data sets (to eliminate the redundant and irrelevant features and overcome the curse of dimensionality), by using NF-ABC. The proposed work is processed on ABC and ANFIS for overcoming the issues that remove inappropriate and redundant features, as well as reduce the size of the dimension. Finally, the NF-ABC hybrid feature selection method selected the best feature subset with significant features and attained the best classification accuracy, which are evaluated with two classification algorithms, namely, kNN and SVM. The results show the efficiency of NF-ABC method which is yielding the best results rather than individual efficiency of ABC and ANFIS.

## References

1. J. Li, P. Duan, H. Sang, S. Wang, Z. Liu, P. Duan, An efficient optimization algorithm for resource-constrained steelmaking scheduling problems. *IEEE Access* **6**, 33883–33894 (2018)
2. Chabaa S et al., Application of adaptive neuro-fuzzy inference systems for analyzing non-gaussian signal, in *2009 International Conference on Multimedia Computing and Systems* (IEEE Explore)

3. S.M. Odeh, Using an adaptive neuro-fuzzy inference system (AnFis) algorithm for automatic diagnosis of skin cancer. *J. Commun. Computer* **8**, 751–755 (2011)
4. Y. Marinakis, A hybrid ACO-GRASP algorithm for clustering analysis. *Ann. Oper. Res* **188**(1), 343–358 (2011)
5. P. Ganesh Kumar, C. Rani, D. Devaraj, A. Albert Victorie, Hybrid Ant Bee algorithm for fuzzy expert system based sample classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **11**(2), 347–360 (2014)
6. H. Shah, R. Ghazali, N. Mohd Nawi, *Hybrid Ant Bee Colony Algorithm for Volcano Temperature Prediction* (Springer, Berlin, 2012), pp. 453–465
7. J.S.R. Jang, C.T. Sun, E. Mizutani, *Neuro-Fuzzy and Soft Computing* (Prentice-Hall, Upper Saddle River, NJ, 1997)
8. X. Zong, Z. Yong, J. Li-Min, H. Wei-Li, Construct interpretable fuzzy classification system based on fuzzy clustering initialization. *Int. J. Inform. Technol.* **11**(6), 91–107 (2005)
9. P. Woolf, Y. Wang, A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics* **3**, 9–15 (2000)
10. S. Vinterbo, Small, fuzzy and interpretable gene expression based classifiers. *Bioinformatics* **21**(9), 1964–1970 (2005)
11. A.C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinform.* **2**, 75–83 (2003)
12. S. Haddou Bouazza, N. Hamdi, A. Zeroual, Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers, in *2015 Intelligent Systems and Computer Vision (ISCV)*, vol. 1 (IEEE), pp. 1–6. <https://www.computer.org/csdl/proceedings-article/iscv/2015/07106168/12OmNzCwG8q>
13. T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000)
14. F. Chu, L. Wang, Applications of support vector machines to cancer classification with microarray data. *Int. J. Neural Syst.* **15**(06), 475–484 (2005)
15. G. Schaefer, Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recogn.* **42**(6), 1133–1137 (2009)
16. UCI machine learning repository, <http://www.archive.ics.uci.edu/ml/>

# Chapter 5

## Entity Resolution for Maintaining Electronic Medical Record Using OYSTER



Tanya Gupta and Varad Deshpande

### 5.1 Introduction

The process of medical record-keeping has moved from paper based to electronic medium. There has been a significant advancement of technology that is used to keep an update on the patient's medical record. Technology is also dealing efficiently with the privacy of these personal documents. With all these advancements, hospitals are getting better and bigger to handle large number of patients in a single day. Handling the medical data of a myriad of patients, with many patients having similar names, diseases, date of births, etc., can be difficult. Thus, we have introduced a method using Entity Resolution (ER) to manage this data that can otherwise lead to anomalies while processing, reassessing, and updating it into the database. The concept of ER is based on the idea of linking data of identical entities together into clusters and distinguishing them from other separate entities by detecting relationships. Entities in our case are the patients in the hospital and Cluster is a group of all medical entries containing attributes like name, surname, date of birth, visit details, treatment involved, etc., in database that point to a single unique entity. ER can, hence, be viewed as a method of de-duplication which has a huge impact on the efficient organization, storage, and updating of Big Data. As we have entered into the world of Database management, closed source software or paid software can be costly for a seemingly small yet cumbersome and essential task of maintenance of medical records. OYSTER (Open sYSTEM Entity Resolution) is an open-source software which enables a user to perform ER with ease on large amounts of unorganized data. By this all the medical records are clearly visible after performing entity resolution, within their respective clusters for evaluation, analysis,

---

T. Gupta (✉) · V. Deshpande  
Dwarkanadas J. Sanghvi College of Engineering, Mumbai University, Mumbai, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_5](https://doi.org/10.1007/978-3-030-19562-5_5)

41

and future modifications. Big hospitals encounter thousands of patients every day and maintaining a follow-up of every patient's medical record can be a tedious task. Moreover, low cost and easy handling of data will encourage many hospitals who rely on paper to maintain the medical records to go digital. We will use ER to make this cumbersome job easier for the hospitals that handle a great deal of patients.

## **5.2 Literature Review**

### **5.2.1 Medical Record System**

Medical information of patients in the healthcare system is still recorded on paper in many parts of the world. Medical records that are recorded manually on paper can go missing, resulting in the patients taking repeated medical tests, lack of proper diagnosis and security issues related to the medical data [1]. Also, such incomplete medical records can give rise to excessive medication, unwanted drug consumption, and incomplete medication due to incorrect interpretation of the patient's medical problem [2]. A patient's medical record might have to be transferred to some other location and making copies or transportation of such records can be difficult and time-consuming which can also prove to be fatal for the patient's health in case of emergencies. To overcome the disadvantages of the paper-based medical record system, many electronic medical record systems have been devised and designed. One such system involves feeding the necessary data into the patient's record and retrieving the data from a designated database using appropriate hardware and software [1]. Some proposed systems made use of CD-ROM (Compact Disks-Read Only Memory) as repositories of medical data that enabled recording, storage, and real-time retrieval of the required medical data of a patient [3]. However, there were certain shortcomings in these implemented and proposed systems. Information retrieval should also constitute quality assurance of the data retrieved, which is essential for the long-term functioning of a particular electronic medical record system in patient healthcare and was particularly neglected in the reviewed systems [4]. Moreover, high initial costs were witnessed in the reviewed systems due to requirement of additional hardware for meeting the high storage requirements for maintaining medical record of a large number of patients [1, 3, 5].

### **5.2.2 Entity Resolution**

Entity resolution is the process of identifying and matching the manifestation of same real-world entities. There are many situations wherein redundancies and inconsistencies in large amounts of data can burden the storage and processing of this data. ER aims to overcome these problems by ensuring de-duplication of data



by the formation of clusters. At times, the heterogeneity of data sources becomes the major reason leading to unintentional duplication and issues in data integration [6, 7]. This occurs due to differences in text formats, interpretations, and human errors like spelling mistakes. Many methods within ER like the Levenstein edit-distance metric help in getting rid of the effects of such dissimilarities and in obtaining accurate matches, and hence accurate clusters [7, 8]. ER enables the formation of distinct clusters with the help of manually and logically designed match rules which focus on the threshold of similarities and dissimilarities which when applied, bring together and separate two or more records in the given data, respectively [9]. Further, techniques for entity summarization can help in recognizing the underlying entity within a given cluster by ranking the attributes according to their importance to avoid wastage of time in going through lengthy descriptions of an entity [10]. The provision of these beneficial methods to assure information quality makes Entity Resolution a great option for maintaining electronic medical records.

### 5.2.3 OYSTER

OYSTER is an easily configurable open-source software that utilizes several runtime XML scripts that include details about the format and location of the source records to be processed, access to files containing previously identified clusters, match rules and associated matching algorithms along with certain parameters system performance to particular ER applications [11].

Entity Resolution using OYSTER proves to be a savior due to the following reasons:

1. Using Identity Capture, a text file is generated that consists of the different clusters formed from the input data with a unique OYSTER ID generated for each distinct cluster. This file takes up limited space (around 1 MB for 0.2 million records) and solves the purpose of differentiating separate entities.
2. Using appropriate Match Rules, two distinct entities having similar attributes can be separated leading to elimination of ambiguity and procurement of higher accuracy in limited time.

Thus, ER using OYSTER is highly cost-effective and memory efficient and focuses on Information Quality making it a viable process for the implementation of an efficient electronic medical record system.

## 5.3 Methodology

We aim at designing and providing a system to hospitals functioning on a large scale in order to create a reliable, efficient, and compact method of maintaining digital medical records. When a document is prepared with the entries of patients skewed

around with respect to time, it becomes tough for any doctor to go through any one patient's record in an efficient manner. Using Entity Resolution every patient can be identified uniquely without any duplications. Moreover, our system will make data assembling, monitoring, and updating medical records easier and far less tedious for the hospital administration. The following flowchart explains the processes that the system follows.

### 5.3.1 *Flowchart*

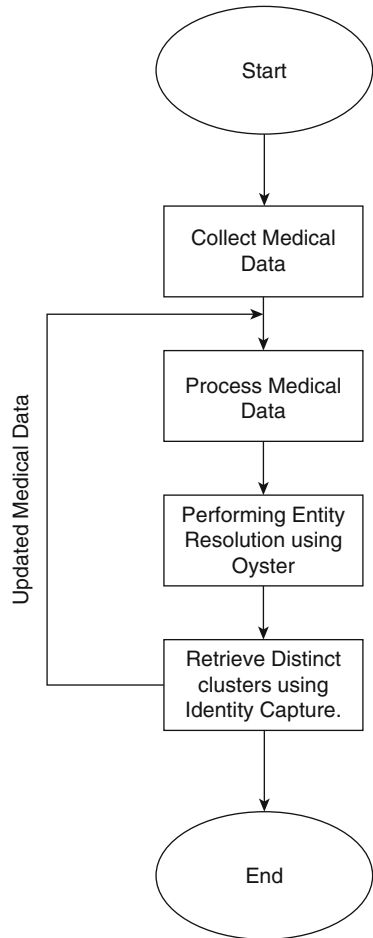
The flow of the process starts with the hospital staff collecting the medical data of patients such as personal details, medical reports, past prescriptions, test results, and further recommendations. This data can be collected by the normal interface which can create normal database in comma separated values. Data processing monitors the assembly of database and its format before entity resolution is performed. As mentioned before, OYSTER utilizes several run-time XML scripts which contains the location and the details of the records. A set of rules called Match Rules are put to use to process all the records and obtain the minimum number of accurate clusters. OYSTER consists of eight different modes or architectures which aid in linking records and form clusters as per specific requirement. Identity Capture, which is one of the eight modes, will be used to maintain medical record of patients electronically and it is the form of Entity Resolution in which the system creates an identity file that acts as a knowledge base which contains all the clusters constructed from the source records during the run [11]. Once the clusters are formed, they constitute the updated, accurate data which can be used for present diagnosis and can be stored for future reference when the patient visits the hospital again so that the past medical record can be analyzed along with the present symptoms and conditions to observe patterns and relate medical history for quicker and effective treatment (Fig. 5.1).

### 5.3.2 *Entity Resolution Using OYSTER*

*Identity Capture:* To understand the concept of Identity Capture, we will consider a sample dataset with arbitrarily arranged records (Fig. 5.2).

The attributes pertaining to this dataset are "RecID," "Name," "Address," "City State Zip," "PO Box," "PO City State Zip," "SSN," and "DOB." A specific number of match rules are applied to obtain clusters. All attributes within each match rule should be identical between two or more records for them to be a part of the same cluster. Two or more records should satisfy at least one match rule to come together in a single cluster. Match rules are applied to the sample dataset to generate clusters that maximize true-positives and true-negatives, avoiding the creation of false-positives and false-negatives [5]. The concept of true-positives indicates that

Fig. 5.1 Flowchart



"RecID","Name","Address","City State Zip","PO Box","POCity State Zip","SSN","DOB"  
 "A998999","Richard Arias","519 Circle Dr","quincy, fl 32351","PO BOX 1709","QUINCY, FL 32353",124-27-1512,""  
 "A998998","tess a stewart","914 S Park Ave","SANFORD, FLA. 32771","pop box 924","SANFORD, FSLA. 32772",102275293,""  
 "A998997","leanen finsiter","4646 E FLORENCE AVE","fresno, ca 93725","",,"070433509,""  
 "A998996","NANCY R ROBERTS","3473 clark rd apt 171","Soarasota, Fla. 34231","PO BOX 1257","Sarasota, Florida 34230",,""  
 "A998995","AMADOR VILLABROZA","3500 vincednt ct","BAKERSFIELD, CA 93304","Caller 60103","Bakersfield, Ca 93386",139846105,""  
 "A998994","aaron zuniga","20414 HAYNES ST","winnetka, ca 91306","",,"403-45-7780,""  
 "A998993","ANASTASIA H SALINAS","2941 Meridian Bay Ln","DICKINSON, TX 77539",,"",170-99-9476,""  
 "A998992","NINA M HA","5322 ASPEN POINT DR","katy, tx 77449","po box 110","KATY, TX 77492",193251838,""  
 "A998991","margaret c mejia","25685 SPRING DR APT 4","Hayward, Ca 94542","",,"086958232,""  
 "A998990","eric murillo","2121 Gus Thomasson Road","DALLAS, TXU 75228","Po Box 810892","dallas, texas 75381",196-40-2570,""

Fig. 5.2 A small part of sample dataset

all the records in a cluster have been rightly brought together to form a cluster and all of them belong to the entity corresponding to that cluster while true-negatives refer to records that have been accurately separated and placed in distinct clusters corresponding to different entities. On the other hand, false-positives refer to the

```

<?xml version="1.0" encoding="UTF-8"?>
- <OysterAttributes System="School">
  <Attribute Item="RecID" Algo="none"/>
  <Attribute Item="Name" Algo="none"/>
  <Attribute Item="Address" Algo="none"/>
  <Attribute Item="Place" Algo="none"/>
  <Attribute Item="POBox" Algo="none"/>
  <Attribute Item="POplace" Algo="none"/>
  <Attribute Item="SSN" Algo="none"/>
  <Attribute Item="DOB" Algo="none"/>
  <!-- -->
- <Indices>
  - <Index Ident="X1">
    <Segment Item="Name" Hash="Soundex"/>
  </Index>
</Indices>
- <IdentityRules>
  - <Rule Ident="1">
    <Term Item="Name" MatchResult="Soundex"/>
    <Term Item="Address" MatchResult="Scan(LR, ALPHA, 0, ToUpper, SameOrder)"/>
    <Term Item="Place" MatchResult="Scan(LR, DIGIT, 0, KeepCase, SameOrder)"/>
  </Rule>
  - <Rule Ident="2">
    <Term Item="Name" MatchResult="Soundex"/>
    <Term Item="SSN" MatchResult="Scan(LR, DIGIT, 0, KeepCase, SameOrder)"/>
  </Rule>
  - <Rule Ident="3">
    <Term Item="Name" MatchResult="Soundex"/>
    <Term Item="DOB" MatchResult="Scan(LR, ALPHA, 0, KeepCase, SameOrder)"/>
  </Rule>
  - <Rule Ident="4">
    <Term Item="Name" MatchResult="Soundex"/>
    <Term Item="Address" MatchResult="Scan(LR, DIGIT, 0, ToUpper, SameOrder)"/>
    <Term Item="POBox" MatchResult="Scan(LR, DIGIT, 0, KeepCase, SameOrder)"/>
  </Rule>
</IdentityRules>
</OysterAttributes>

```

Fig. 5.3 XML code for Match Rules and Attributes

situation wherein two or more records belonging to different entities form a cluster and false-negatives refer to the case when two or more records belonging to the same entity do not occur together in a cluster. The term “positive” signifies the records coming together in a single cluster while “negative” indicates the occurrence of two or more records in separate clusters. To understand this better, let us take an example of three records with the names “Tomm Boy,” “Tomm Boy,” and “Jen Pearl.” For understanding this example, let us assume that the first two records map to just one person named “Tomm Boy” and these are not identical names belonging to two different persons. Now, if the records corresponding to the first two names form a cluster, then it is a case of true-positive. The record with the name “Jen Pearl” forming a separate cluster from the first two records is the case of a true-negative. If at all records with names “Jen Pearl” and “Tomm Boy” are contained within a single cluster, then it is a case of false-positive as two records have falsely come together to form an inaccurate cluster. Also, if the first two records that belong to one entity are contained in separate clusters, then they have been falsely separated, forming a case of a false-negative. Let us view the XML file consisting of match rules and attribute details along with a snapshot of the number of clusters obtained by applying those match rules on the sample dataset (Figs. 5.3, 5.4, and 5.5).

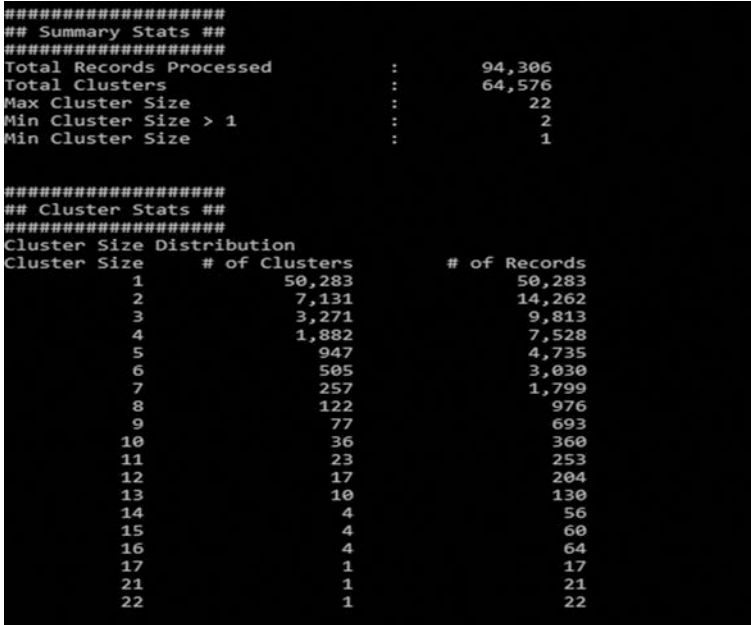


Fig. 5.4 Number of clusters formed

```

Identifier="00FKVM9Z0EZYAN2" (Date="2017-07-06")
:References)
  <Reference>
    <Value>A^source1.A911652|C^RACHELLE A GONZALES|D^6440 WINDSOR LN|E^SAN JOSE, CALI 95129|H^147068420</Value>
    <Traces>
      <Trace OID="00FKVM9Z0EZYAN2" RunID="1" Rule="[2]"/>
    </Traces>
  </Reference>
  <Reference>
    <Value>A^source1.A912775|C^RACHELLE A GONZALES|D^487kay currie rd|E^HUNTINGTON, VTEXAS 75949|H^147068420</Value>
    <Traces>
      <Trace OID="00FKVM9Z0EZYAN2" RunID="1" Rule="[2, 1]"/>
    </Traces>
  </Reference>
  <Reference>
    <Value>A^source1.A913852|C^RACHELLE A GONZALES|D^487 kay currie rd|E^Huntington, Tx 75949|H^147068402</Value>
    <Traces>
      <Trace OID="00FKVM9Z0EZYAN2" RunID="1" Rule="[1]"/>
    </Traces>
  </Reference>
  <Reference>
    <Value>A^source1.A916826|C^RACHELLE A GONZALES|D^351 MIMOSA AVENIDA|E^EL PASO, TX 79915|F^PO BOX 27096|G^EL PASO, TX 709926|H^147068420|I^1923-04-20</Value>
    <Traces>
      <Trace OID="00FKVM9Z0EZYAN2" RunID="1" Rule="[0]"/>
    </Traces>
  </Reference>
  <Reference>
    <Value>A^source1.A986993|C^RACHELLE A GONZALES|D^6440 WINDSOR LN|E^san jose, c 95129|H^147-06-8402|I^04-20-1923</Value>
    <Traces>
      <Trace OID="00FKVM9Z0EZYAN2" RunID="1" Rule="[1]"/>
    </Traces>
  </Reference>
:References)

```

Fig. 5.5 One of the clusters obtained as output

After we've come across all these aspects of OYSTER, we finally will see how Identity Capture comes in the picture. Identity Capture outputs the actual clusters formed along with the corresponding match rule for each record, OYSTER ID (OID) and metadata of each cluster in a text file. A snapshot of this file displaying one cluster formed from the sample dataset is given below.

Thus, Identity Capture helps in the detailed monitoring of each cluster to understand how accurately Entity Resolution has been performed. It also helps us to view the segregation based on distinct entities. This is exactly the point that helps us to successfully associate Entity Resolution with medical health record handling and analysis.

Let us consider a patient "xyz" with an assigned unique ID "abc" who has been visiting a hospital for several years for the treatment of various health issues. The details provided by this patient on each visit are his unique ID, full name, birthdate, contact number, e-mail address, and residential address. The details filled by the doctor include the symptoms, medical condition, status of the treatment for the particular condition, medicines prescribed, and the date and time of visit. If a doctor is willing to procure the medical record of "xyz," then a number of match rules can be applied to obtain all the records corresponding to "xyz." The match rules in this case will be as follows:

1. Match rule with only the unique ID.
2. In case a patient forgets the unique ID, then name and e-mail address can be one match rule.
3. In case the e-mail address of the patient has changed and we still want to identify him/her in the cluster, his name and address can form one match rule.
4. If the address has also changed, we can identify the patient using name, birthdate and contact number as one match rule.

Similarly, many match rules can be formed and the analysis of the clusters visible using Identity Capture can help in deciding whether these match rules are viable giving the correct clusters required as the output. Once, the appropriate match rules are designed as per the attributes specific to the patients in a hospital, the doctors can easily obtain all the necessary information pertaining to the medical history of each patient that visits the hospital.

*Handling Ambiguous Data:* Let us consider a rare yet possible coincidence of two patients that have the same name and birthdate. There is a high chance that the current electronic medical record systems in hospitals confuse these similar yet discreet entities to be a single entity which will not only affect the diagnosis of these patients but in worst case scenarios, can be dangerous and life-threatening. Medical profession is the one that cannot afford any kind of frivolousness. Entity Resolution using OYSTER gives a perfect solution to this problem. Assigning unique IDs to patients can prove to be very useful in overcoming confusion. Customized match rules once designed, considering every possibility, can permanently eradicate ambiguity in data. This will help in the entire focus shifting successfully towards providing proper healthcare services due to quick and efficient diagnosis.

## 5.4 Conclusions

The field of Healthcare can be considered as the one which needs to update itself with time and evolution of technology to contribute in the most effective manner towards the good health of the people. Medical History of a patient can prove to be very helpful in correctly diagnosing and treating any illness. It has been demonstrated how Entity Resolution using OYSTER can revolutionize electronic medical records in hospitals catering to a myriad of patients. The resulting system will work very efficiently, leading to de-duplication of medical data. The concept of Identity Capture accompanied by the designing of good match rules will create clusters of unique entities and merge the attributes pointing to a particular patient. This aims at improving the quality of records and making it easily accessible. By improving system space and time utilization, this method will not only speed up the evaluation process of the doctors to treat the patients but also improve the functioning of the electronic devices being used in the hospitals. A significant yet unintentionally neglected issue of the rare occurrence of ambiguous data, which can be common in case of large numbers of patients in a hospital, can also be solved with ease by the proposed method. With the rapid and advanced modification of medical equipment, devices and tools used for proper treatment and conducting specialized operations, it also becomes essential to focus on introducing new and better paths that help each doctor to strengthen their diagnosis and improve their illness detection skills for any given set of symptoms witnessed by each patient visiting a hospital on a busy day.

## 5.5 Limitations

The major limitation of developing this kind of system is the privacy of patient's medical data [12]. Lawfully, a patient has the right to decide whether he/she is willing to share his/her medical information with a third person other than the doctor. Medical data of a patient is supposed to be a private document, and hence there is a threat of data theft and counterfeit. Thus, data security is a major limitation pertaining to the proposed method. One more limitation is that the match rules are to be applied manually, and hence further automation is needed.

## 5.6 Future Prospect

The proposed model is applicable for large-scale hospitals and medical colleges. This model can be further implemented on the national level where every citizen can be brought on a single portal which can accommodate medical records of every person in a timely manner. This extension will help in integrating data from all

hospitals in the country corresponding to each citizen. In case any patient requires to discontinue visiting his/her regular hospital due to change in residence or change in preference of doctors, this portal will aid them in doing so without any hassle or complications. Moreover, the manual match rules selection can be further automated by the use of machine learning and Artificial Intelligence and make complete system self-reliable and secured.

**Acknowledgement** We are grateful to Dr. John R. Talburt, Professor of Information Science at University of Arkansas, Little Rock, USA, for teaching us the concepts of entity resolution and getting handy with OYSTER. We would also like to thank him for providing us with the sample data for getting the results. Further we would like to thank Prof. Neha Katre and Prof. Vinaya Sawant, Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai for reviewing our work and making it better and more presentable.

## References

1. S.A. Asabe, N.D. Oye, M. Goji, Hospital patient database management. *COMPUSOFT Int. J. Adv. Comput. Technol.* **2**(3), 65–73 (2013)
2. H.S. Lau, C. Florax, A.J. Porsius, A. de Boer, The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. *Br. J. Clin. Pharmacol.* **49**, 597–603 (2001)
3. T.J. Hannan, Electronic medical record. *Canad. Med. Assoc. J.* 1–15 (2008)
4. K. Häyrynen, K. Saranto, P. Nykänen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int. J. Med. Inform.* **77**(5), 291–304 (2008)
5. R.H. Miller, I. Sim, Physicians’ use of electronic medical records: barriers and solutions. *Health Affairs* **23**(2), 116–126 (2004)
6. I. Bhattacharya, L. Getoor, Iterative record linkage for cleaning and integration, in *Proc. SIGMOD-04 DMKD Workshop*, 2004
7. W. Cohen, J. Richman, Learning to match and cluster large high-dimensional data sets for data integration, in *Proc. KDD-02*, 2002, pp. 475–480
8. W. Cohen, P. Ravikumar, S. Fienberg. A comparison of string metrics for matching names and records, in *Proc. KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 2003, pp. 13–18
9. G. Cheng, X. Danyun, Y. Qu, C3D+P: a summarization method for interactive entity resolution. *J. Web. Semant.* **35**(4), 203–213 (2015)
10. G. Cheng, T. Tran, Y. Qu, RELIN: relatedness and informativeness-based centrality for entity summarization, in *Proceedings of the Tenth International Semantic Web Conference, Part I*, ed. by L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, E. Blomqvist, (Springer, Berlin, 2011), pp. 114–129
11. J.R. Talburt, Y. Zhou, A practical guide to entity resolution with OYSTER: handbook of data quality (2013), pp. 235–270
12. J.L. Fernández-Alemán, I.C. Señor, P.Á.O. Lozoya, A. Toval, Security and privacy in electronic health records: a systematic literature review. *J. Biomed. Inform.* **46**(3), 541–562 (2013)



# Chapter 6

## Lifetime Improvement of Wireless Sensor Networks Using Tree-Based Routing Protocol



Sushaptha Rajagopal, R. Vani, J. C. Kavitha, and R. Saravanan

### 6.1 Introduction

Wireless sensor networks (WSN), also called as wireless sensor and actuator networks, are the autonomous sensors used to monitor physical/environmental conditions using sensors and to send their data through the network to the main location. Wireless sensor networks are networks with sensors distributed to sense some physical phenomenon and then the information gathered is processed to get relevant results. Wireless sensor networks consist of protocols and algorithms with self-organizing capabilities. The development of wireless sensor networks is targeted for usage in military applications for battlefield surveillance and are used in many industrial and consumer applications such as machine health monitoring and so on.

The main aim of our paper is to minimize the total energy consumption and increase the load balance by means of a dynamic tree-based routing protocol. Moreover, data compression is provided to improve the performance and the lifetime of the network is greatly improved. To achieve our aim, we not only minimize the total energy consumption but also improve the balance of WSN load. In this paper, we have increased the lifetime of the network by balancing the load as General Self-Organized Tree-Based Energy Balance (GSTEB) routing protocol is

---

S. Rajagopal (✉) · R. Saravanan

Department of Electronics and Communication Engineering, Meenakshi College of Engineering, Chennai, India

R. Vani

Department of Electronics and Communication Engineering, SRM University, Chennai, India

J. C. Kavitha

Department of Information Technology, RMD Engineering College, Chennai, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_6](https://doi.org/10.1007/978-3-030-19562-5_6)

51

used to achieve a longer network lifetime. During each round, the base station assigns a root node and sends the same to all sensor nodes. Each node selects its parent by considering only itself and its neighbor's details, thus making it a dynamic protocol. The key feature of the tree-based data aggregation scheme is to improve the network lifetime. A Lempel Ziv Welch algorithm is used to support data compression in WSNs using sensors and is computationally simple with no transmission overhead. Thus, the protocol used in the proposed system prolongs the overall lifetime of the network. The typical applications include air traffic control, surveillance, environmental monitoring, etc. The paper is organized as follows: Sect. 6.2 summarizes the literature survey related to work. Section 6.3 explains the proposed system and Sect. 6.4 provides results and discussion are done. Finally Sect. 6.5 concludes the work.

## 6.2 Literature Survey

A lot of work has been contributed for prolonging the lifetime of wireless sensor network. The Lempel-Ziv-Welch (LZW) compression algorithm is widely used because it achieves an excellent compromise between compression performance and speed of execution. The algorithm implements lossless compression without degrading the speed and performance. Improvement depends upon the nature of the source and the algorithm achieves a very good compression performance especially on shorter files [1]. Aiming at the problem of sensors limited energy, the algorithm developed in [2] is an improvement based on the classic clustering algorithm LEACH. The load balanced clustering algorithm incorporates two thresholds in DECSA in terms of supporting minimum and maximum acceptable load. Thus, when cluster load in one frame is smaller than the minimum threshold, cluster's frame is expanded, and when it is bigger than the maximum threshold, a fraction of cluster's load in each frame is transferred to a second channel node.

Clustering is used for wireless sensor networks due to less availability of storage (memory), energy, and resources. Clusters are controlled by cluster head which organizes the operation of the entire cluster. The cluster head is responsible for collecting, combining, and transferring the data to the base station. The algorithm makes use of two phases, namely setup phase and steady-state phase. In setup phase, the cluster head will broadcast an advertisement message to the remaining nodes for the selection of head. In steady-state phase, cluster head will perform its operation. TDMA scheduling is implemented. There is reselection of cluster head which prolongs the lifetime of the network. The transmission of packets can be controlled. There is an enhancement in the delivery of data. Fixed clusters are used to decrease the communication overhead. Thus, the protocol minimizes the overhead of cluster formation and latency in data delivery [3]. The algorithm discussed in [4] makes use of two phases, namely setup phase and steady-state phase. TDMA scheduling is implemented and there is reselection of cluster head which prolongs the lifetime of the network. The protocol minimizes the overhead of cluster formation and latency

in data delivery. By applying routing algorithm, protocol was designed to improve the network time and the energy consumption is reduced [5].

Energy saving is used to calculate the performance of WSNs. Decreased energy consumption reduces the energy dissipation, thereby improving the lifetime of the network. Hierarchical routing protocol (DAIC) is better than flat routing protocol (SPIN) as it achieves greater energy efficiency. DAIC splits the network into tiers and selects the channels (cluster head) having higher energy which we use at the nearest distance from the BS (base station). It is an intelligent protocol which dynamically computes the number of cluster heads depending upon the surviving nodes of the network. The energy consumption on the network mainly depends upon the data transmission distance. The channel in the primary tries to collect the data from nodes and channel in the secondary tier transmits the data to the base station. The lifetime is prolonged by means of channel retention. Therefore the DAIC protocol is more energy efficient and can be adapted to a routing protocol for energy-sensitive WSN applications owing to its energy-efficient features [6]. The concept of cluster and graph theory is used to develop energy-efficient algorithm in [7]. This optimization is achieved in cluster head election and clusters routing. This improves the link quality, transmission efficiency, and network lifetime.

An innovative method of using an evolutionary algorithm for the purpose of centralized clustering is used. Evolution means an initial population of solutions is generated which through some generations converge to a population that contains an optimum goal. The chromosomes are selected and initialized and the survivors are evaluated with the existing population. The algorithm utilizes the two fitness functions which is basically a target function containing the maximum or minimum value. Fitness function depends upon the distance because the energy consumption depends upon the distance. Parents are selected and the crossover is implemented which is followed by mutation. After the survivor selection, the process is terminated. Random changing of cluster heads improves the lifetime of the network. Survivor selection retains the suitable chromosomes of the original population and also keeps the chromosomes of the previous generation to improve the efficiency of the algorithm. In the evolutionary algorithm, the fitness function and operators reached global whereas in simulated annealing fitness function and operators will be placed only at local minimal. So evolutionary algorithm used here is for better compared to simulate annealing. The search ability is improved to increase the network lifetime [8]. The nonadaptive measurements have the character of "random" linear combinations of basis/frame elements. The results in [9] use the notions of optimal recovery, of  $n$ -widths and the information-based complexity. They show that "most" subspaces are near optimal and show that convex optimization (Basis Pursuit) is a near optimal way to extract information derived from these near optical subspaces.

Multi-hop communication used in this algorithm increases the energy efficiency of the network. The nodes are deployed and the sink (base station) makes use of broadcasting to produce clusters. A cluster will have a cluster head which utilizes TDMA for its member nodes (sensor). The sensor node of each cluster will send the data to its respective cluster head via intra-cluster communication. The cluster

heads accumulate all the data and finally sends the information via multi-hops to other cluster heads as well as the sink (base station). This is done by inter-cluster communication. If the distance between the cluster head and the base station is small (less than TD-MAX), then the data is transmitted directly, else a relay node is used to transmit the data from cluster head to base station. In this algorithm, interference is minimized by means of a locality TDMA schedule. The stability and scalability are improved by means of a distributed decision which is based on the location information of the network [10].

Wireless sensor networks are utilized in fields where the users require raw data. A centralized clustering geographic energy-aware routing (GEAR-CC) protocol is used to prolong the lifetime of the wireless sensor networks. GEAR-CC has the influence of both hierarchical routing and geographic energy-aware routing. In this protocol the sensor nodes do not care about the routing protocol but simply forwards the data to the next node. GEAR-CC utilizes the global positioning system and directional antennas to provide information about every node to the next base station. The optimum route for transmission is obtained by using the Dijkstra's algorithm. GEAR-CC is a centralized algorithm which works in three phases, namely best route finding phase, next-hop setting phase, and data transmission phase. The energy consumption model provides the optimization of energy in this algorithm. GEAR-CC is the best algorithm because the cluster head will transmit data via the optimal path and optimization is done among all nodes than cluster heads [4]. The research work in [11] focuses on design and development of new Energy-Aware Multicast Cluster (EAMC)-based routing to enhance the QoS of the MANETs. The metrics such as average delay, energy consumption, packet delivery ratio, loss, and throughput are considered in this study to measure the behavior and to enhance the service quality of the MANET.

Wireless sensor networks make use of the ELDC protocol which is nothing but an artificial neural network-based energy-efficient and robust routing scheme for environmental monitoring. The protocol achieves its objectives by means of an artificial neural network (ANN) which is basically arithmetic algorithms capable of learning the complicated mappings between input and output. ELDC helps in the selection of chief node and reduces the selection time required, thereby performing the desired task. The protocol follows a three-layered structure and so it is an extension of Energy-Efficient Unequal Clustering (EEUC) and Energy-Efficient Multiple Distance-Aware Clustering (EEMDC) protocols. ELDC suffers from a large amount of packet loss due to the absence of load balance which reduces the reliability of the network. The random selection of cluster heads reduces the energy level and due to heavy energy burden there is an early death of the cluster head. In ELDC only intraprocess communication takes place successfully and there is a great possibility of errors which might occur in interprocess communication. This is because in interprocess communication any external node consisting of some malicious or unwanted data might enter a group and corrupt the nodes having the data required for transmission and so there is no proper security of data. These factors greatly reduce the lifetime of the network [12].

### 6.3 Proposed System

In order to overcome the limitations of ELDC protocol, we propose the use of GSTEB protocol which is a General Self-Organization Tree-Based Energy-Balancing protocol that implements efficient data operations on a large amount of data in wireless sensor networks (WSN). The network is a system composed of a large number of low-cost micro-sensors which is used to collect and send various kinds of message(data) to the base station(destination). WSN comprises a large number of low-cost nodes with limited battery power and so it is highly difficult to replace the batteries of all the nodes in the network. The only solution to offer a long-life work time is to improve the energy efficiency of the network and so GSTEB can be used to minimize the total energy efficiency and offers load balance which prolongs the lifetime of the network. The performance is further enhanced by means of data compression and this data compression is implemented using S-LZW compression algorithm which is a lossless compression and S stands for Sensors used in the network. The proposed algorithm increases the reliability of the WSNs by increasing its lifetime.

The block diagram of GSTEB protocol is shown in Fig. 6.1. The brief explanation of each block is as follows.

- Source: The source consists of the sensor nodes which act as a client. The data is transmitted from the source.
- Initial phase: In this phase, the basic parameters of the network are initialized. In the beginning, BS broadcasts a packet to all the nodes and each node sends its packet with a particular radius, and finally the neighbors receive the packet and store the information in the memory
- Tree Constructing Phase: This stage is very important as the routing tree is built in this phase. The BS assigns a node as root and broadcasts root ID and root coordinates to all sensor nodes. Each node tries to select a parent from its

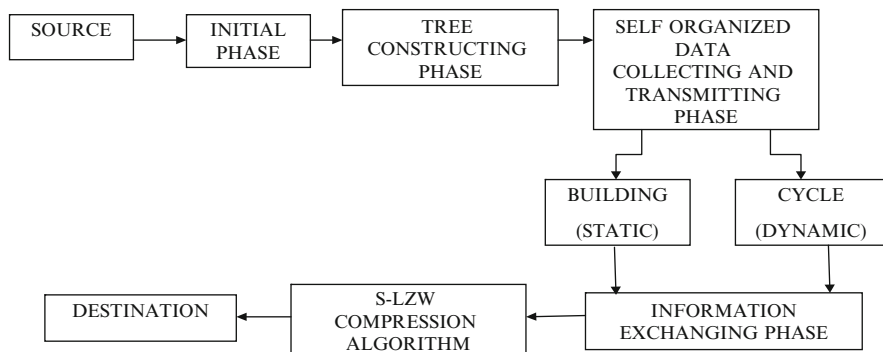


Fig. 6.1 Block diagram of GSTEB protocol

neighbors, and if a node has no child node, then it defines itself as a leaf node from which the data is transmitted.

- **Self-Organized Data Collecting and Transmitting Phase:** After the construction of routing tree, each node collects the information to generate a data packet to be transmitted to the base station. If the information is fixed (static), then additional information cannot be included and it is called as BUILDING. If the information is not fixed (dynamic), then information can be added or removed and it is called a CYCLE.
- **Information exchanging phase:** The dying sensor nodes greatly influence the topography of the network and so the nodes that are going to die must inform the other nodes so that the topography remains unaffected. This phenomenon of information exchange is done in this phase.
- **S-LZW Compression Algorithm:** The data is compressed after the construction of tree and sent to the destination. The algorithm implements lossless compression for the sensor nodes in the network.
- **Destination:** The destination consists of the base station which is the server and receives the compressed data from the source.

### ***6.3.1 Tree-Based Data Aggregation Scheme***

The tree-based data aggregation concept focuses to maximize the network lifetime where the sensor nodes are distributed in tree structure. Data aggregation operation is done at the intermediate nodes along the tree and the final aggregated data is sent to the root node.

The proposed system (GSTEB) maintains a data aggregation tree in a WSN. For tree construction, the tree root first broadcasts a control message which contains information related to the sensor node identification, the parent node, the residual power, the status of the sensor node, and the hop counts from the sink. Based on these data, sensor uses a timer to count down to identify the idle channel. The neighboring node with more residual power and shorter path is found along the tree. The process is repeated till all sensor nodes are added to the tree.

After tree is built using data aggregation concept, a residual power threshold is compared with threshold  $P_{th}$  to maintain the tree structure and its child nodes. The network lifetime is maximized in terms of the number of rounds where each round corresponds to the aggregation of sensing data transmitted from different sensor nodes. To achieve this objective, the proposed GSTEB aims to reduce the total energy consumption of sensor nodes in each round by calculating a minimum spanning tree over the network with link costs.

### 6.3.2 *S-LZW Compression Algorithm*

The compression concept with sensors implies that the data compressed can be recovered from a small number of nonadaptive, randomized linear projection samples. Thus, they can exploit compressibility without relying on any prior knowledge or assumption on sensing data. Lossless compression and data aggregation techniques are combined to select a subset of sensor nodes to collect and fuse the sensing data sent from their neighboring nodes and to transmit the small-sized aggregated data to the sink node of the tree. The LZW algorithm is computationally simple and has no transmission overhead as the same initial dictionary entries are used at both ends of transmitter and the receiver. Based on input and the existing dictionary data, the receiver can recover the complete dictionary from the compressed data.

## 6.4 Results and Discussion

This section deals with the results obtained from our paper, the results observed in terms of graphs after the simulation of the existing and proposed system.

It is clearly evident from the simulation window of the existing system shown in Fig. 6.2 that the existing system (ELDC) suffers from packet loss and there is no possibility of simultaneous transmission and reception of data. Also, the load is not properly balanced.

The simulation window of the proposed system shown in Fig. 6.3 depicts the dynamic tree-based routing used in the proposed system (GSTEB) and there is no packet loss. The simultaneous transmission and reception of data is possible and so GSTEB protocol achieves load balance and prolongs the lifetime of the network. There is a security feature in our proposed system where the protocol prevents the entry of unauthorized node containing malicious data by means certificate revocation. So the sensor nodes are prevented from malicious attack.

Wireless network with communication model is established using TCL script. End-to-end delay is calculated using awk script which processes the trace file and produces the result in a file. Delay is the difference between the time at which the sender generates the packet and the time at which the receiver receives the packet. Delay is calculated using awk script which processes the trace file and produces the result. The end-to-end time delay plot shown in Fig. 6.4 compares the delay of the existing system (ELDC-RED) and proposed system (GSTEB-GREEN) and it is observed that the red curve has a greater delay than the green curve and so GSTEB is more efficient.

A node loses a particular amount of energy for every packet transmitted and also for every packet received. The packet delivery ratio for a particular number of nodes must be high to ensure the energy efficiency of the nodes. The node packet delivery plot shown in Fig. 6.5 gives the packet delivery ratio for the nodes in both

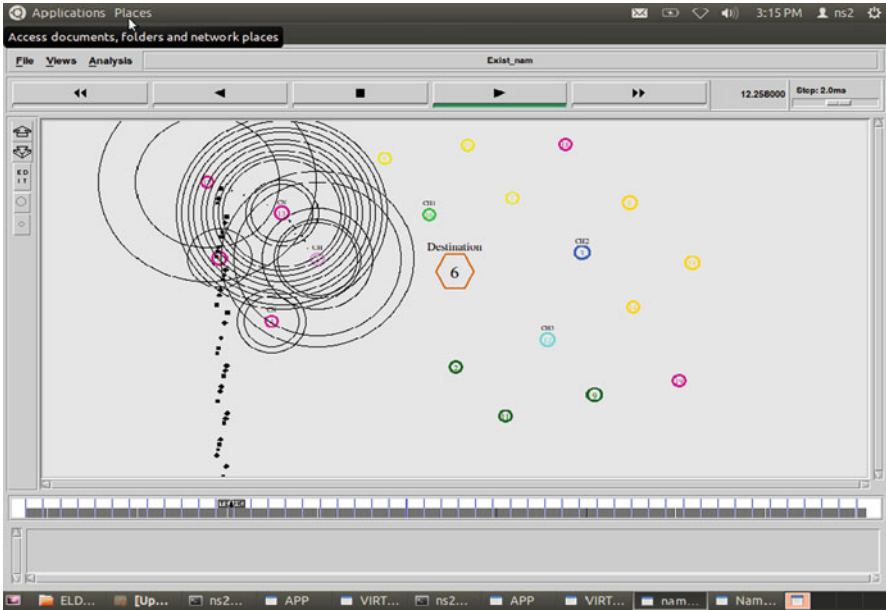


Fig. 6.2 Simulation window of the existing system

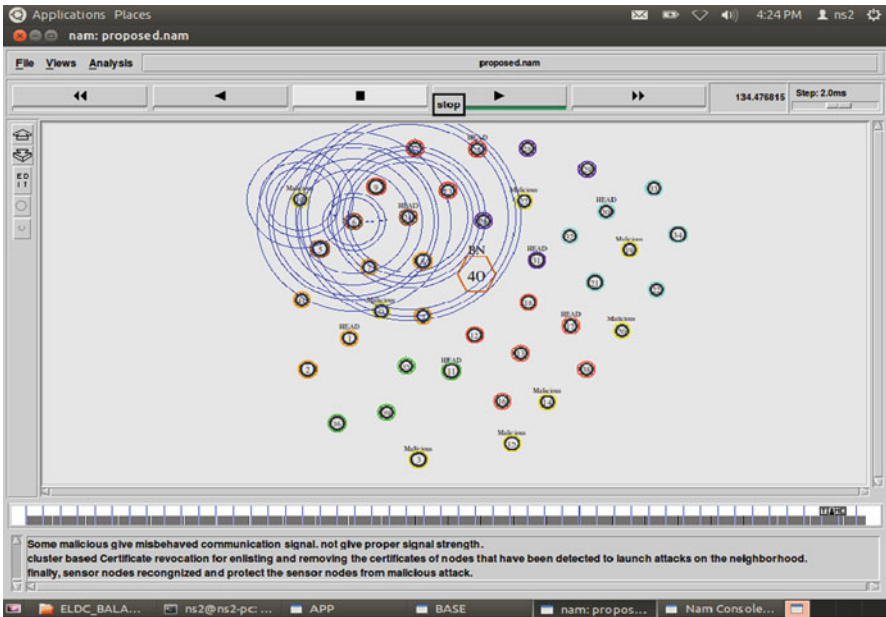


Fig. 6.3 Simulation window of the proposed system



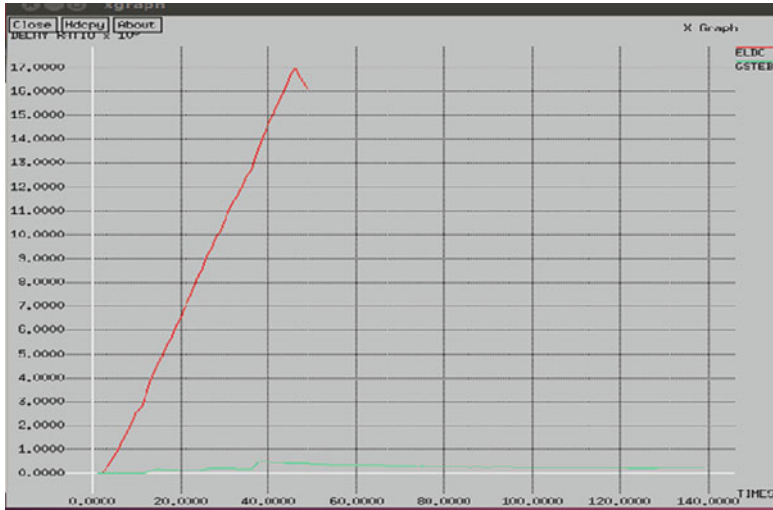


Fig. 6.4 Delay plot

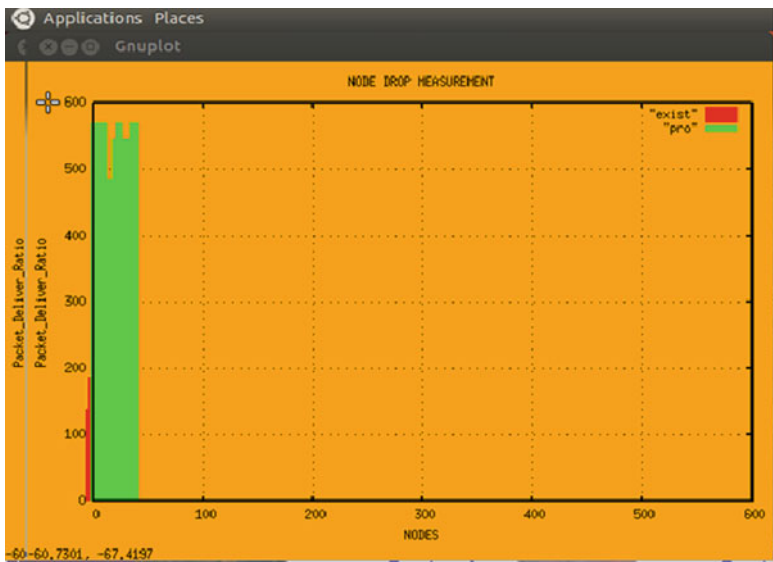


Fig. 6.5 Node packet delivery plot

the existing system (ELDC-RED) and proposed system (GSTEB-GREEN). It is observed that the packet delivery ratio is extremely high for the proposed system (GSTEB-GREEN), thereby achieving greater energy efficiency.

## 6.5 Conclusion

In this paper, we have successfully provided the load balance for 40 nodes by means of the tree-based routing algorithm. The protocol used has minimal packet loss and so the reliability of data transmission is improved. Moreover, S-LZW compression algorithm offers excellent data compression which further enhances the performance of the network. It could be inferred that the load balance provided by GSTEB is 100% more than that of ELDC and so the lifetime of the network is greatly improved. An important advantage of our proposed system is that the GSTEB protocol prevents the unauthorized entry of node during the inter process communication by means of certificate revocation. So the sensor nodes are prevented from malicious attack. Additionally, we have provided a QR scanning before the beginning of the simulation of the proposed system. Higher versions of the GSTEB protocol used in our paper can be developed by improvising the methodology utilized. This can be done by updating the branches of the tree without any degradation in the energy. The protocol can be implemented on large servers and finds its application in big data analytics. Real-time environmental monitoring, surveillance, air traffic control, etc. are some of the applications of our paper.

## References

1. R. Nigel Horspool, Improving LZW, in *Data Compression Conference*, 1991, pp. 332–341
2. M. Rangchi, H. Bakhshi, A new energy efficient routing algorithm based on load balancing for wireless sensor networks, in *Seventh International Symposium on Telecommunications*, 2014, pp. 1201–1205
3. F. Bajaber, I. Awan, Centralized dynamic clustering for wireless sensor network, in *International Conference on Advanced Information Networking and Applications Workshops*, 2013, pp. 193–198
4. B. Tang, D. Wang, H. Zhang, A centralized clustering geographic energy aware routing for wireless sensor networks, in *IEEE International Conference on Systems, Man and Cybernetics*, 2013, pp. 1–6
5. M. Baniata, J. Hong, Energy-efficient unequal chain length clustering for wireless sensor networks in smart cities. *Wirel. Commun. Mob. Comput.* **2017**, 1–9 (2017)
6. N. Gautam, J.Y. Pyun, Distance aware intelligent clustering protocol for wireless sensor networks. *J. Commun. Networks* **12**(2), 122–129 (2010)
7. H. Xia, R.H. Zhang, J. Yu, Z.K. Pan, Energy-efficient routing algorithm based on unequal clustering and connected graph in wireless sensor networks. *Int. J. Wireless Inf. Networks* **23**(2), 141–150 (2016)
8. A. Rahmanian, H. Omranpour, M. Akbari, K. Raahemifar, A novel genetic algorithm in LEACH-C routing protocol for sensor networks, in *IEEE in Electrical and Computer Engineering*, 2011
9. L. David, Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
10. R. Zhang, L. Ju, Z. Jia, X. Li, Energy efficient routing algorithm for WSNs via unequal clustering, in *International Conference on Embedded Software and Systems*, 2012, pp. 1226–1231

11. M. Vijayalakshmi, D.S. Rao, Energy-Aware Multicast Clustering (EAMC) with increased Quality of Service (QoS) in MANETs, in *Applied and Theoretical Computing and Communication Technology*, 2016, pp. 793–798
12. A. Mehmood, Z. Lv, J. Lloret, M.M. Umar, ELDC: an artificial neural network based energy-efficient and robust routing scheme for pollution monitoring in WSNs. *IEEE Trans. Emerg. Top. Comput.* **99**, 1 (2017)

# Chapter 7

## An Energy-Efficient Distributed Unequal Clustering Approach for Lifetime Maximization in Wireless Sensor Network



S. Manikandan and M. Jeyakarthic

### 7.1 Introduction

Wireless sensor networks (WSNs) have been widely used in various applications such as surveillance, disaster management, and monitoring the physical conditions of the environment [1] through a group of spatially dispersed and dedicated wireless sensors. The deployed sensor nodes are programmed to perform the task of collecting and forwarding critical sensed data. Since the network experiences large-scale deployment of low-power sensor nodes, it is important to conserve the residual energy of the nodes to prolong the lifetime of the WSN. Clustering [2, 3] is seen as an efficient technique to conserve the scarce energy of the sensor nodes. In clustering, certain nodes are elected as Cluster Heads (CHs) and other nodes are elected as Cluster Members (CMs), which are called as the member nodes of the elected CHs. Through single or multi-hop forwarding, the CMs forward the sensed data to their associated CHs. Furthermore, the CHs aggregate the received data from their CMs and forward the aggregated data to the sink node. Based on the route discovery mechanism [4, 5] of AODV (Ad hoc on-demand Distance Vector Routing) protocol, the forwarders are selected for relaying the data to the sink. The major advantage of allowing the nodes to communicate in multi-hop forwarding is that the transmission congestion could be avoided by reducing the number of transmission links which is achieved by allowing the CMs to simultaneously communicate only with their respective CHs and the CH is instructed to manage the communication of its associated CMs. Through multi-hop communication, the CMs reduce the workload of the CHs by sharing the CH's work of data fusion which helps in minimizing the energy depletion in CHs and eventually extend the lifetime of the

---

S. Manikandan (✉) · M. Jeyakarthic

Department of Computer and Information Sciences, Annamalai University, Chidambaram, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_7](https://doi.org/10.1007/978-3-030-19562-5_7)

63

WSN. For the final data upload at the sink, the CHs forward the fused data to the sink node.

The node deployments in a clustered WSN are generally non-uniform where the distance between the various nodes differs and the depletion of energy among the nodes varies. Thus, it is important that the sensor nodes should not be clustered into equal sized clusters since it will lead to uneven energy depletion among the sensor nodes. Importantly, for the nodes acting as the CHs, the energy depletion will be higher since they are loaded with additional responsibilities of managing the data from the CMs and forwarding the aggregated data to the sink which increases the consumption of energy in the CHs. Therefore, it is important to implement unequal clustering [6–8] over the deployed sensor nodes to achieve fair energy consumption by balancing the load among the nodes. The major advantage of such distributed clustering technique is that it does not need global network-related topological information since the nodes forms clusters among themselves based on their own information and the information obtained from the neighboring nodes. Thus, compared to the centralized technique, the communication overhead in cluster-based technique is greatly reduced by eliminating the need to frequently communicate with the sink node for updating the status regarding the nodes in the network and hence it is more practical for implementing in WSN.

It is obvious that the nodes elected as CHs consume more energy compared to the nodes acting as the CMs. During intra-cluster data forwarding which is performed in multiple-hops, the sensor nodes closer to the CH are assigned with the task of frequently relaying the forwarded data from the farther CMs to the CH of that particular cluster. Such frequent forwarding leads to reduced network lifetime due to greater energy loss of the CMs which are frequently involved as relay sensor nodes within the cluster. Particularly, in applications such as vehicular WSNs [9, 10], the processes of establishing routing links demands higher energy consumption and hence in case of premature relay node death, additional energy consumption is incurred due to frequent reestablishment of routing links. This highly disturbs the stability of data transmission in WSNs. Thus, to ensure efficient data transmission in WSN, it is important to conserve the residual energy of the neighboring sensor nodes of the CHs which frequently participates in the process of data delivery.

To design an Energy-Efficient Distributed Unequal Clustering (EEDUC) technique, it is necessary to consider various important factors like node degree, energy of current and neighboring nodes, etc. Thus, through consideration of multiple essential factors, the cluster stability could be drastically improved by electing appropriate CHs [11]. However, such a CH election leads to high degree of uncertainty. Fuzzy logic is seen as an optimal solution for such multi-parameter decision problem where the CH should be elected by integrating different clustering parameters.

Figure 7.1 shows the framework of the fuzzy logic inference system. Through the fuzzifier, the crisp input parameters such as nodes' residual energy are converted into the fuzzy linguistic input variables, and furthermore the process of defuzzification is performed by the defuzzifier. The inference system is used to convert the fuzzy linguistic input variables into crisp values. The knowledge base contains the

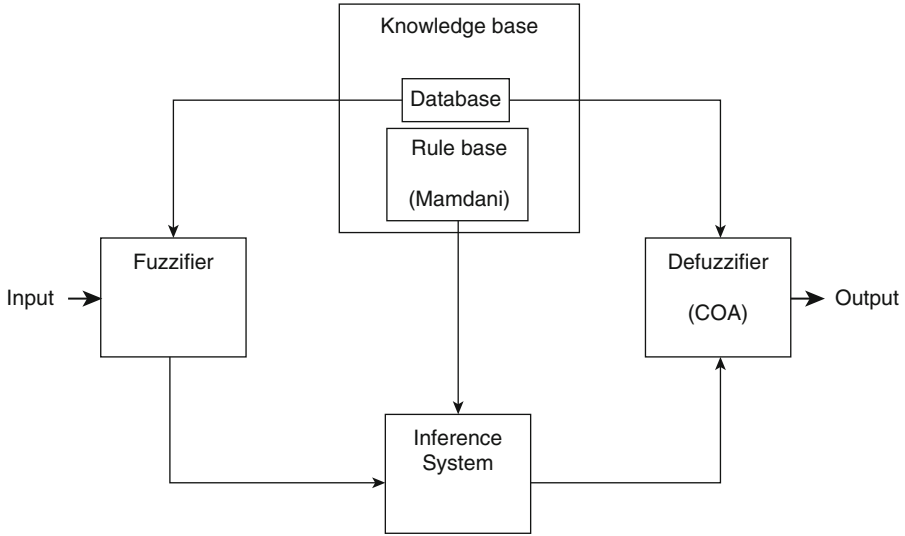


Fig. 7.1 Fuzzy inference system of EEDUC

database and rule base. The database contains the necessary membership functions for processing the input variables. The rule base contains the required fuzzy *if then* rules for analyzing the input variables.

The Mamdani [12] inference system is used by most of the fuzzy systems in which the fuzzy sets define the consequent and premise parameters. The fuzzy logic’s theoretical foundation is laid by the Mamdani system. The form of the fuzzy rules is depicted below:

$$\begin{array}{l} \text{IF } g_1 \text{ is } I_{1a} (m_{1a}, \sigma_{1a}) \text{ and } g_2 \text{ is } I_{2a} (m_{2a}, \sigma_{2a}) \dots g_n \text{ is } I_{na} (m_{na}, \sigma_{na}) \\ \text{THEN } h \text{ is } J_a(m_a, \sigma_a) \end{array}$$

where the Gaussian membership function is represented with mean  $m_{na}$  and deviation  $\sigma_{na}$ . The dimension is represented as  $n$  and rule as  $a$ . In *if then* rules, the consequence parameter is translated into a fuzzy set.

## 7.2 Overview of Existing Works

The Akyildiz and Su proposed a Low-Energy Adaptive Clustering Hierarchy (LEACH) [13] is a classical cluster-based routing algorithm implemented for WSN. LEACH optimally elects the CH through which the network wide energy balance is achieved and maximizes the network scalability. In LEACH, CH election is based on a random number, which is generated between 0 and 1 and a probability  $P_i$ . The

communication among the nodes is scheduled using the Time Division Multiple Access (TDMA) technique. The node is elected as a CH if the generated random number falls within  $P_i$ .  $P_i$  is defined according to Eq. (7.1).

$$P_i = \begin{cases} \frac{k}{N - k * \left( r \bmod \frac{N}{K} \right)} & C_i = 1 \\ 0 & C_i = 0 \end{cases} \quad (7.1)$$

where the total number of sensor nodes is denoted by  $N$ , the desired average number of CHs per round is denoted by  $k$ , the current round number is denoting by  $r$ , the indicator function  $C_i$  determines whether sensor node  $i$  has been a CH for the most recent  $r \bmod (N/K)$  rounds. Thus,  $P_i$  ensures each sensor node is chosen as a CH once per round of operation and balances the network workload among the node which helps in delaying the death of the first node. LEACH provided the benchmark for designing future practical clustering techniques for implementation in WSN. Though LEACH balances the transmission energy consumption in the real-time communication environment, the major drawback is the random selection of CHs which is considered inefficient and may not ensure optimal selection of CH.

The authors [14] proposed a hybrid energy-efficient distributed (HEED) clustering algorithm, which performs CH selection based on the residual energy of a node and the cost incurred for performing the intra-cluster communication. For performing inter-cluster communication, multi-hop communication among the cluster heads is used. HEED is achieved in maximizing the lifetime of WSN but the transmission load is not effectively balanced since the sensor nodes nearer to the sink are forced to premature death. Moreover, HEED demands additional energy consumption owing to higher number of control messages being broadcasted for establishing clusters. In distributed energy-efficient clustering algorithm (DEEC) [15], the CH election was performed by considering the ratio of node's residual energy and the average residual energy of the network. However, DEEC failed to efficiently avoid the energy-hole problem of many-to-one data gathering WSNs.

The authors [16] proposed a Cluster Head Election mechanism using Fuzzy logic (CHEF) which integrates the concept of fuzzy logic with clustering for optimizing the conservation of energy in WSNs. The input fuzzy parameters chosen by CHEF are distance and residual energy of the node to compute its probability to act as CH. However, CHEF ignored the impact of node's degree in affecting the network's energy consumption. The authors [17] used a distributed fuzzy logic control (DFLC) approach which uses the fuzzy logic engine to minimize the number of messages transmitted. DFCL maximizes the accuracy of the candidate CH nodes by adding a filtering system prior to the election of the CHs. The input fuzzy parameters considered in DFCL scheme are distance to sink, residual energy, and density of the node. The COA method is used to convert probability, which is the output parameter to crisp values. Though, DFCL do not efficiently balance the network load which is due to the ignorance of the hotspot issue in multi-hop data transmission.

The authors [18–23] proposed Distributed Unequal Clustering using Fuzzy logic (DUCF) scheme which performs clustering through the concept of fuzzy

logic. The selection of CHs is performed on the basis of fuzzy input parameters, namely, distance to sink, residual energy, and node temperature. The authors [24–36] proposed an Energy-Efficient Distributed Clustering algorithm using Fuzzy approach (EEDCF) which follows non-uniform distribution for implementation in WSNs. The fuzzy input parameters considered for selection of CHs are node degree, residual energy, and residual energy of neighbor nodes. Considering the energy of the neighbor nodes allows the EEDCF scheme in optimizing the network load balance of the forwarder nodes and thus avoids the problems of hotspot which occurs due to operating in multi-hop transmission mode. However, the ignorance of important parameters such as distance and node centrality affect the efficiency of optimal CH selection which further leads to inefficient energy conservation in WSNs.

### 7.3 Proposed Method

Despite the fact that the related works show enhancement when compared with one another, there is no algorithm which minimizes the overall depletion of energy in the WSN and simultaneously maximize the network stability. Hence, to overcome the above issues, a novel fuzzy inference system-based energy-efficient technique called EEDUC is developed in this paper. EEDUC primarily aims at maximizing the lifetime of the network which is achieved by assuring equal consumption of energy among the nodes deployed in WSN. Thus, ensuring equal energy consumption will ensure the delay of node death which in turn increases the network stability period. The following shows the important features of EEDUC:

EEDUC follows unequal clustering where the fuzzy logic is applied for electing the appropriate in a two-phase system.

EEDUC utilizes the Mamdani fuzzy inference system for selecting the CH among the neighboring nodes. For maintaining the local topology, the nodes only communicate among its neighbors.

#### 7.3.1 EEDUC: Fuzzy Parameters

##### *Input parameters*

*Residual Energy (RE)*: It indicates the sensor node's total remaining energy.

*Node Degree (ND)*: It indicates the number of neighboring sensor nodes of a sensor node. Intra-cluster communication cost increases with larger number of sensor nodes. Therefore, the sensor nodes with higher node degree should comprise smaller cluster size.



*Distance to sink (DS)*: It refers to the distance from the node to the sink. The idea behind considering this parameter is that the CHs nearer to the sink should have smaller sized clusters than the CHs far away from the sink.

*Node Centrality (NC)*: It indicates the extent of the sensor node to be located centrally to its neighboring nodes. The intra-cluster distance for communication decreases with higher node centrality.

*Neighbor nodes' Residual Energy (NRE)*: It indicates the average residual energy of the neighbor nodes. Since the nodes operate in multi-hop mode for data transmission within the cluster, the neighboring nodes should have larger residual energy. The cluster size increases with higher residual energy neighboring nodes.

#### *Output parameters*

*Cluster size*: It indicates the number of nodes a CH can accommodate. An energy-efficient CH can accommodate higher number of nodes.

*Chance*: It indicates the probability of a sensor node to act as a CH.

Both cluster size and chance depend on the above-discussed five input parameters.

### **7.3.2 Network Assumptions**

The following assumptions are considered while designing the EEDUC clustering protocol.

The nodes are randomly deployed in the WSN area with each node having a distinctive node ID.

The node death is possible only after complete depletion of its residual energy.

The nodes and sink are static after deployment and distance between the nodes and sink is computed based on the received signal strength indicator (RSSI).

The transmission between the CH and its associated CMs is performed in multi-hop transmission mode.

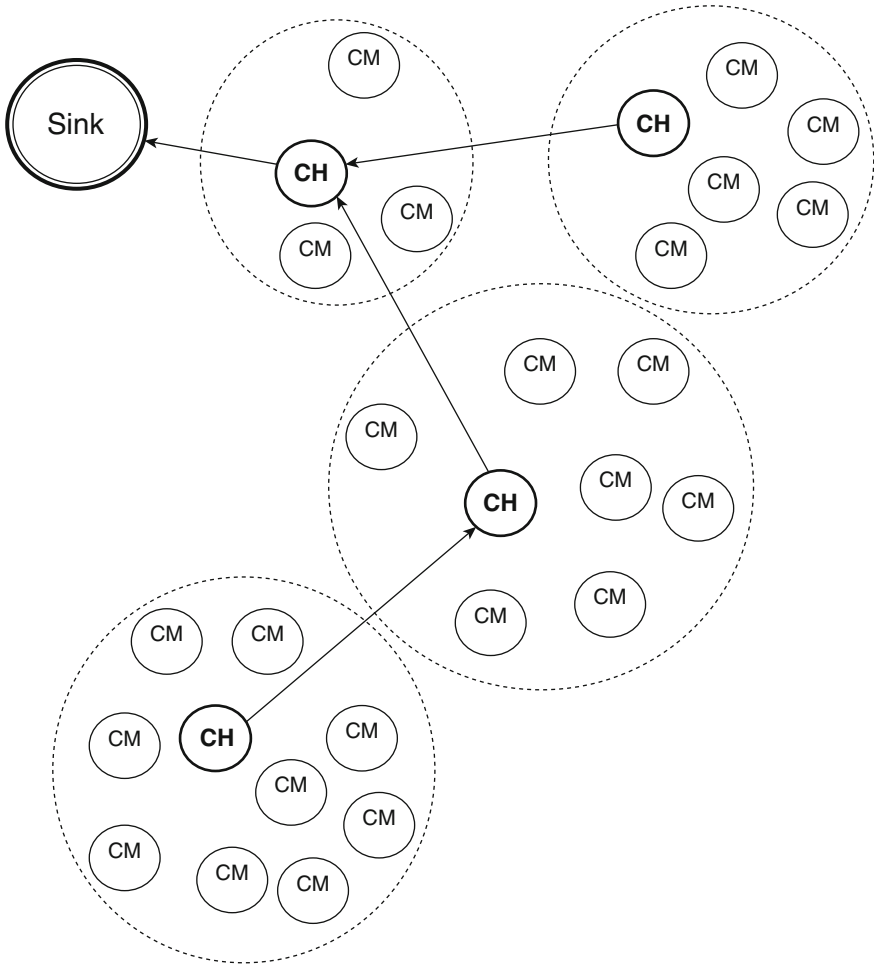
The MAC layer is implemented with Carrier-sense multiple access with collision avoidance (CSMA/CA technique).

### **7.3.3 Concepts**

The concept of using fuzzy inference system logic is to split the entire network into different clusters with suitable cluster size which could allow them to balance the depletion of energy. Figure 7.2 illustrates the process of unequal clustering.

EEDUC is implemented on the basis of four major states for each node.

*Start state*: A node is in an independent state.



**Fig. 7.2** Unequal clustering

*Candidate CH state:* A node is fit enough to enter the competition to become the final CH. Initially all the deployed sensor nodes are considered to be in the candidate CH state.

*CH state:* A node is selected as one of the final CH.

*CM state:* A node is selected as the cluster member and transmits the sensed data to its associated CH.

In every round, initial tentative cluster heads are chosen by creating an arbitrary number for each node. If the generated arbitrary number is less than the probability value (TH) of the nodes given in Eq. (7.3), then it becomes a tentative cluster head.

### 7.3.4 EEDUC: CH Election

On election of candidate CHs, the fuzzy linguistic variable is assigned to the corresponding fuzzy input variable. RE, ND, DN, NC, NRE, RE: Low, Medium, High, ND: Less, Average, Enormous, DS: Near, Reachable, Far, NC: Low, Medium, High NRE: Low, Medium, High. Table 7.1 describes the input parameters and the linguistic variables for selecting the CH and the size of the cluster. The boundary variables, namely low, high, near, far, less, enormous, very low, very high, very small, rather large, rather small, rather high, and very large, are represented using the trapezoidal membership function. The middle variables are represented using the triangular membership function.

$$\mu_A(x) = \begin{cases} 0 & x \leq a_1 \\ \frac{x-a_1}{b_1-a_1} & a_1 \leq x \leq b_1 \\ \frac{c_1-x}{c_1-b_1} & b_1 \leq x \leq c_1 \\ 0 & c_1 \leq x \end{cases} \tag{7.2}$$

$$\mu_A(x) = \begin{cases} 0 & x \leq a_1 \\ \frac{x-a_2}{b_2-a_2} & a_2 \leq x \leq b_2 \\ 1 & b_2 \leq x \leq c_2 \\ \frac{d_2-x}{d_2-c_2} & c_2 \leq x \leq d_2 \end{cases} \tag{7.3}$$

**Table 7.1** Fuzzy IF-THEN rules

Rule	Input					Output	
	RE	ND	DS	NC	NRE	Chance	Cluster size
1.	High	Enormous	Near	High	High	Very high	Rather large
2.	High	Enormous	Near	High	Medium	High	Large
3.	High	Enormous	Near	High	Low	Rather high	Very large
4.	High	Enormous	Near	Medium	High	Very high	Medium
5.	High	Enormous	Near	Medium	Medium	High	Medium
6.	High	Enormous	Near	Medium	Low	Rather high	Medium
7.	High	Enormous	Near	Low	High	Very high	Very small
8.	High	Enormous	Near	Low	Medium	High	Small
9.	High	Enormous	Near	Low	Low	Rather high	Rather small
10.	High	Enormous	Reachable	High	High	Very high	Rather large
11.	etc.						
240	Low	Less	Far	Medium	Low	Very low	Medium
241	Low	Less	Far	Low	High	Rather low	Very small
242	Low	Less	Far	Low	Medium	Low	Small
243	Low	Less	Far	Low	Low	Very low	Rather small

For the process of defuzzification, center of area (COA) method is used as to obtain the crisp value, cluster size, and chance, as shown in Eq. 7.4.

$$\text{COA} = \frac{\int \mu_A(x) \cdot x dx}{\int \mu_A(x) dx} \quad (7.4)$$

In our application areas, proficient information is designated in terms of imprecise (“fuzzy”) words from a natural language such as “small” or “large” or “high,” etc. To define such words in computer-understandable terms, we can use fuzzy logic techniques. In fuzzy logic, each natural-language word is described by an association function  $\mu(x)$ , a function that assigns, to every number  $x$ , the degree  $\mu(x) \in [0, 1]$  to what this number satisfies the corresponding property (e.g., the degree to which the number  $x$  is small).

## 7.4 Simulation Setup and Performance Metrics

The performance of EEDUC algorithm is compared with a traditional clustering technique called LEACH and two distributed fuzzy logic-based clustering protocols, CHEF and DUCF which are proposed for implementation in WSNs. The abovementioned algorithms are tested in two different network scenarios as described below.

*Scenario 1:* Sink located outside the region of interest (ROI). *Scenario 2:* Sink located in the middle of ROI. The simulation parameters are shown in Table 7.2. The residual energy of all the sensor nodes is equal during initial deployment. The simulation was performed using MATLAB. Using the toolbox environment of MATLAB, the FIS for CHEF, DUCF and EEDUC are built.

The performance metrics used to compare different clustering algorithms are described below:

1. Network lifetime: The completion of number of rounds before the First Node Death (FND) and Half Node Death (HND).
2. Per round energy consumption: For a single round of communication, the total energy consumed for protocol operation which involve the CHs and CMs.
3. Message transmission count: Till HND, the total number of data messages transmitted from CHs to the sink node. Performance analysis.

### 7.4.1 Network Lifetime

It is important for any WSN-based algorithm to consider the residual energy of each deployed sensor node for enhancing the sensing efficiency since sudden node deaths greatly triggers the degradation of collecting quality sensor information. Importantly, after HND, the performance of the sensor network greatly reduces

**Table 7.2** Simulation parameters

Parameter	Value
Node deployment	Uniform random
Number of nodes	100
Antenna	OmniAntenna
Transmission rate	1 Mbps
Sensors' transmission range	40 m
Data packet size	500 bytes
Control packet size	25 bytes
Number of bits to be transmitted	4000
$d_{th}$	87 m
Traffic type	CBR
Simulation time	1000 s
Total number of simulations runs	75
Number of simulations runs for each scenario	25
Back-off mechanism	CSMA/CA
Sensor initial energy	1000 mJ
$E_{elec}$	50 nJ/bit
$\epsilon_{fs}$	10 pJ/bit
$\epsilon_{mp}$	0.0013 pJ/bit
Desired percentage of CHs in LEACH, CHEF	[0.1, 0.3]

which further triggers the collapse of the entire network. Therefore, it is important for any clustering algorithm deployable in WSN to ensure the delay of both FND and HND. It is observed from the simulation results that EEDUC shows better performance than LEACH, CHEF, and DUCF in terms of both FND and HND.

According to Fig. 7.3, in scenario 1, the distance between the CH and the sink increases, the clustering algorithms, namely, LEACH, CHEF, and DUCF, shows lowliest performance in terms of delaying the node death since they are involved in single hop communication which eventually increases the communication cost of the CHs to transmit the data to the sink node. Hence, as a result, the energy of the CHs gets rapidly reduced which increases the chances of sudden network collapse. However, both DUCF and EEDUC follow multichip data forwarding and hence the energy consumption is minimized to a certain extent and eventually delays the death of the sensor nodes. Though, in all the four clustering algorithms, the larger distance between the sink and the nodes has a negative effect in terms of the communication cost and thus incurs higher energy consumption.

According to Fig. 7.4, in scenario 2, it is clear that EEDUC extends the FND by around 300 rounds in all 25 simulations as compared to LEACH clustering. CHEF performs better clustering performance than LEACH due to its ability to select the CHs based on energy and also efficiently avoids the competition among the nodes within each of their transmission range. Though, CHEF delays the FND metric, it shows decreased performance than LEACH when half of the nodes in the network die and from that instance the remaining nodes start to act as independent CHs which

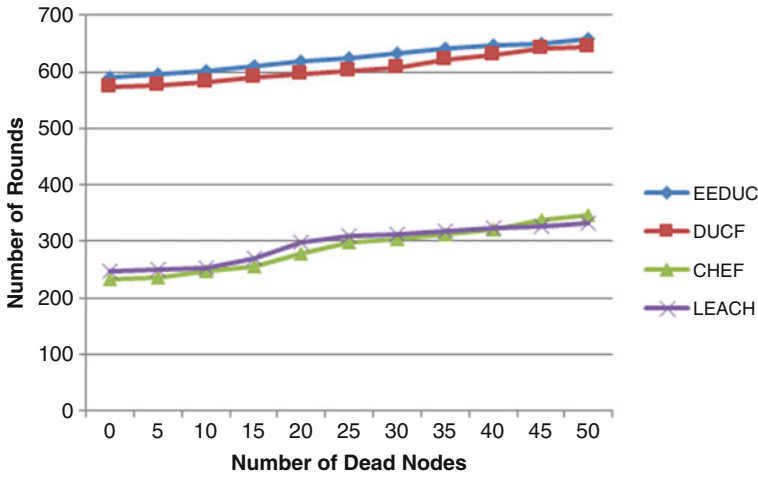


Fig. 7.3 Scenario 1—Node death in terms of number of rounds

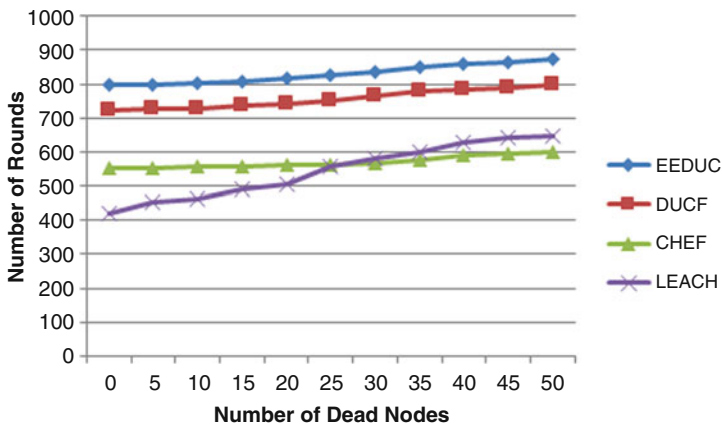


Fig. 7.4 Scenario 2—Node death in terms of number of rounds

eventually increases the communication cost for direct transmission and increases the chances of sudden network death. The reason behind EEDUC showing improved performance than the other two fuzzy-based clustering techniques is because of the consideration of two essential parameters, namely, node centrality (NC) and neighbor nodes' average residual energy (NNE), in addition to residual energy of a node, distance to sink, and node degree. Considering these two additional parameters into FIS improves the intra-cluster transmission cost and efficiently assigns the optimal number of CMs to each elected CH.

### **7.4.2 Per Round Energy Consumption**

For the two scenarios, the energy consumed for one round of protocol operation in LEACH, CHEF, DUCF, and EEDUC is shown in Fig. 7.4. It is observed that the energy consumption in LEACH clustering technique is higher which is due to the following reasons. The LEACH technique does not perform CH election considering any important energy-related parameters. Also, the CHs are allowed to transmit the aggregated data directly to the sink node which greatly increases the energy consumption of the CHs. Though, CHEF considers energy-related parameters during CH election which shows better performance than LEACH, it is affected by the policy of direct transmission like LEACH and hence increases the energy consumption during transmission. DUCF shows better performance than LEACH and CHEF through consideration of important parameters like Node degree and moreover allowing multi-hop transmission for communicating with the sink node. However, ignoring the other important energy-related factors such neighbor node's energy and the centrality of a node during CH election results in non-optimal selection of CHs and hence incurs unnecessary energy consumption. EEDUC minimizes energy depletion by considering node centrality and neighbor nodes' residual energy for selecting optimal CHs and also allows the transmission to the sink node from the CHs through multi-hop transmissions.

## **7.5 Conclusion and Future Works**

This research work proposes EEDUC, which is a fuzzy interface-based energy-efficient clustering technique for implementation in WSN. It is important for any WSN to receive appropriate valuable information during the course of its operation. The data messages aggregated from the nodes will be valuable until HND. EEDUC minimizes energy depletion by considering node centrality and neighbor nodes' residual energy for selecting optimal CHs and also allows the transmission to the sink node from the CHs through multi-hop transmissions. EEDUC achieves network lifetime maximization through optimal balancing of sensor residual energy. EEDUC allows a cluster to optimally decide its size based on Residual Energy, Distance to sink, Node Degree, Node Centrality, and Neighbor nodes' Residual Energy. Moreover, EEDUC allows multi-hop data transmission for minimizing the energy depletion during data transmission. From the simulation results, it is observed that EEDUC outperforms other existing clustering techniques and proves its suitability for implementation in energy-constrained WSN applications. The future work will consider more additional and relevant clustering parameters such as link quality and also implement the system with multiple sinks and mobile sensor nodes.

## References

1. D. Rajendra Prasad, P.V. Naganjaneyulu, K. Satya Prasad, A hybrid swarm optimization for energy efficient clustering in multi-hop wireless sensor network. *Wirel. Pers. Commun.* **94**, 2459–2471 (2017)
2. A. Pughat, V. Sharma, A review on stochastic approach for dynamic power management in wireless sensor networks. *Hum. Centric Comput. Inf. Sci.* **5**, 4 (2015)
3. N. Kumar, J. Kim, ELACCA: efficient learning automata based cell clustering algorithm for wireless sensor networks. *Wirel. Pers. Commun.* **73**, 1495–1512 (2013)
4. J. Huang, Y. Hong, Z. Zhao, Y. Yuan, An energy-efficient multi-hop routing protocol based on grid clustering for wireless sensor networks. *Clust. Comput.* **20**, 3071–3083 (2017)
5. Energy Efficient Backoff Hierarchical Clustering Algorithms for Multi-Hop Wireless Sensor Networks, <https://link.springer.com/article/10.1007/s11390-011-9435-4>. Accessed 5 Nov 2018
6. R. Priyadarshi, S.K. Soni, V. Nath, Energy efficient cluster head formation in wireless sensor network. *Microsyst. Technol.* **24**, 4775–4784 (2018)
7. W. Zhou, Energy efficient clustering algorithm based on neighbors for wireless sensor networks. *J. Shanghai Univ. Engl. Ed.* **15**, 150–153 (2011)
8. K. Guravaiah, R. Leela Velusamy, Energy efficient clustering algorithm using RFD based multi-hop communication in wireless sensor networks. *Wirel. Pers. Commun.* **95**, 3557–3584 (2017)
9. Energy Efficient Clustering Scheme (EECS) for Wireless Sensor Network with Mobile Sink, <https://link.springer.com/article/10.1007/s11277-018-5653-1>. Accessed 5 Nov 2018
10. M. Ulema, J.M. Nogueira, B. Kozbe, Management of wireless ad hoc networks and wireless sensor networks. *J. Netw. Syst. Manag.* **14**, 327–333 (2006)
11. D. Yun-Zhong, L. Ren-Ze, Research of energy efficient clustering algorithm for multilayer wireless heterogeneous sensor networks prediction research. *Multimed. Tools Appl.* **76**, 19345–19361 (2017)
12. Triangular fuzzy-based spectral clustering for energy-efficient routing in wireless sensor network, <https://link.springer.com/article/10.1007/s11227-018-2357-y>. Accessed 5 Nov 2018
13. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey. *Comput. Netw.* **38**(4), 393–422 (2002)
14. A.A. Abbasi, M. Younis, A survey on clustering algorithms for wireless sensor networks. *Comput. Commun.* **30**(14–15), 2826–2841 (2007)
15. M. Liu, Y. Zheng, J. Cao, G. Chen, L. Chen, H. Gong, An energy-aware protocol for data gathering applications in wireless sensor networks, in *Proceedings of the IEEE International Conference on Communications*, Glasgow, UK, 24–28 Jun 2007, pp. 3629–3635
16. J.M. Kim, S.H. Park, Y.J. Han, T.M. Chung, CHEF: cluster head election mechanism using fuzzy logic in wireless sensor networks, in *Proceedings of the 10th International Conference on Advanced Communication Technology (ICACT)*, Gangwon-Do, Korea, 17–20 Feb 2008, pp. 654–659
17. J. Yu, Y. Qi, G. Wang, Q. Guo, X. Gu, An energy-aware distributed unequal clustering protocol for wireless sensor networks. *Int. J. Distrib. Sens. Networks* **2011**, 202145 (2011)
18. F. Bajaber, I. Awan, Adaptive decentralized re-clustering protocol for wireless sensor networks. *J. Comput. Syst. Sci.* **77**(2), 282–292 (2011)
19. C.E. Perkins, E.M. Belding-Royer, S.R. Das, Ad hoc on demand distance vector (AODV) routing. IETF RFC 3561, 2003, pp. 1–67
20. A. Yadav, Y.N. Singh, R.R. Singh, Improving routing performance in AODV with link prediction in mobile adhoc networks. *Wirel. Pers. Commun.* **83**(1), 603–618 (2015)
21. V. Gupta, R. Pandey, An improved energy aware distributed unequal clustering protocol for heterogeneous wireless sensor networks. *Eng. Sci. Technol. Int. J.* **19**(2), 1050–1058 (2016)



22. S. Thompson, K. Suresh Joseph, Particle swarm optimization-based energy efficient channel assignment technique for clustered cognitive radio sensor networks. *Comput. J.* **61**(6), 926–936 (2018)
23. S. Thompson, K. Suresh Joseph, Cognitive radio assisted OLSR routing for vehicular sensor networks. *Proc. Comput. Sci.* **89**, 271–282 (2016)
24. S. Thompson, K. Suresh Joseph, PSO assisted OLSR routing for cognitive radio vehicular sensor networks, in *Proceedings of the International Conference on Informatics and Analytics*, 2016, pp. 1–8
25. S. Naeimi, H. Ghafghazi, C.O. Chow, H. Ishii, A survey on the taxonomy of cluster-based routing protocols for homogeneous wireless sensor networks. *Sensors (Switzerland)* **12**(6), 7350–7409 (2012)
26. E.H. Mamdani, Application of fuzzy logic to approximate reasoning using linguistic synthesis, in *Proceedings of the 1997 27th International Symposium on Multiple-Valued Logic*, Los Alamitos, CA, 28–30 May 1997
27. W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocols for wireless microsensor networks, in *Proceedings of the Hawaii International Conference on Systems Sciences*, Maui, HI, 4–7 Jan 2000, pp. 1–10
28. O. Younis, S. Fahmy, HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans. Mob. Comput.* **3**, 366–379 (2004)
29. L. Qing, Q. Zhu, M. Wang, Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Comput. Commun.* **29**, 2230–2237 (2006)
30. Y. Liao, H. Qi, W. Li, Load-balanced clustering algorithm with distributed self-organization for wireless sensor networks. *IEEE Sensors J.* **13**, 1498–1506 (2013)
31. D. Lin, Q. Wang, D. Lin, Y.A. Deng, Energy-efficient clustering routing protocol based on evolutionary game theory in wireless sensor networks. *Int. J. Distrib. Sens. Netw.* **2015**, 1–12 (2015)
32. A. Alaybeyoglu, A distributed fuzzy logic-based root selection algorithm for wireless sensor networks. *Comput. Electr. Eng.* **41**, 216–225 (2015)
33. R. Dutta, S. Gupta, M. Das, Low-energy adaptive unequal clustering protocol using fuzzy c-means in wireless sensor networks. *Wirel. Pers. Commun.* **79**, 1187–1209 (2014)
34. D.M.S. Bhatti, N. Saeed, H. Nam, Fuzzy C-means clustering and energy efficient cluster head selection for cooperative sensor network. *Sensors* **16**, E1459 (2016)
35. B. Baranidharan, B. Santhi, DUCF: distributed load balancing unequal clustering in wireless sensor networks using fuzzy approach. *Appl. Soft Comput.* **40**, 495–506 (2016)
36. Y. Zhang, J. Wang, D. Han, H. Wu, R. Zhou, Fuzzy-logic based distributed energy-efficient clustering algorithm for wireless sensor networks. *Sensors* **17**, 1554 (2017)

# Chapter 8

## An Effective Big Data and Blockchain (BD-BC) Based Decision Support Model for Sustainable Agriculture System



M. Dakshayini and B. V. Balaji Prabhu

### 8.1 Introduction

Agriculture has a major share in the economy of developing countries. India being a developing country, over 58% of the rural households depends on agriculture as their principal means of livelihood, and agriculture is one of the largest contributors to the GDP of the country [1]. But the agricultural system in the developing countries is lagging ineffective use of advanced technologies available, and hence facing many hurdles. According to National Crime Records Bureau's latest farmer suicides data, over 6867 farmers had committed suicide across the country India in 2015–16 [2]. This is primarily due to the failure to pay back loans raised from banks and microfinance institutions [3, 4].

There is a myth in the agriculture that more yields give more profit in crop business. Everyone is working towards improving the yield and production of the crop without bothering about the actual demand for the same. With the lack of actual demand information and the other farmer's choices, if the farmers produce more goods in the verge of making more profit, there will be more supply than demand resulting in losses. Conversely, if there is less supply in the market, then the consumer suffers from the high prices.

This is mainly because of the lack of management between supply and demand for various food produces in the agricultural system. This gap can be minimized by developing an efficient demand-supply management service system. This service system maintains the demand and supply for all the food produces and guides the farmers in making a wise decision in selecting appropriate crop for cultivation, con-

---

M. Dakshayini · B. V. Balaji Prabhu (✉)  
Department of Information Science and Engineering, BMS College of Engineering, Bangalore,  
India  
e-mail: [dakshayini.ise@bmsce.ac.in](mailto:dakshayini.ise@bmsce.ac.in)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_8](https://doi.org/10.1007/978-3-030-19562-5_8)

77

sidering the actual desires of the society and hence decreasing the demand-supply miss match which may lead to unexpected variations in the market conditions. Hence, it is imperative for farmers to adopt technology-based management service system, make critical decisions thereby getting better yields and profits.

Upon the production of needy harvest the farmer problem will not end. The big problem is supply chain management. In most of the developing countries, the supply chain of food commodities is inefficient due to information asymmetry which is also a main reason for low farmer income. The current food supply chain is a complex network, as it involves various characters between farmers to consumers range from brokers, distributors, processors, retailers, regulators, etc. There is no transparency in any phase of this supply chain and avoiding a middleman in each phase is a headache. There are certain bodies which will hold the supply of some commodities for a certain time period to create a demand in the market with the intention of making a profit. All these limitations make the farmers suffers in their farm business or consumers for their daily usage.

The integration of Big Data [5], Cloud [6], and Blockchain [7] technology in agriculture will bring revolutionary changes. This paper proposed a Big Data, Blockchain, and Cloud based efficient crop management system achieving effective demand-based decision support and simplified, transparent, and secure supply chain system aiming to provide technological solutions for the present problems of agricultural sectors.

The Blockchain technology is revolutionizing the different domains of the society such as financial services [8, 9], health care [10], Supply chain management [11], and the land ownership registration [12]. Blockchain technology is also transforming other domains of the society like education, entertainment, government, protection of rights, human resource management, retail industry, business and the list goes on.

Agriculture is a field which also requires a face shift from traditional farming methods to modern technological methods. There is a need of a forum or a system which could assist the farmers' right from the crop selection till they receive their proper shares for their harvest from the market. This avoids the middleman's interruption in all stages of crop farming. Blockchain and Big data are such kinds of technologies which can bring the radical changes in the field of agriculture to make our farmers happy in their farm business. The main contributions of this work are as follows.

- The proposed system synchronizes the demand and supply, thereby reducing the gap between supply and demand. Hence it reduces the loss for the farmers and also the price inflation for food crops making consumers happy.
- A Blockchain-based crop-trading platform is proposed to trade the yields in different phases like village, Taluks, and District levels to avoid the middleman intervention. The proposed system will manage the supply at various levels based on the demand.

The remaining portion of the paper is organized as follows: Sect. 8.2 illustrates the proposed crop management system with different modules. Section 8.3 discusses the implementation and results, Sect. 8.4 concludes the work.

## 8.2 An Effective BD-BC-Based Crop Management System

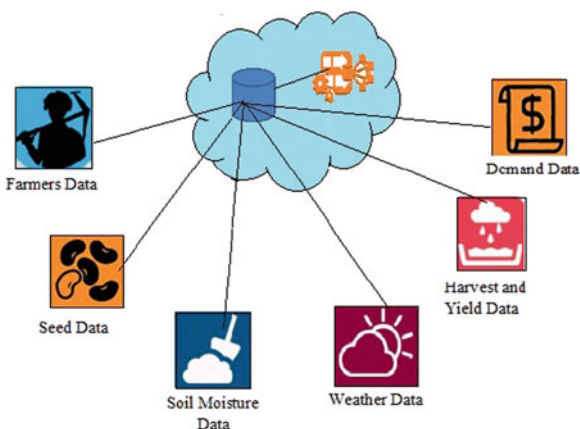
In order to match the supply of the food commodity from the farmers as per the need of the consumer, the demand-based crop selecting system is proposed. The proposed decision support system predicts the demand and assists the farmers in cultivating the proper crops required to match the demand, thereby maintaining the equilibrium in the market conditions.

### 8.2.1 Demand-Based Efficient Decision Support System for Suitable Crop Selection

The proposed system maintains various data like farmer’s land-related information, soil moisture data, weather and environmental data, harvest and yield data, and demand and supply data in a cloud-based framework as shown in Fig. 8.1. These data helps the data analytics module of the system to suggest the best suitable crop for the farmer to harvest for which there will be a demand. Keeping track of the list of crops with the corresponding demand and also the total amount of yield expected for each crop in the list by various farmers avoids the overabundance by impeding farmers not to select the crop which has already been matched with the predicted demand.

The system architecture of the proposed crop selection system is shown in Fig. 8.2. A farmer can avail the services provided by this system by registering

**Fig. 8.1** Cloud-based framework to assimilate the land-related information in cloud



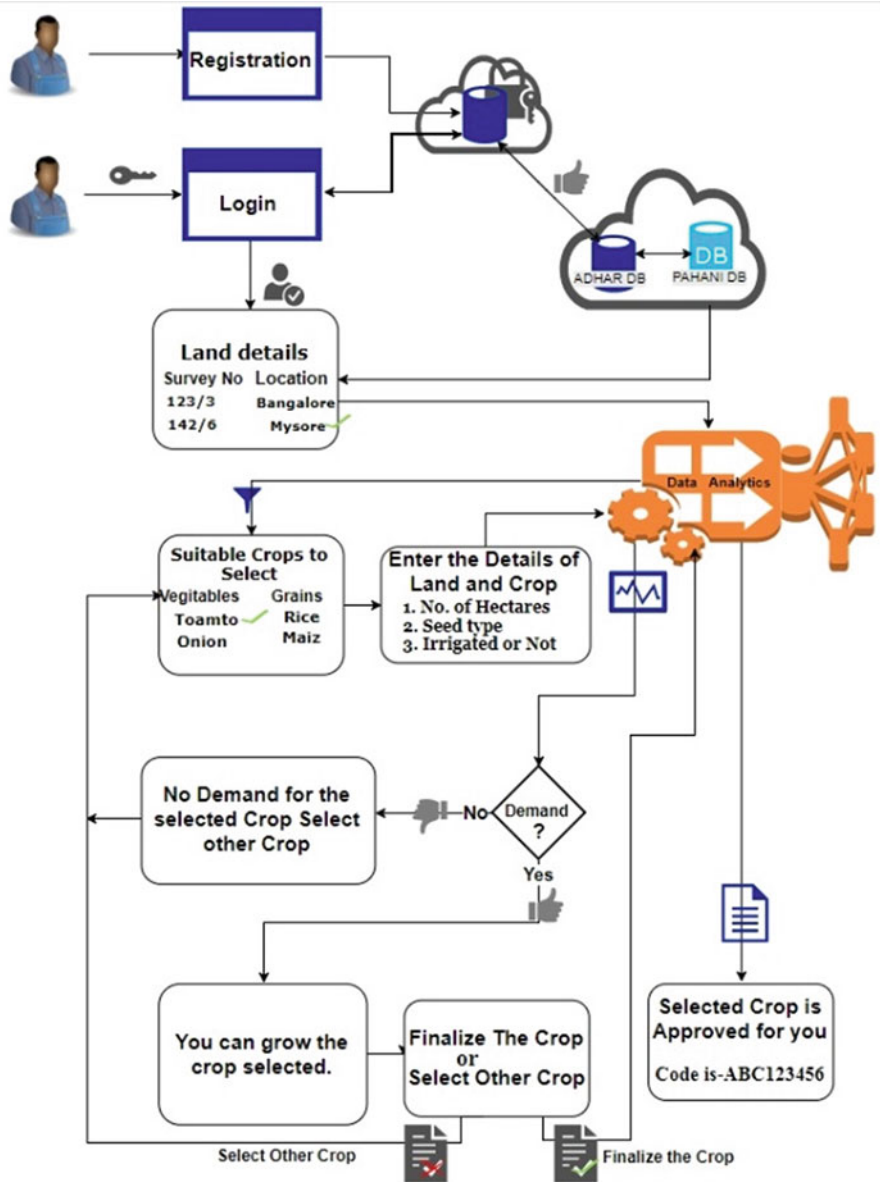


Fig. 8.2 System Architecture of the proposed crop selection model

with the system furnishing “Aadhaar” details, thus allowing only legitimate users and avoiding the unauthorized access to the system. Once the user gets registered with the system, a user ID and password are sent to their registered mobile number.

The registered user can utilize the crop suggestion system to select the crop suitable for his land specifications which will have a demand in future.

When the user logs into the system, the system displays the list of various lands owned by the user. The land details are obtained through the Pahani (Pahani or RTC is an important land record that contains details of land) database, where the Aadhaar data is integrated with Pahani records to get the user-specific land details. The user has to select the land from the list displayed in which the crop is planned to grow.

Based on the land selected and other related parameters like soil type, rainfall details, the season of the year, and the sources of irrigation, the data analytics module of crop suggestion system render a list of suitable crops for the selected land. In the displayed list of suitable crops, the user has to specify one or more crops of his interest and enter the number of acres in which he wishes to grow. Once these details are specified, analytics module will evaluate the expected yield for the selected crop. Analytics module also estimates the harvesting period required to get the yield and forecasts the demand for the selected crop considering the estimated harvesting period. Once the quantity of yield and the harvesting period has been estimated, the estimated quantity of yield is equated with the forecasted demand for the expected period of the harvesting. If the estimated quantity of yield is less than the forecasted demand, the system will recommend the user to go with the crop he had opted to grow. If the estimated quantity of yield is greater than the forecasted demand, the system will alert him and he can adopt another crop from the displayed list of suitable crops and the process repeats.

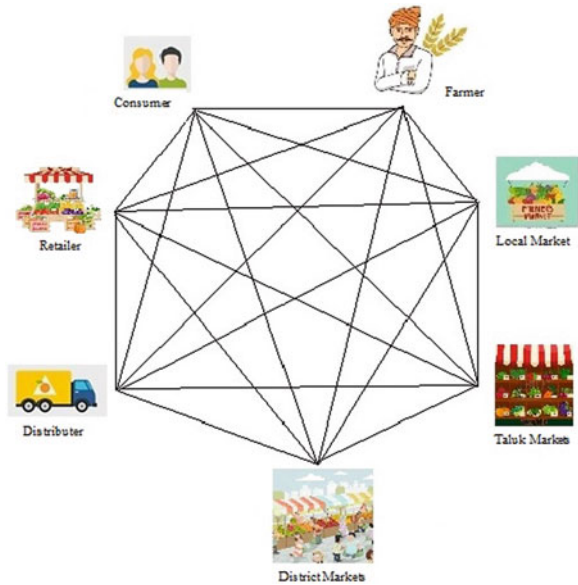
Once the crop has been finalized by the user for cultivation, the demand for that crop is updated by subtracting the estimated quantity of yield of the same which will be used for the next demand comparison. At the end of the inquiry, the user will get a unique code for the successful selection of the crop which is added to ledger of Blockchain.

Even after a valid registration and crop selection process, if the farmer fails to meet the projected yield, then careful appraisal for the possible reasons of failure in production of an estimated quantity of yield is done. If the reason is due to natural calamities or disasters, then the farmers in that area could be considered for compensation from the Government. This system could also be instrumental in helping the Government for identifying the farmer's plight and make necessary policy interpretations tailored to the needs for smart agriculture.

## ***8.2.2 Blockchain-Based Crop Management System***

The proposed food crop supply chain management system implemented using Blockchain technology is shown in Fig. 8.3. The proposed system builds the chain of all different bodies involved in food supply chain ranging from farmer to consumer connected through Blockchain. Each participant in the supply chain requires certification. Certification is performed by accredited agencies. After a

**Fig. 8.3** Blockchain-based food supply chain management system



successful audit, the certification body uploads the audit report to cloud. Every transaction at each phase is registered in ledger which brings the transparency and also the food crop supply can be identified at each phase which gives food security and avoids middleman each phase.

When the Farmers get the yield, the same could be delivered at the Government supported local market in the village. The Manager at the market takes the yield after verifying the farmer with the unique transaction code and the volume of yield estimated for his land using the ledger data with data analytics module. When the farmer makes the trade, corresponding transaction will be added to ledger. After the verifier acknowledges the transaction, farmer account is credited with the rate fixed. After all farmers submitted their harvest, the village market manager sends the collected harvests into Taluk market.

Each Taluk market receives the harvest of different crops from different villages and the same is recorded in a ledger through Blockchain. The Blockchain technology allows the retailers and distributors to purchase the selected yields directly from the farmers and ensures the appropriate and immediate payment for farmers by avoiding brokers and be fraudulent from any illegal sources. The remaining yields will be transported to District level Markets for the distribution and then to state level. As the data about the demand is pre-estimated and the supply data can also be estimated on the number of farmers registered to grow, the export can also be planned earlier to avoid price variation in market.

Blockchain will maintain a traceability in supply chains of agriculture, as the Blockchain ledger could record and update the status of harvest from each and every market. The upside for large operations is a secure, immutable ledger that ensures the never lose a supply. The status of all the crops is available in real time. The

BD-BC-based decision support model manage the food crop wastage in case of excess supply, crop scarcity in case of less supply and middleman's intrusion very efficiently.

### 8.3 Implementation

The data set about the demand and supply for certain crops for the state Karnataka has been gathered from the authorized government websites [13, 14] to analyze the scenario of demand and supply for the year 2016–17. The data values are tabulated as shown in Table 8.1, and the table shows the excess supply and scarcity for the different food crops for the state Karnataka during the year 2016–17. Excess supply for crops like wheat, sugarcane, etc., and results in wastage of food crops and hence farmers will not get an expected profit. In contrary, when there is less supply for any food crops like Rice, Maze, etc., it creates a scarcity for food crops and consumers suffer with the high price.

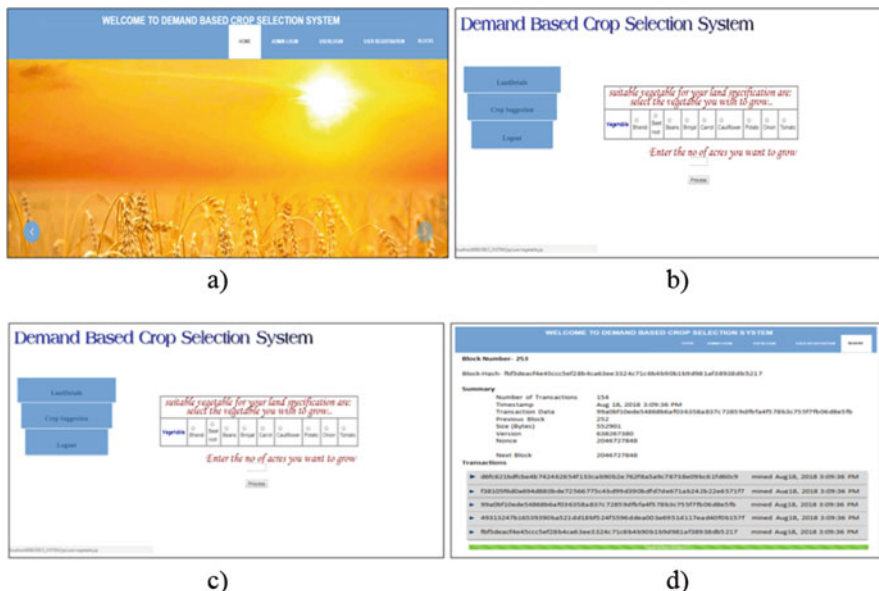
The main aim of BD-BC system is to reduce the gap among the supply and demand of food produces and to bring the transparency in trading of the same. The proposed system has been implemented using Java, which consists of a Web server and client browser. The Web server was implemented using PHP. The client browser was implemented using JavaScript. The Blockchain is designed to record information about crop selection, estimated supply, crop trading, and crop transportation data along with the time stamp. A Web server is accountable for collecting and storing land-related data, demand data, crops data, all transaction data, block-related data, crops supply data, and supply transportation data. The client browser is an interface for farmers to interact with the system for crop selection and allocation as shown in Fig. 8.4.

Datasets used in the process of crop suggestion like land details, rainfall pattern, crops suitable for different land conditions, harvesting durations of different crops are taken from "Profile of agriculture statistics Karnataka state, department of agriculture Bengaluru." The yield for different crops at different regions is taken from the government website [15]. Data sets used for Demand forecasting are

**Table 8.1** Demand supply gap for different food crops for the state Karnataka for the year 2016–17

Crops	Demand	Supply	Gap	Result
Rice	110	102	−8	Scarcity
Wheat	89	108	18	Excess
Maize	19	12	−7	Scarcity
Cereals	235	245	10	Excess
Pulses	22	40	18	Scarcity
Food grains	257	365	108	Excess
Oilseeds	59	26	−33	Scarcity
Sugarcane	279	380	101	Excess





**Fig. 8.4** (a) Welcome page. (b) Crop suggestion phase. (c) Crop selection phase. (d) Block information page

**Table 8.2** Expected reduced gap for the food crops using the proposed methodology

Crops	Demand	Supply	Gap
Rice	110	112	2
Wheat	89	90	1
Maize	19	18	-1
Cereals	235	237	2
Pulses	22	24	2
Food grains	257	260	3
Oilseeds	59	60	1
Sugarcane	279	280	1

considered from government website [16, 17]. For evaluating this model, 15 different vegetables, 20 different cereal crops have been considered. 50 farmers are made to register to the system also demand and supply data of previous 10 years have been considered. Blockchain technology has been used to eliminate the brokerage system and has been implemented using python with a group of 4 farmers, 3 distributors, retailers, consumers and Bank at 2 Taluks and one District level.

The proposed system could successfully map the supply and demand of food crops by efficiently assisting the farmers and regulating the supply according to the demand, thereby reduce the chances of either food wastage or scarcity due to more supply or less supply as shown in Table 8.2.

**Table 8.3** Price inflation through halting the supply

Month	Demand	Supply	Wholesale price	Retail price
Jan	467123.3	3934	15.25	20.32
Feb	421917.9	2295	16.54	20.73
Mar	467123.3	2798	16.78	22.85
Apr	452054.9	2908	13.04	20.47
May	467123.3	3826	16.65	24.34
Jun	452054.9	4258	15.98	23.86
Jul	467123.3	3770	14.27	22.49

**Table 8.4** Regulated demand supply by the proposed system

Demand	Supply	Price
17	15	9
18	17	10
15	17	8
18	14	11
18	16	8
17	18	8

The proposed system could avoid the middleman’s intervention and improve the returns for farmers. Table 8.3 shows the scenario of price inflation due to the intrusion of middleman even when the supply and demand were almost inline. Brokers may supply the yields illicitly in the market with the intention of making the profit. From Table 8.3 it can be observed that when the supply is more than the demand, the wholesale price which the farmer gets is less. When the supply is less than the demand, there is a hike in price and consumers suffers with high price. It can also be observed in that table that, in the month of June and October even when the supply is almost equal to the demand, the retail price is high; this may be due to holding the supply to create a demand.

The proposed system could provide the transparency and traceability at each step of supply chain. This avoids the middleman’s interruption and illegal holding of food supply, thereby reduces the uneven price inflations in the market and also improves the return for farmers.

Table 8.4 shows the reduced demand supply gap and reduced price inflation which is a result of using the proposed BD-BC system. This helps farmers from the loss and also the consumers from high price from food crops.

### 8.4 Conclusion

The Big Data, Blockchain, and Cloud technologies brings revolutionary changes in agriculture system. By making use of technologies available, this proposed model effectively improved the quality of the agricultural system by effectively achieving 90% to 92% match in demand and supply of food crops required by the society from the farmer’s end, thus avoiding the loss for farmers and catering the needs

of consumers. This leads to a gainful crop business for farmers and satisfactory fulfillment of the societal needs. Thus, a judicious mix of extensive physical outreach and interactive methods of information technology could be used for sustainable and better agricultural practices. BD-BC-based decision support model proposed in this work effectively facilitated real-time monitoring of supply chain bringing total transparency and security to agricultural transactions also eliminating the nuisance created by middlemen and curbing the wastage of agricultural produce. In future, an entire network of all the stake holders including farmers, regulatory bodies, processing units, etc. could be build, in the Blockchain. All this will be possible with the help of regulation and consensus system known as smart contracts leading to very less scope for corruption.

## References

1. Ministry of Finance, Govt. of India, An overview of india's economic performance in 2017–18. Economic Survey 2017–18, vol. 2, 2018, pp. 1–27
2. Farmers Suicide Trends in India, <https://thewire.in/wp-content/uploads/2017/04/Annex-1-%E2%80%93-Farmer-suicides-2016-and-2017.pdf>. Accessed 2018/8/15
3. A. Addae-Korankye, Causes and control of loan default/delinquency in microfinance institutions in Ghana. *Am. Int. J. Contemp. Res.* **4**(12), 36–45 (2014)
4. The Indian Express, <https://indianexpress.com/article/india/in-80-farmer-suicides-due-to-debt-loans-from-banks-not-moneylenders-4462930/>. Accessed 2018/8/15
5. K. Taylor-Sakyi, Big data: understanding big data, 2016
6. F. Duraõ, A systematic review on cloud computing. *J. Super Comput.* **68**(3), 1321–1346 (2014)
7. Z. Zheng, An overview of blockchain technology: architecture, consensus, and future trends, in *IEEE International Congress on Big Data*, 2017, pp. 557–564
8. R. Lewis, J. McPartland, R. Ranjan, Blockchain and financial market innovation. *Economic Perspectives*, 2017, pp. 1–15
9. Blockchain for Financial Leaders: opportunity vs. reality. Financial Executives International, 2018. [www.financialexecutives.org](http://www.financialexecutives.org)
10. V. Rawal, P. Mascarenhas, Blockchain for healthcare. Citius Tech, White Paper, 2018, pp. 1–12
11. P. Brody, How Blockchain is revolutionizing supply chain management. *Digitalist Magazine*, 2017, pp. 1–7
12. J. Michael Graglia, C. Mellon, Blockchain and property in 2018: at the end of the beginning, in *World Bank Conference on Land and Poverty*, Washington, DC, 2018
13. Agmarknet, <http://agmarknet.gov.in/PriceTrends/Default.aspx>. Accessed 2018/6/9
14. Horticulture, <http://nhb.gov.in/OnlineClient/MonthwiseAnnualPriceandArrivalReport.aspx>. Accessed 2018/5/12
15. Government of Karnataka, Final Estimates of Area, Production and Yield of Principal Crops in Karnataka for the Year 2012–13, 2015
16. Government of Karnataka, Household Consumer Expenditure in [State Sample] NSS 64th Round (July 2007–June 2008). Directorate of Economics and Statistics, No. 4, 2012
17. India National Sample Survey Office, Household Consumption of Various Goods and Services in India. NSS 66th Round July 2009 to June 2010, 2010

# Chapter 9

## An SDN-Based Strategy for Reliable Data Transmission in Mobile Wireless Sensor Networks



V. Shubha Rao and M. Dakshayini

### 9.1 Introduction

In today's world, wireless sensor networks (WSN) have been gaining popularity because of its applications in an environment where humans cannot intervene [1]. It is a network consisting of tiny devices called sensor nodes connected for a purpose. These sensor nodes are equipped with a sensor, a microcontroller, a transceiver, and ADC. These nodes are limited in resources. Thus, WSNs need low track communication techniques that might use minimum resources and achieves the quality of service. The limited resources of WSN poses many strategic issues, energy efficacy is one important consideration. These networks are planned to work in an atmosphere with inadequate bandwidth and ability to communicate. The base station is also responsible to communicate with the traditional networks. The network is equipped with a base station with more computational power and more communication resources. The applications of WSNs are numerous and in areas mainly meant for monitoring the physical environment for various purposes. The monitored information if reached the appropriate destination and proper actions are taken then it would avoid or prevent damages caused if otherwise. The sensed data thus will aid the choice makers to take proper timely decisions by analyzing the received data. The timely transmission of critical data is very prominent for

---

V. Shubha Rao (✉)

Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, India

e-mail: [Shubha.ise@bmsce.ac.in](mailto:Shubha.ise@bmsce.ac.in)

M. Dakshayini

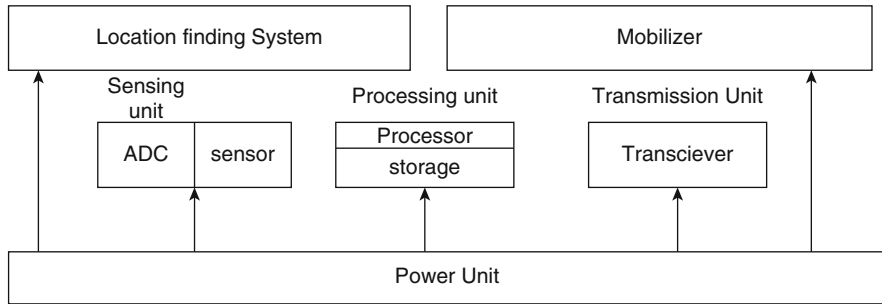
Department of Information Science and Engineering, BMS College of Engineering, Bangalore, India

e-mail: [dakshayini.ise@bmsce.ac.in](mailto:dakshayini.ise@bmsce.ac.in)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_9](https://doi.org/10.1007/978-3-030-19562-5_9)

87



**Fig. 9.1** Components of a mobile sensor node

human life-saving applications. Thus, the timely delivery of data without delay is the need of the hour. Formulating a system, which aids in timely decisions in the controlled environment of wireless sensor networks, is a challenging job. Thus, there is a requirement for designing a model that reduces delay, data loss, and energy consumption. Mobile wireless sensor networks (MWSN) could be a feasible solution for reducing the delay. There are various applications of MWSNs such as traffic management and invigilating disaster environment. They are defined as wireless sensor networks with mobility provided to the sensor node for monitoring the physical zone of interest. The components of a mobile sensor node as shown in Fig. 9.1 are sensor unit which has sensors, and Analog to Digital converters (ADC), processing unit comprising of a microcontroller and memory, transceiver unit along with some additional components to support mobility such as location finder unit that identifies the location of the sensor node, a mobilizer which enables mobility of the sensor node and a power unit that is in charge for generating power to fulfill the mobile node's energy requirements [2].

The ability and flexibility of WSN to reduce the delay can be enhanced by introducing mobility to some/all the nodes. The mobility of sensors that are mobile could be controlled to perform different tasks. The characteristics of MWSN are very similar to that of WSN. The major differences between the MWSN and static WSN are more dynamic in topology as related with WSN, the challenges exist in protocols of routing, MAC and other layers for MWSN. The highly unreliable communication channel, detecting the location of the mobile node, is also challenging. The advantages of MWSN are efficient energy consumption, better channel utilization, efficient network connectivity, and data reliability as a number of hops for transmission of packet reduces the error occurrence and increases the data quality and reduces the energy utilized for retransmissions that might occur because of errors. Thus, this paper proposes a novel model with mobility introduced for increasing the data reliability with lesser number of retransmissions that might have occurred because of errors also. Software Defined Networking [3, 4] is employed for managing network resources with two planes, data plane and controller plane. The control plane is centralized and is responsible for managing the resources based on the priority and reduces the delay, has the sensor node is

relieved with functionalities of resources management and controlled by the central controller energy consumption is reduced. Thus, increases the system performance. The rest of the paper is planned as follows: Sect. 9.2 provides a brief of the existing literature. Section 9.3 discusses the proposed model and methodology. Section 9.4 discusses the results and Sect. 9.5 concludes the paper.

## 9.2 Related Work

This section provides a review of a few of the existing work in the literature. In [5] authors have discussed combining lightweight coding and compressed sensing techniques to enhance the performance of the system for real-time data and reliability of resource managing in the mobile Internet of Things. The parameters like sample signal, signal and hops are set up by using compressed sensing scheme that is based on an adaptive frame format definition. The global or local network resource scheduling is managed by building nonlinear relationship matrixes among the resource information of sensors and Quality of Service. In [2] authors have presented a hierarchical multi-tiered architecture for mobile wireless sensor networks. They have discussed the effect of mobility on diverse performance metrics. They also have made a survey on applications of mobile WSN. In [6] authors have implemented an intelligent distributed network where sensors and robots cooperate to resolve cooperative jobs. This is achieved by communicating wireless among the fixed and mobile nodes. The experiment is carried out for assessing the viability of the proposed technique and also the trustworthiness of the system in a well-lit point in a room. In [7] authors have proposed an overview of plans that can be used in the assessment of mobile communication in WSN. Challenges that need to be addressed for mobile wireless sensor networks are also discussed by the authors.

In [8] authors have presented an assessment for mobile WSN by employing position cluster-based and non-position-based routing protocols based on numerous factors such as control overhead, packet delivery ratio, and network lifetime. In [9] authors have discussed various challenges addressed for mobile WSN in the existing literature. In [10] an on-demand multipath routing ad hoc protocol to discover numerous paths is employed to forward the data to the destination node. The authors have considered energy-efficient routing for static sink node and mobile nodes and, they have compared this with ad hoc on-demand distance vector routing protocol. In [11] authors have presented various characteristics, challenges, and application and various mobility models in introducing mobility in wireless sensor networks. In [12] a cluster head selection strategy has been proposed for efficient energy utilization that is analyzed and validated based on remaining energy and randomized selection of the node. It is correlated with the LEACH protocol. In [13] Javier and others have estimated lightweight component-based service platform for WSN. The software resources of WSN are managed by the proposed architecture that takes in to consideration an SLA specification. To dynamically

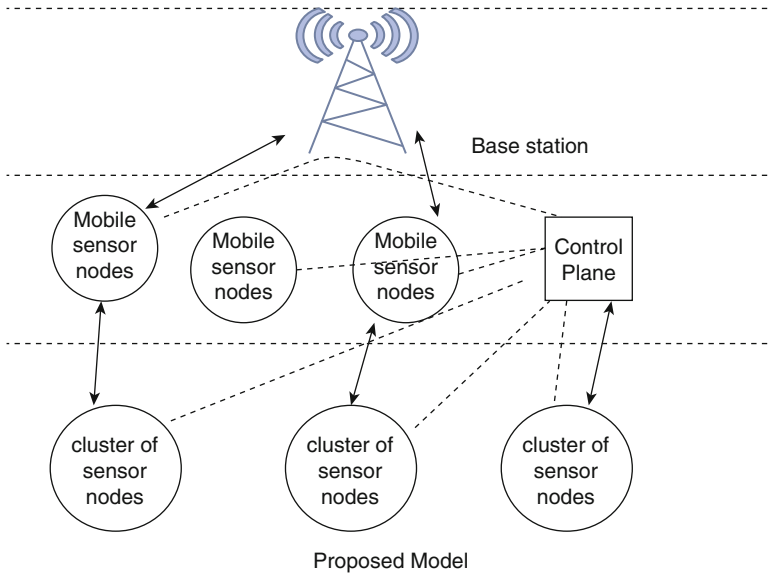
optimize the component behavior and the enactment of SLA, contextual information collected from the network is employed. In [14] Yu and his team have developed and evaluated the performance of a medium access protocol for transmission of multimedia data over wireless networks. In [15] Vinothini and Umamakeswari have proposed a novel approach for addressing reliability. The approach utilizes techniques to find connectivity delay and rebroadcasting delay for efficient packet transmission. In [16] authors have developed a protocol for conserving energy and to avoid congestion for multicast traffic in wireless sensor networks comprising mobile nodes. The protocol is using linear and binary feedback method to compute congestion in intra and inter cluster levels. In [17] Aikebaier and his fellow authors have discussed a redundant data transmission protocol to transmit sensed value to the actuator reliably and efficiently. The protocol eliminates the redundancy and forwards fewer number of values. In [18, 19] authors have provided with an analysis of root causes for unreliable transmission. Expressions related to packet reception rate have been derived as a function of distance using environmental and radio parameters like path loss exponent and shadowing variance of the channel. In [20] authors have proposed a reliable data transmission scheme which is distributed in nature based on a plug-able modular approach using neighbor nodes. This method decreases the amount of retransmissions by transferring the retransmission job from the sensor node to the neighbor node which has a better link quality.

## 9.3 Proposed System

### 9.3.1 System Model

The system considered for the research work is shown in Fig. 9.2. It is deployed in a three-layered architecture which is based on the SDN concept. The first layer consists of  $n$  static sensor nodes deployed in cluster form, the second layer contains  $m$  mobile sensor nodes and control plane (CP), and the next layer comprises a base station that is linked to the internet through the gateway.

The static sensor nodes are embedded with local controller responsible for the packet from the sensed data and forwards the data are deployed in a clustered fashion with a designated cluster head (CH). The CH is in charge of eliminating the redundant data, prioritizing the data and forwarding the data utilizing the resources allocated by the control plane (CP) towards the next layer. The next layer consists of  $m$  mobile sensor nodes and CP which depicts the control plane. The entire control logic and management of resources of the entire network is handled by the control plane. It consists of modules to check the availability of the resources for that priority data and if so allocates the resources and sends this control information to CH and through CH to the sensor node to utilize the reserved resources and communicate the data towards the nearest mobile sensor node. The mobile sensor nodes are responsible to obtain and forward the data received by the cluster heads



**Fig. 9.2** Proposed system

utilizing the resources according to the instructions of the CP to the base station with mobility. The third layer contains a base station that is linked to the internet through a gateway. It forwards the data to the intended users through the internet.

### 9.3.2 Methodology

The proposed system works in three phases: (1) initialization phase, (2) priority, and control information generation phase, (3) and data forwarding phase.

#### 9.3.2.1 Initialization Phase

In the first layer  $n$  static sensor nodes are deployed to form a cluster. CH is identified and in the next layer  $m$  mobile sensor nodes are deployed and CP is identified which is a static CH. All the CHs transmits the information about the resources available to CP which utilizes to allocate the resources depending upon the priority of the data. Updating of resources available at each CH is done periodically.

Sensors are deployed in random fashion and clusters are formed. Each sensor node is equipped with LPM and CPM is identified at each cluster. One of the CPMs is designated with MPM. All the CPMs transmits the amount of available resources to MPM which stores it in  $r$ -table of the MPM. Updating of resources available at



each CPM in done periodically. Also, the handshaking of priority data between the CH and CP is also exchanged.

Input: sensors

Output: network formed, and resources are collected at CP.

1. n static sensor nodes are deployed.
2. C clusters are formed
3. M mobile sensor nodes are deployed
4. For(i=1 to C)
  - Transmit the resources from CH to CP
  - End for
5. Repeat step 4 periodically.

**9.3.2.2 Priority and Control Information Generation Phase**

In this phase the priority is assigned based on the type of information (1) real-time data, (2) non-real time, and (3) local data. This information is transmitted to CP. The CP now generates the control information about the amount of resources to be allocated to data flow depending upon the priority. This information is transmitted to CH and to the local controller and to the mobile sensor nodes before the data forwarding phase starts.

1. Depending on the type of data assign the priority
  2. Transmit the information to the CP and CH
  3. CP decides the resources depending upon the priority
  4. Transmit the control information to the CHS
1. Refresh the resources availability
  2. If(requested resources are available in the path)
    - Then
      - If(priority-received ==priority obtained from updater) && requested resources)
        - Reserve the resources for the data flow
        - Transmit the control information to the CH
- Endif

**9.3.2.3 Data Forwarding Phase**

The sensor nodes sense the data periodically and form packets then forwards it to the CH. Redundant sensed data is eliminated and based on the type of data priority is assigned and based on the control information received from the CP CH stored in the suitable queues Q1 representing the highest priority. And forwards the data utilizing the resources reserved for that particular data priority towards the next layer which consists of mobile sensor nodes.

1. Packets are formed
2. Eliminate the redundant data and store it different priority queues(Q1, Q2, Q3) based on priority information.
3. For(i=1 to C)
  - a. Transmit high priority data from CH to MS utilizing the resources allocated to that particular data
  - b. End for

#### At Mobile Sensor Nodes

The mobile sensor nodes utilize the resources allocated for that data and carry towards the base station by moving towards the base station and reducing the number of hops, which in turn reduces the retransmission rates and thus achieves reliability. Since the functional burdens of the sensor nodes, CH, and mobile sensor nodes are reduced energy consumption in those nodes are also reduced. Also, higher priority data is reserved for more resources and thus reduces the delay in transmission of these packets.

1. Carries and transmits the packets towards the base station.

*Performance Metrics:* The proposed system is evaluated for its performance using the packet delivery function and delay for an end-to-end delivery and energy efficacy metrics. The packet transmission ratio is increased when related with traditional networks since the mobile sensor node lessens the number of hops for transmission of the data packets and which in turn reduces the packet retransmissions. End-to-end delay is reduced as resources are reserved for the high priority data and it reaches the destination with in time. Consumption of energy for each sensor node is also reduced as the resource management of all sensor nodes are centralized at the control plane and all sensor nodes are relieved from this burden.

## 9.4 Results and Discussions

The proposed framework was simulated, and the performance metric reliability is evaluated by conducting several simulations. A single base station and five clusters over an area of  $1500 \times 1500 \text{ m}^2$  are considered. Each cluster consists of 50–60 sensor nodes deployed and in the middle layer 3–4 mobile sensor nodes are assumed. There are three queues Q1, Q2, and Q3 at CH where Q1 is the highest priority queue. The proposed model aims at achieving better reliability as the mobile sensor node reduces the number of hops for transmission which in turn reduces the errors in transmission and thus increases the reliability. The sensed data reaches the base station by getting forwarded from the intermediate node until it reaches the base station by visiting the cluster head and utilizing the resources allocated by the CP

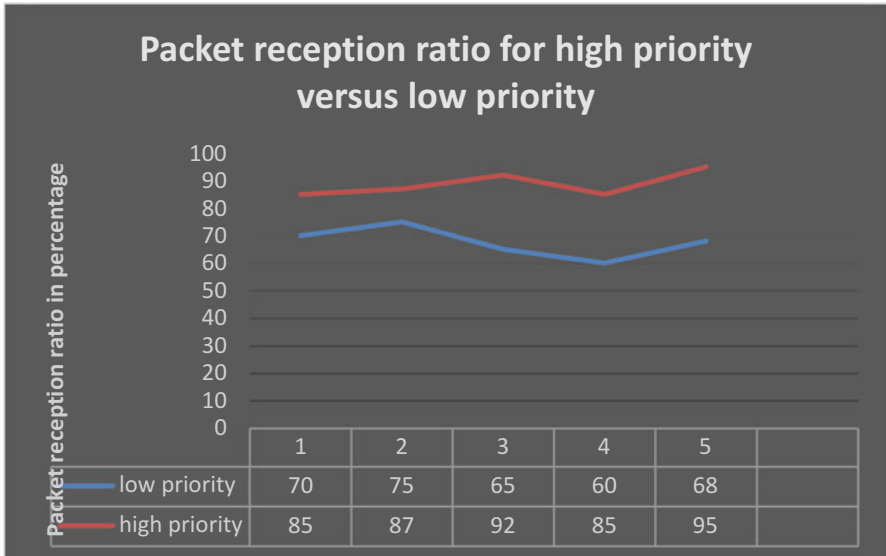


Fig. 9.3 Packet reception ratio for high priority versus low priority

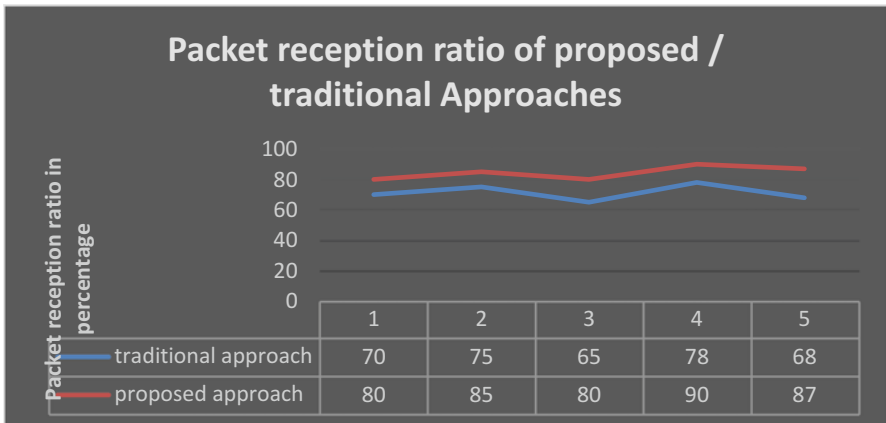


Fig. 9.4 Packet reception ratio of proposed/traditional approaches

and carried upon by the nearest mobile sensor node. Figure 9.3 shows that the packet reception ratio at the base station is more for high priority data when compared with packet reception ratio of low priority data. Figure 9.4 shows the packet reception ratio of the proposed approach is more when compared with packet reception ratio when using other traditional approaches. The proposed model increases the reliability of high priority data because of mobile sensor node which transmits data even channel link is weak. Thereby the overall reliability of the system is increased.

## 9.5 Conclusion

The resource-constrained nature of wireless sensor networks makes it less reliable when compared with traditional networks. However, they have numerous applications particularly in the risky scenarios data sensed is very critical and important which demands high reliability and with minimal delay. The proposed work employed mobile sensor nodes and SDN approach for reducing the number of hops and thus increased the probability of successful transmissions. The results of simulations have proven that the proposed approach outperforms with respect to reliability when compared with traditional approaches.

## References

1. M.R. Ahmed et al., Wireless sensor network: Characteristics and architectures. World Acad. Sci. Eng. Technol. Int. J. Electr. Comput. Energ. Electron. Commun. Eng. **6**(12), 1398–1401 (2012)
2. S.A. Munir, B. Ren, W. Jiao, B. Wang, D. Xie, J. Ma, Mobile wireless sensor network: architecture and enabling technologies for ubiquitous computing, in *Proc. 21st Int. Conf. Adv. Inf. Netw. Appl. Work. (AINAW'07)*, vol. 1, 2007, pp. 113–120
3. M. Jacobsson, C. Orfanidis, Using software-defined networking principles for wireless sensor networks, in *Proc. 11th Swedish Natl. Comput. Netw. Work. (SNCNW 2015)*, Karlstad, May 28–29, 2015, pp. 1–5
4. R. Article, M.A. Hassan, Q. Vien, M. Aiash, Software defined networking for wireless sensor networks: a survey. *Adv. Wirel. Commun. Netw.* **3**(2), 10–22 (2017)
5. Z. Jianming, L. Fan, L. Qiuyuan, Resource management technique based on lightweight and compressed sensing for mobile internet of things. *J. Sensors* **2014**, 690521 (2014)
6. F. Viani, M. Donelli, G. Oliveri, A. Massa, A mobile wireless sensor network architecture for collaborative tasks achievement by means of autonomous robot swarm, in *2010 IEEE Antennas Propag. Soc. Int. Symp. (APSURSI)*, 2010, pp. 1–4
7. J. Rezazadeh, M. Moradi, A.S. Ismail, Mobile wireless sensor networks overview. *Int. J. Comput. Commun. Networks* **2**(1), 17–22 (2012)
8. S. Parvin, M.S. Rahim, Routing protocols for wireless sensor networks: a comparative study. *ICECC* **1**, 891–894 (2008)
9. Y. Sun, S. Zhang, H. Xu, S. Lin, New technologies and research trends for mobile wireless sensor networks. *Int. J. Distrib. Sens. Networks* **2014**, 929121 (2014)
10. T. Hu, Y. Fei, QELAR: a machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks. *IEEE Trans. Mob. Comput.* **9**(6), 796–809 (2010)
11. M. Bouaziz, A. Rachedi, A survey on mobility management protocols in wireless sensor networks based on 6LoWPAN technology. *Comput. Commun.* **74**, 3–15 (2014)
12. R.U. Anitha, P. Kamalakkannan, Energy-efficient cluster head selection algorithm in mobile wireless sensor networks. *Int. Conf. Comput. Commun. Informatics* **9**(6), 1–5 (2013)
13. D. Cid, P. Javier, et al., DARMA: Adaptable service and resource management for wireless sensor networks, in *Proceedings of the 4th International Workshop on Middleware Tools, Services and Run-Time Support for Sensor Networks*, (ACM, 2009)
14. H. Yu, A multiple access protocol for multimedia transmission over wireless networks. arXiv preprint arXiv:1205.4959 (2012)

15. M. Vinothini, A. Umamakeswari, Reliable data transmission using efficient neighbor coverage routing protocol in wireless sensor network. *Indian J. Sci. Technol.* **7**(12), 2118–2123 (2014)
16. R. Beulah Jayakumari, V. Jawahar Senthilkumar, Priority based congestion control dynamic clustering protocol in mobile wireless sensor networks. *Sci. World J.* **2015** (2015)
17. K. Morita, A. Aikebaier, T. Enokido, M. Takizawa, Evaluation of the reliable data transmission protocol in wireless sensor-actuator networks, in *Int. Conf. Complex, Intell. Softw. Intensive Syst. 2008 (CISIS 2008)*, vol. 4, 2008, pp. 12–18
18. M.Z. Zamalloa, B. Krishnamachari, An analysis of unreliability and asymmetry in low-power wireless links. *ACM Trans. Sens. Networks* **3**(2), 7 (2007)
19. M. Zuniga, B. Krishnamachari, Analyzing the transitional region in low power wireless links, in *2004 First Annual IEEE Commun. Soc. Conf. Sens. Ad Hoc Commun. Networks, 2004 (IEEE SECON 2004)*, 2004, pp. 517–526
20. J. Seo, M. Kim, I. Hur, W. Choi, H. Choo, DRDT: distributed and reliable data transmission with cooperative nodes for lossy wireless sensor networks. *Sensors* **10**(4), 2793–2811 (2010)

# Chapter 10

## Different Aspects of 5G Wireless Network: An Overview



Akash R. Kathavate, Bhanu Priya, Rajeshwari Hegde, and Sharath Kumar

### 10.1 Introduction

The remarkable development of cell phones in the recent years has encouraged the researchers to move towards developing 5G technology which is aimed to overcome drawbacks of 4G [1]. There is a dire requirement for wireless technology for wide variety of applications such as IoT (Internet of Things), infotainment systems, and security in vehicles, to name a few [2]. 1G was first afloat in 1979 by Nippon Telegraph and Telephone (NTT) and employed the use of analog signals for data transmission which led to many problems such as data encryption and security [3]. 1G technology provided seamless mobile connectivity introducing voice services. To overcome the challenges of 1G, 2G network Concept was launched in 1991 by Radiolinja in Finland. It had a Data bandwidth of 64 kbps and used TDMA (Time Division Multiple Access) multiplexing. It worked on Circuit switching and PSTN (Public Switched Telephone Network) core network. There was a remarkable improvement in the quality of phone calls and increased the voice capacity [4]. Due to low bandwidth of 2G, the constraint in communication led to concept of packet switching which was used in 3G. Thus 3G was launched as a pre-commercial network in 1998 by NTT Docomo in Japan. It has Data bandwidth of 2 Mbps and uses CDMA (Code Division Multiple Access) multiplexing in the core network. But the spectrum and latency being on the lower side, 4G concept was launched in 2015. It uses Data bandwidth of 1 Gbps and has CDMA multiplexing done in this standard. It works on Packet switching giving mobile ultra-broadband access. Under this

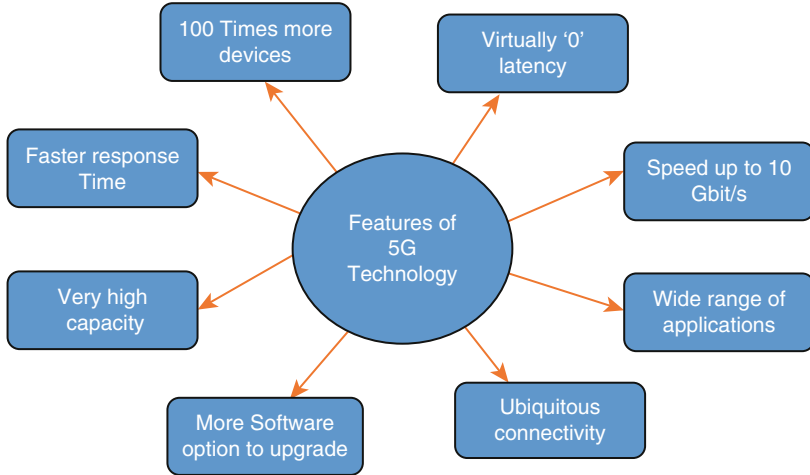
---

A. R. Kathavate · B. Priya · R. Hegde (✉)  
Department of Telecommunication Engineering, BMS College of Engineering, Bangalore, India  
e-mail: [IBM15TE001@bmsce.ac.in](mailto:IBM15TE001@bmsce.ac.in); [IBM15TE007@bmsce.ac.in](mailto:IBM15TE007@bmsce.ac.in)

S. Kumar  
Reliance JIO, Bangalore, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_10](https://doi.org/10.1007/978-3-030-19562-5_10)

97



**Fig. 10.1** Features of 5G

standard Traditional voice calls were replaced by IP telephony. To increase the data speed and bandwidth usage, 5G was introduced in 2012 at Mobile World Congress. It will have a Data bandwidth of more than 1 Gbps where CDMA multiplexing is done. It works on the concept of packet switching. The 5G mobile network uses OFDM technique and comprises wireless systems which are packet switched and has area coverage which is vast. 5G ranges in frequency from 30 to 300 GHz with high throughput in millimeter, which enables a data speed of 20 Mbps up to 2 km [5]. Wireless World Wide Web (WWW) applications can be provided by these specifications of 5G [6].

We would like to give different aspects of 5G network in this chapter and discuss on it. This chapter is structured in the following manner. Section 10.2 deals with the related work. Section 10.3 gives an overview of 5G. Section 10.4 deals with the performance parameters of 5G and its evaluation. Section 10.5 deals with implementation of 5G. Section 10.6 presents the challenges in deploying 5G. The chapter is concluded in Sect. 10.7 (Fig. 10.1).

## 10.2 Related Work

In April 2008, NASA started to develop 5G technology under the supervision of G. Brown and was implemented by Machine-to-Machine Intelligence (M2Mi) Corp [7]. European Union (EU) began exploring the possibilities and options related to 5G technology by launching eight such projects. 5G presented an overview of wireless future communication and €50m was granted for research in this field by EU for 5G deployment by 2020 [8]. The 5G project ought to be a smart and

economical wireless infrastructure which shall use small and light antennas which makes use of directional beam forming that can bounce back signals off buildings using the high-frequency spectrum [9]. EU started another group known as METIS which started on 1 November 2012, in regard of 5G. METIS has proposed these 5G schemes that put forth the challenges in future and shall act as a reference guiding future work. An augmented number of connected devices are handled proficiently along with competent user experience, extended life of battery, less latency and authenticity.

Thus, METIS had a crucial role before global standardization, of awakening consensus for major external stakeholders [10]. The iJOIN EU project was launched in November 2012 which mainly focused on “small cell” technology, which makes use of radio wave spectrum and related limited and planned resources [11]. In the year 2013, Samsung Electronics announced their plan of bringing in a 5G wireless technology. But during testing, the transfer speed was 1.056 Gbit/s for the sent data of 5G network [12]. NTT DoCoMo merged with the following companies and the institutes to come up with the following results in the field of 5G technology.

### ***10.2.1 Ericsson***

The area of technology which the trails were related on was architecture of “small cell” which comprised of network which was variegated, 15 GHz frequency bands which included high-speed and high-capacity transmission for the broadband communication. 5G system would be an amalgamation of associated Radio Access technologies which includes some LTE versions [13, 14].

### ***10.2.2 Nokia***

NTT Docomo and Nokia consented to cooperate on 5G technologies research and work together on a Concept system of 5G Proof. These two companies continued to work together on the future of radio access systems and to research on potentials of the technology of wave in millimeters at very-high-frequency spectrum band [5].

### ***10.2.3 NEC***

Its main agenda was to testify large number of antennas for “small cells” with amplified time-domain beam-forming technologies. Thus, this technology was expected to enhance MIMO technology. MIMO is expected to support mobile data coverage for many users at once and mitigating interference while enabling 5G features such as high speed, better communication, and capacity [5].



### ***10.2.4 Tokyo Institute of Technology***

NTT Docomo and TIT worked on a joint experiment and invariably achieved a packet transmission uplink rate of 10 Gbps which is almost 1000 times the rate of LTE prevalent today. A spectrum of 11 GHz and bandwidth of 400 MHz was relayed by a mobile station. Multiplexing various streams of data using 8 transmitting and 16 receiving antennas of similar frequency used MIMO technology [15].

### ***10.2.5 Alcatel-Lucent***

Under their perspective, 5G telecom networks shall cater to as per user requirements to build the network the user desires. An advanced, flexible network infrastructure uses interface as air which benefits from both virtual network and networking outlined by software [5].

### ***10.2.6 Fujitsu and DOCOMO***

Both built a collaboration for the 5G realization. With the help of experimental practices with DOCOMO, they aspire to testify 5G and thus further endow to society by bringing in further enhancements of IoT and Big Data [16].

### ***10.2.7 Samsung***

It designed the world's very first technology using flexible transceiver arrays which operates in Ka bands in the millimeter-wave, it has 28 GHz frequency, ranging till 1.056 Gbps of speed and extending up to 2 km. The technology using flexible transceiver arrays uses around 64 antenna elements, which is used to centralize radio energy in close to surmount the weaker propagated characteristics of millimeter bands, directional bands are used [16].

### ***10.2.8 The Federal Communications Commission***

On 14 July 2016, FCC initiated to start using advanced bandwidth in the high-band spectrum which is underutilized and can be used for 5G wireless communications. The Snapdragon X50, the first 5G modem by Qualcomm, was announced on 17 October 2016 as the first commercial 5G mobile chipset. The first ever 5G

deployment was done on 9 February 2018 at Winter Olympics in South Korea. European Union law makers on 2 March 2018 proposed onto a deal of bringing in the 3.6 and 26 GHz bandwidths by 2020 to make adjustments for 5G. Other countries where 5G deployment happened are Australia (by Telstra), Bangladesh (by Huawei), Finland and Estonia (by Elisa), Indonesia (XL Axiata with Nokia), Norway (by Telenor), Philippines (by Global Telecom), and Qatar (by Ooredoo) [17].

### 10.3 What is 5G?

With the idea of upgrading the present telecommunication standards, 5G concept was proposed to provide large broadcasting of data along with significant improvements in performance parameters. It ought to be a packet switched wireless system to support Virtual Private Network (VPN). It uses CDMA as well as BDMA (Beam Division Multiple Access). The data speed and capacity is expected to be higher than 4G. It aims at providing ubiquitous connectivity, more software options to upgrade, and wide range of applications. Appropriate QoS (Quality of Service) is provided to the people according to their requirements. The main goal of QoS is to give priority to networks with less latency, a checked jitter, and dedicated bandwidth. There is presently no standard for 5G deployment; however international agencies like IEEE, IET, ITU, and FCC are working on the standardization of 5G. The International Telecommunications Union (ITU) has lately started researches which outline stipulations for International Mobile Telecommunications (IMT) 2020.

### 10.4 Performance Parameters and its Evaluation

Several parameters are considered for inspecting and monitoring quality and performance of the networks.

#### 10.4.1 *Network Performance Parameters*

*Network performance* refers to analysis and review of certain attributes in a collective network which help in advancing the service quality. The performance level of a given network can be measured using it as a qualitative and quantitative process. We experienced a varied change in parameters from 4G to 5G which can be stated as under (Table 10.1).

**Table 10.1** Changes in network parameters from 4G to 5G

Parameters	4G LTE	5G
Data rates	500 Mbps in 4G	1–10 Gbps in 5G
Capacity	100 s GB/user	36 TB/user
Latency	About 10 ms	About 1 ms
Frequency bands	700–2100 MHz	28–40 GHz
Spectral efficiency (DL)	15 bps/Hz	30 bps/Hz

### 10.4.2 QoS (Quality of Service) Parameters

QoS Parameters are used to obtain the overall performance of a network a user primarily observes. The parameters being Bit rate, IP packet loss, transmission delay, throughput, and availability.

*IP Packet loss:* ITU-T recommends that the main stiff QoS packet-loss purpose should be less than  $1 \times 10^{-5}$  for a large end-to-end QoE [18].

*Network Availability:* Network Availability can determine the total free time of a network which includes the network peripherals such as routers, multiplexers, and switches.

*Contention Expected ratio:* Around 50:1.

### 10.4.3 Evaluation of Performance Parameters

Evaluation of any technology plays a crucial role before implementing and maintaining the standard. Henceforth the performance parameters of 5G can be analyzed using the following simulation tools:

*WiSE* (Wireless Simulator Evolution): It is a dynamic system-level simulator used in evaluation of 4G/LTE with beam-formed channel state information-reference signal (CSI-RS) transmission, Class A precoder for 32 antenna ports and advanced CSI feedback. It has been validated with the Third Generation Partnership Project (3GPP) calibration campaigns [19].

*NS3 Network Simulator:* It is an open-source network simulator written in C++ and python. The mmWave model is used for 5G network simulation using Evolved Packet Core (EPC). It is written in C++ and provides support for TDD and OFDM.

*Opnet Simulator:* This simulation tool analyzes the behavior and performance of any given network. It is an event-driven simulator which uses LTE-A model along with IEEE 802.15c standard.

## 10.5 Implementation of 5G

With an aim to surpass the challenges and successfully deploy the 5G wireless system, certain design notions need to be followed by the 5G architecture. They can be stated as under.

### 10.5.1 *Massive MIMO*

It is a sub-six GHz physical layer technology designed for wireless access in future. It uses a large array of antenna elements at base station to serve numerous sovereign terminals at the same time. The benefits of massive MIMO can be stated as excellent spectral efficiency and superior energy efficiency. The main characteristics are [20]:

1. Absolute digital processing with antennas having their own RF.
2. Computationally inexpensive decoding algorithms.
3. Array gain which results in closed loop link budget enhancement.

### 10.5.2 *Ultradense Networks (UDN)*

UDN has high density of radio resources when compared to current networks. In this network the base station density possibly reaches the user density and the inter site distance is only a few meters. It ought to increment the capacity, competence of radio links energy and obtain an enhanced victimization of spectrum. UDNs can use the prevalent benefits given out by direct transmissions, and density of large nodes brings out novel challenges. Here interference in an UDN which becomes more severe as volatility increases. There may be a large number of strong interferers. Hence the future of 5G wireless network is invariably an ultradense cellular network [20].

### 10.5.3 *Moving Networks (MN)*

By moving network, we mean a moving entity, possibly with high or average speed and carrying a few or hundreds of passengers. Thus, it will be required in 5G that moving network users under high mobility can communicate subject to specific Quality of Experience (QoE) constraints, i.e., the communication experience should be similar to non-moving cases. Consequently, more innovative services for moving network users can be realized [21].

### 10.5.4 D-2-D Communications

Packets of data are exchanged locally in between the devices which makes use of proximity based services. The potential gains involved in D2D cellular communication are:

1. Capacity gain: Practical sharing of spectrum resources between cellular and D2D users leads to this capacity gain.
2. Latency gain: Direct communication between devices without the presence of intermediate infrastructure decreases the inherent latency.
3. User data rate gain: With increase in proximity and convenient propagation conditions with high peak rates, user data rate gain increases [22] (Fig. 10.2).

### 10.5.5 URC (Ultrareliable Communication)

URC is a communication service which has certain level of enabling a tremendous degree of availability and reliability. Some of the applications of URC are industrial automation connectivity, reliable connectivity in the cloud, and reliable vehicle coordination through wireless communication [23].

### 10.5.6 Massive Machine Communications (MMC)

It aims at providing measurable connectivity solutions intended for a vast number of network-enabled devices. MMC communication will be connecting more than billions of IP-based devices via 5G wireless network. The concept of MMC comprises a group of radio ICT and approaches which includes access technologies

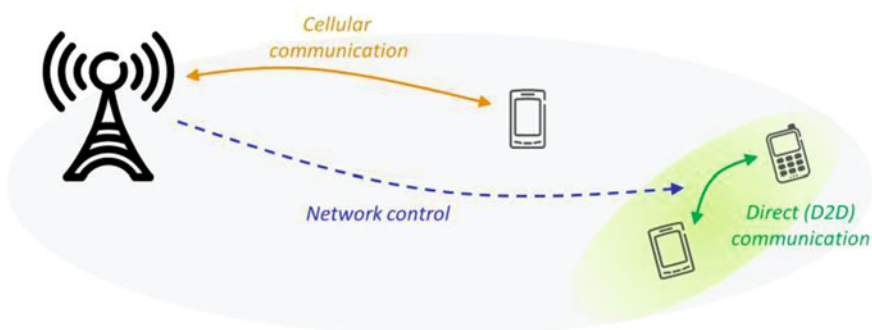


Fig. 10.2 Working of device to device (D2D) communication

such as direct access through a junction, access through an aggregated point, and fiber-optic communication between devices [24–26].

## 10.6 Challenges in Lieu of 5G Deployment

Though the 5G technology standard seems to be a massive leap in the field of telecom industry, there are certain challenges related to the deployment of this standard. A faction known as RAN (Radio Access Network) Research furnished a report in the year 2015 that coherently predicted 5G implementation to be slow. They stated that until the year 2030, 4G will dominate the wireless technology market. RAN Research even quoted few reasons for the same:

1. Till date 4G hasn't been developed entirely and the 3GPP which is a union of standards and implementation proposes to continue developing 4G till 2020.
2. In the USA, the deployed 4G still doesn't entirely meet the standards of 4G. The wireless companies should continue to make sound advancements to upgrade the existent 4G and thus improve the cellular telecom technology according to RAN Research [27].

### 10.6.1 Radiation Hazards

As cited in the Los Angeles Times, there is a potential increase in the radiation due to the up shoot of the number of transmitters and receivers and an array of new Internet-enabled devices is deployed to bring in new evolving telecom standards. The harmful effects of radio frequency radiation have left a grim persistence in mobile technology [27, 28].

Studies show that the biological effects which are caused due to the exposure to the radiations of RF are disturbance in cell metabolism, decrease in melatonin, and breakage of DNA strands.

### 10.6.2 Bandwidth Utilization

5G shall use both low (lower than 1 GHz) and high frequency (between 1 and 6 GHz) and frequencies greater than 6 GHz, referred to as “millimeter wave” frequencies. The 5G spectrum guarantees extended coverage due to the presence of low frequencies, very less power consumption, and high speed owing to the large channels in VHF bands. Thus, the diversity in the bands will be useful in meeting up every aspect of 5G and provide a harmonized global framework [19].

### ***10.6.3 Efficient Medium Access Control***

In a network, which consists of a large amount of access nodes and terminals of user, the user throughput will inherently be low, there will be an increased latency, and the number of hotspots in cellular technology won't be enough to cater high throughput. So there is a need for extensive research to optimize this efficient medium access control technology [29].

### ***10.6.4 Traffic Management***

Due to the presence of a large amount of Machine to Machine (M2M) devices a cell contains, there will be a serious system challenge which will give rise to overload and congestion, when compared to inherent human to human traffic in a wireless telecom network [29].

### ***10.6.5 Communication, Navigation, and Sensing***

5G technology in spite of having a strong computational power in order to employ the tremendous volume of data coming from various sources requires large infrastructure support [29].

### ***10.6.6 Security and Privacy***

Encryption and protection of personal data is by default the most important challenge 5G shall face. Ambiguity related to privacy, cyber security, and security threats has to be clearly defined by the 5G standards [29].

### ***10.6.7 Legislation of Cyber Law***

With the increase in data speed in 5G technology, it may lead to an increase in Cybercrime and other online fraud. Hence a proper cyber law should be drafted which would legislate such online crimes and reduce the effect of cybercrimes in critical agencies like government and political, which is a national and international issue [29].

## 10.7 Conclusion

In this chapter, we have brought into picture the aspects of various universities and industrial organizations to establish the standardizations of 5G. A number of conglomerates have executed considerable work in order to hasten the process of launching this standard. With the ever advancing and unpredictable future, we should anticipate an augmented pace in change of technology in spite of the presence of numerous hindrances and scope in development of 5G, with a great reliance on the consequences. But this 5G theory makes room for the challenges on which there is scope for further development.

**Acknowledgement** The work represented in this chapter is assisted by the college through the Technical Education Quality Improvement Programme [TEQIP-III] of the MHRD, Government of India.

## References

1. A. Kondur, M. Rao, B.S. Pavan Kumar, R. Hegde, Evolution of wireless mobile communication networks and future of cellular market in India, *Computer Science & Information Technology (CS & IT)*, 2012, pp. 453–462
2. What is 1G or First generation of wireless telecommunication technology? <http://www.clear doubts.com/technology/what-is-1g-or-first-generation-of-wireless-telecommunication-technology/>
3. M. Meraj ud in Mir, S. Kumar, Evolution of mobile wireless technology from 0G to 5G. *Int. J. Comput. Sci. Inform. Technol.* **6**, 2546–2551 (2015)
4. M. Sarfraz, Project report on 4G wireless. BETE-2nd (10588), 15 Apr 2014, [https://www.academia.edu/7032267/Project\\_Report\\_on\\_4G\\_Wireless](https://www.academia.edu/7032267/Project_Report_on_4G_Wireless)
5. E. AlMousa, F. AlShahwan, Performance enhancement in 5G mobile network processing. *Lect. Notes Inform. Theory* **3**(1), 19–24 (2015)
6. F.C. de Gouveia, T. Magedanz, Quality of service in telecommunication networks. *Telecommun. Syst. Technol. EOLSS* **2**, 1–8 (2009)
7. White paper: Current activity in 5G, <https://www.keysight.com/main/editorial.jsp?cc=IN&ckey=2311424&id=2311424&lc=eng>
8. NYU and NYU-Poly launch world's first academic research center to combine medicine with wireless communications and computing, <https://engineering.nyu.edu/news/nyu-and-nyu-poly-launch-worlds-first-academic-research-center-combine-medicine-wireless>
9. D. Gandla, Study of recent developments in 5G wireless technology. *Int. J. Electron. Commun. Eng. Technol.* **4**(5), 39–46 (2013)
10. Speech at mobile world congress: The road to 5G, [https://ec.europa.eu/commission/commissioners/content/speech-mobile-world-congress-road-5g-0\\_en](https://ec.europa.eu/commission/commissioners/content/speech-mobile-world-congress-road-5g-0_en)
11. General METIS presentations available for public, <https://metis2020.com/documents/presentations/index.html>
12. J. Atkinson, NTT DOCOMO to trial 5G technology with six suppliers [online], <http://www.wirelessmag.com/News/29288/ntt-docomo-to-trial-5g-technology-withsix-suppliers-.aspx#sthash.e8eO3kix.dpuf>
13. Ericsson white paper, 5G radio access, <https://www.ericsson.com/en/networks/trending/hot-topics/5g-radio-access>



14. DOCOMO to conduct 5G experimental trials with world-leading mobile technology vendors [online], 8 May 2014. [https://www.nttdocomo.co.jp/english/info/media\\_center/pr/2014/0508\\_00.html](https://www.nttdocomo.co.jp/english/info/media_center/pr/2014/0508_00.html)
15. P. Demestichas, A. Georgakopoulos, D.K.K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, J. Yao, 5G on the horizon: key challenges for the radio-access network. *IEEE Veh. Technol. Mag.* **8**(3), 47–53 (2013)
16. E.H. AlMousa, F. AlShahwan, R. Alhajri, The Deployment of the Future Mobile Network, vol. 5(2), Mar 2016
17. 5G Wikipedia page, [https://en.wikipedia.org/wiki/5G#cite\\_ref-23](https://en.wikipedia.org/wiki/5G#cite_ref-23)
18. H.-W. Chang, C.-L. Lai, K.-Y. Lin, H.-T. Chien, Moving networks for 5G communication systems. *J. Inform. Commun. Technol.* **16**2, 11–15 (2015)
19. C.-K. Jao, C.-Y. Wang, T.-Y. Yeh, C.-C. Tsai, L.-C. Lo, J.-H. Chen, W.-C. Pao, W.-H. Sheen, WiSE: a system-level simulator for 5G mobile networks. *IEEE Wirel. Commun.* **25**, 4–7 (2018)
20. E.G. Larsson, L. Van der Perre, Massive MIMO for 5G, *IEEE 5G Tech Focus*, vol. 1(1), Mar 2017
21. 5G Vision/The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services, European Commission, Heidelberg, 2015, EC/5G PPP
22. H. Shariatmadari, R. Ratasuk, S. Irabi, Machine-type communications: current status and future perspectives toward 5G systems. *IEEE Commun. Mag.* **53**(9), 10–17 (2015)
23. M.N. Tehrani, M. Uysal, H. Yanikomeroglu, Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions, vol. 52(5)
24. J.F. Monserrat, G. Mange, V. Braun, H. Tullberg, G. Zimmermann, O. Bulakci, METIS research advances towards the 5G mobile and wireless system definition. *EURASIP J. Wirel. Commun. Netw.* **53**(1), 20 (2015)
25. A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M.A. Uusitalo, B. Timus, M. Fallgren, Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *IEEE Commun. Mag.* **52**(5), 26–35 (2014)
26. J. Liu, W. Xiao, I. Chih-Lin, C. Yang, A. Soong, Ultra-dense networks in 5G, *IEEE 5G Tech Focus*, vol. 1(1), Mar 2017
27. G. Mascot, Introduction—5G paving the way to the fourth industrial revolution, Consultation paper, 2017
28. A. Kumar, M. Gupta, Key technologies and problems in deployment of 5G mobile communication system. *Commun. Appl. Electron.* **1**(3), 4–7 (2015)
29. 5G Radiation Dangers—11 Reasons To Be Concerned [online], <https://www.electricsense.com/12399/5g-radiation-dangers/>

# Chapter 11

## Intelligent Systems for Volumetric Feature Recognition from CAD Mesh Models



Vaibhav Hase, Yogesh Bhalerao, Saurabh Verma, and G. J. Vikhe

### 11.1 Introduction

Volumetric features are ubiquitous in mechanical engineering applications from design to manufacturing cycle. In many mechanical engineering parts, blends and holes constitute a significant percentage of features. Recognizing volumetric features in Computer Aided Design (CAD) mesh models are vital in applications such as mesh simplification, design, manufacturing, and finite element analysis.

Mesh models constructed from 3D scan data are called scan-derived mesh and those generated from B-rep models using CAD software are called CAD mesh models (CMM). The focus of this chapter is the CMM.

Segmentation aims to partition CMM into “meaningful” regions [1]. Each region can be fitted to a distinct, mathematically analyzable form [2]. Literature reveals the availability of many mesh segmentation algorithms. However, most of them are not suitable for CMM as scan-derived mesh is dense and streamlined whereas CAD mesh is sparse, non-uniform, and non-streamlined. Several mesh segmentation approaches in the literature have relied on information such as curvature or sharp edges. Huge time is needed for curvature computation. The curvature is sensitive to noise, variations in dimensions, and randomly disseminated triangulations [2]. It is

---

V. Hase (✉) · G. J. Vikhe

SPPU, Department of Mechanical Engineering, Amrutvahini College of Engineering, Sangamner, India

e-mail: [vaibhav.hase@avcoe.org](mailto:vaibhav.hase@avcoe.org)

Y. Bhalerao

Department of Mechanical Engineering, MIT Academy of Engineering, Pune, India

S. Verma

Centre for Computational Technologies (CCTech), Pune, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_11](https://doi.org/10.1007/978-3-030-19562-5_11)

109

difficult to establish one global threshold [3, 4] and so several mesh segmentation methods set local threshold while computing curvature.

The last three decades witnessed significant research work in extracting volumetric and free-form features. However, most feature recognition (FR) tools work on B-rep models while innovative design and manufacturing systems are mesh based [5, 6]. Therefore a need exists to develop FR from the mesh model. STL (Standard triangulated language) is globally accepted by all CAD/CAM system which makes it platform-independent data exchange format [7]. If we recognize features from STL model, it will be a unique data translator service [8, 9].

Above observations have inspired the research work reported in this chapter. The hybrid mesh segmentation approach is used for detecting volumetric features. The proposed algorithm segments the CMM into basic primitives like plane, cylinder, cone, sphere, or torus etc. After extraction of analytical surfaces, rule-based reasoning is used for FR. The innovation lies in the intersecting feature detection in which tedious curvature information and edge detection technique is not required. Further, the results are compared with existing and recent state-of-the-art approaches like Attene et al. [10], Schnabel et al. [11], Li et al. [12], Yan et al. [1], Adhikary and Gurumoorthy [13], and Le and Duan [14].

The main contributions of this research can be summarized as follows:

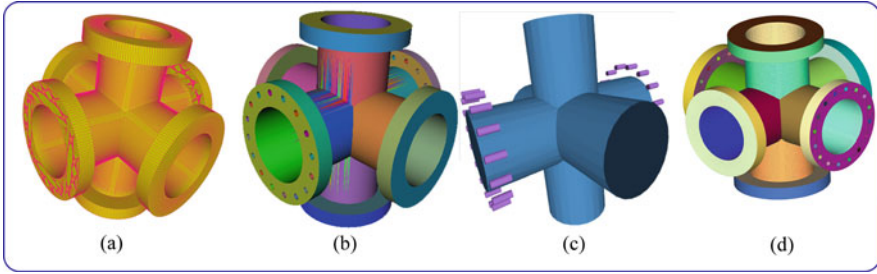
- Intelligent threshold prediction makes hybrid mesh segmentation automatic.
- Complex holes lying on multiple planer regions are detected and separated successfully.
- No curvature information is required for feature detection.
- Features are extracted without edge detection techniques.
- Partitioning criteria used for clustering triangles is “Facet Area.”
- Intersecting features are extracted automatically, and their parameters are also estimated accurately.

The rest of the chapter is structured as follows: Sect. 11.2 provides a comprehensive review of relevant literature; Sect. 11.3 illustrates a proposed methodology for the volumetric feature recognition. Section 11.4 deals with volumetric feature recognition. Discussion based on results is provided in Sect. 11.5. Section 11.6 present conclusion and future scope.

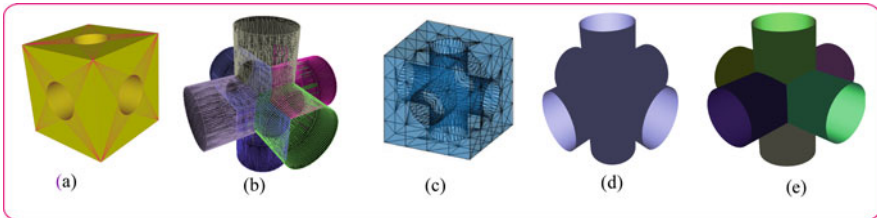
## 11.2 Literature Review

A comprehensive review of various FR approach with their strengths and weaknesses are reviewed in the literature [6, 9, 15–19]. The focus of the current research work is to compare the robustness and consistency of hybrid mesh segmentation algorithm with existing and recent state-of-the-art approaches; the literature review is limited to those approaches only.

Attene et al. [10] designed a Hierarchical Fitting Primitives (HFP) technique of mesh segmentation which needs a number of clusters as an input criterion along with



**Fig. 11.1** Failure cases for intersecting volumetric feature, (a) Input CAD mesh model. (b) Attene et al. [10]. (c) Muraleedharan et al. [20]. (d) Output



**Fig. 11.2** Failure cases for interacting features, (a) input CAD mesh model. (b) Attene et al. [10]. (c) Adhikary et al. [13]. (d) Muraleedharan et al. [20]. (e) Output

visual inspection to carry out segmentation. However, knowing a number of clusters before feature extraction is difficult. Figures 11.1b and 11.2b show the failure case of Attene et al. [10].

Schnabel et al. [11] designed ‘RANSAC’ (RANDOM Sample Consensus)-based framework for recognizing basic primitives. However, the approach either over-segments or under-segments the model. It results in inaccuracy of feature extraction. Li et al. [12] modified the approach of Schnabel et al. [11] and have developed the ‘GlobFit’ method. This approach is primitive fitting based rather than segmentation. They have used parallelism, orthogonality, and equal angle relations to extract primitives. This approach is computationally costlier and heavily depends on ‘RANSAC’ [11] output. Yan et al. [1] invented an algorithm for mesh segmentation of scanned or STL CAD model into non-overlapping patches by fitting quadric surfaces. Each patch was fitted to a general quadrics surface. Criteria used for segmentation was geometric distance based error function. However, the method is suitable for quadric surface only. It is not suitable to identify tori or blends.

Adhikary and Gurumorthy [13] presented an algorithm to recognize free-form volumetric features without segmentation from CMM. They used 2D slicing to identify feature boundaries. Features are identified by extracting feature boundary edges using 3D seed information of those 2D features. Region growing technique is used to find features using 3D seed vertex and feature boundary edges. The algorithm does not depend on mesh geometrical properties and mesh triangle

density. However, the algorithm is unable to detect and extract parameters of volumetric features for test case shown in Fig. 11.2a. Their algorithm depends on the choice of Minimum Feature Dimension (MFD) and must be known in advance before feature extraction. Figure 11.2c shows the failure case of Adhikary and Gurumoorthy [13].

Muraleedharan et al. [20] used a random cutting plane to extract the volumetric features. They blend graph traversal and Gauss map for FR. The algorithm is unable to separate the interacting features. Figure 11.1c shows the limitation of their approach. They used Gaussian curvature for boundary extraction and separating the interacting features. Their algorithm depends on a number of planes for features extraction which is assumed to be known. The feature must have the presence of inner rings which is the major limitation of the algorithm. If a feature does not have inner rings, it will not be detected. Figures 11.1c and 11.2d show examples of volumetric feature recognition but unable to separate into individual features. As feature joints have a complex boundary, segmentation is unable to separate them. However, the propose algorithm detects intersecting features along with geometric parameters.

Le and Duan [14] used uniform slicing along the major direction. They used a dimensional reduction technique which transforms 3D primitives to 2D in order to get a profile curve. The primitives are detected based on profile curve analysis. However, the algorithm is slice thickness dependent, and slicing techniques fail to detect or separate complex interacting features as noted by [13].

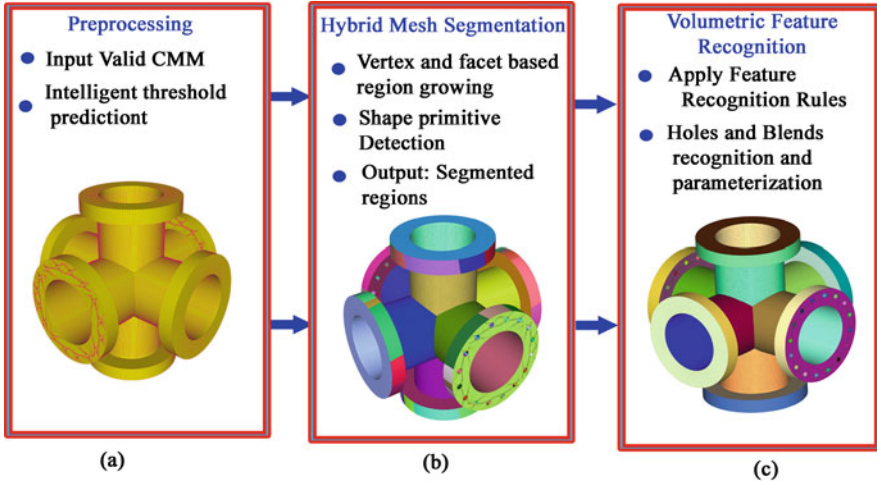
The proposed technique automatically extracts volumetric features like blends and holes along with their geometric parameters. With hybrid mesh segmentation, we can separate the interacting features as well. Figures 11.1d and 11.2e shows examples of volumetric feature recognition. Hybrid mesh segmentation recognized all the features whereas the closest one among others is the Le and Duan [14].

## 11.3 Methodology

The proposed algorithm involves three steps, viz. preprocessing, hybrid mesh segmentation, and volumetric feature recognition. Figure 11.3 illustrates the overall strategy to extract volumetric features from CMM which consists of the following steps:

### 11.3.1 Preprocessing

In preprocessing, topology is built in imported CAD mesh model, and automatic threshold prediction has been carried out.



**Fig. 11.3** The framework of the proposed methodology. (a) Preprocessing. (b) Hybrid mesh segmentation. (c) Volumetric feature recognition

### 11.3.1.1 Input CAD Mesh Model

In this research work, we assume a valid STL model as an input in ASCII (American Standard Code for Information Interchange) or Binary format which is free from errors, hence no need of model healing [9].

### 11.3.1.2 Automatic Threshold Prediction

The facets laying on the same surface have the same quality. We use the “Facet Area” property to segment the model. A significant step in segmentation is to set the appropriate Area Deviation Factor (threshold) at the beginning. It is a cumbersome task of identifying a threshold value for getting the expected results. Most of the time a trial-and-error approach is used to identify a correct threshold [20]. Inadequate threshold leads to over-segmentation (multiple small patches) or under-segmentation. Over-segmentation needs a post-processing merging step which increases processing time whereas under-segmentation leads to deficient results [21]. However, for a layman, setting the appropriate threshold is too complicated. Manual prediction is laborious and errors prone. Therefore, an automatic and intelligent prediction approach is of significance.

As stated above, Area Deviation Factor (ADF) is the decisive factor in segmentation quality. Intelligent prediction of threshold using the artificial neural network (ANN) and a machine learning classifier to partition CMM using hybrid mesh segmentation is proposed and implemented by Hase et al. [22]. A detailed description of automating threshold prediction is beyond the scope of this chapter.

### **11.3.2 Hybrid Mesh Segmentation**

The objective of hybrid mesh segmentation is to partition CMM into basic primitives like a plane, sphere, cylinder, cone and torus. It is difficult to segment CMM by using facet-based region growing or vertex-based region growing alone. Vertex-based region growing technique is used to detect curved surface whereas facet-based growing technique is used to detect curved features and planes. None of these techniques on their own gives a robust solution to recognize feature from CMM, a promising approach wherein intelligent blending of facet-based, vertex-based, rule-based reasoning are combined.

Hybrid mesh segmentation uses the “Facet Area” property to group facets together, using a combination of vertex-based and facet-based region growing algorithms [23]. It uses region growing algorithms to cluster facets into groups. After segmentation, shape primitives detection has been carried out wherein each facet group is subjected to several conformal tests to identify the type of analytical surfaces such as a cylinder, cone, sphere, or tori. After extraction of analytical surfaces, feature boundaries are identified.

#### **11.3.2.1 Iterative Region Merging**

The hybrid mesh segmentation leads to over-segmentation. The over-segmented regions are needed to be merged again to generate the single region. The proposed iterative region merging technique is based on predefined merging criteria. It repeatedly merges the regions that have similar geometric property. Following steps has been carried out in iterative region merging.

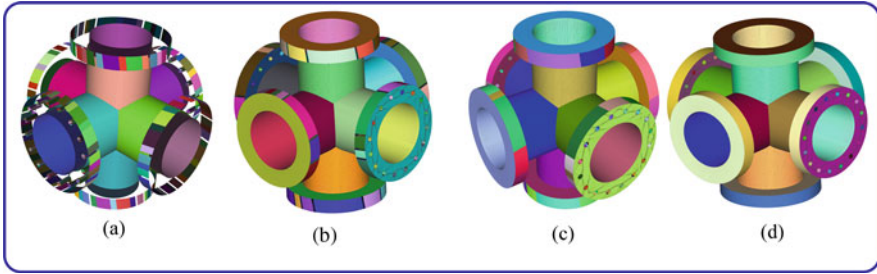
#### **11.3.2.2 Region Merging**

A single pass is not enough to merge all features. Only if two features are adjacent, they will be merged to one on satisfying geometry equality test. After merging, adjacency may be changed, so features that were not eligible for merging in the previous pass will be merged in next pass.

#### **11.3.2.3 Reclamation**

After region merging, small cracks are observed close to the corner and at the region boundaries [24]. To make a watertight model, these uncollected facets are reclaimed into the adjacent identified regions (Feature) based on reclamation criteria.

Figure 11.4 illustrates the cylindrical regions generated by the hybrid mesh segmentation, Fig. 11.3a shows the original mesh models, Fig. 11.4a demonstrates the segmentation results (12 planes and 523 cylindrical patches), Fig. 11.4b demon-



**Fig. 11.4** Hybrid mesh segmentation process. (a) Segmentation. (b) Region merging. (c) Reclamation. (d) Region merging after reclamation

**Table 11.1** A quantitative comparison of CAD mesh model

Test cases	$F$	$V$	$S$	Adf	$N_{Rbrm}$	$N_{Rarm}$	$T$	$C$
Figure 11.5a	1640	812	0.417	0.8	39	20	0.211	100
Figure 11.5c	2472	1230	0.624	0.6	55	29	0.864	99.67
Figure 11.5e	38,932	19,092	9.84	0.7	1169	630	4.257	99.58
Figure 11.5g	1380	690	0.349	0.75	36	25	0.254	99.28
Figure 11.5i	12,068	6034	2.23	0.75	158	69	1.078	100
Figure 11.5k	528	264	0.134	0.75	21	11	0.121	100

$F$ : number of facets,  $V$ : number of vertex,  $S$ : STL size (in MB), Adf: predicted area deviation factor,  $C$ : % coverage,  $T$ : overall timing (in a second),  $N_{Rbrm}$ : number of regions before region merging,  $N_{Rarm}$ : number of regions after region merging

strates the region is merging results, Fig. 11.4c demonstrates the reclamation results, and Fig. 11.4d illustrates the final region merging after reclamation (12 planes and 50 cylinders). The system takes approximately 1.759 s for feature detection.

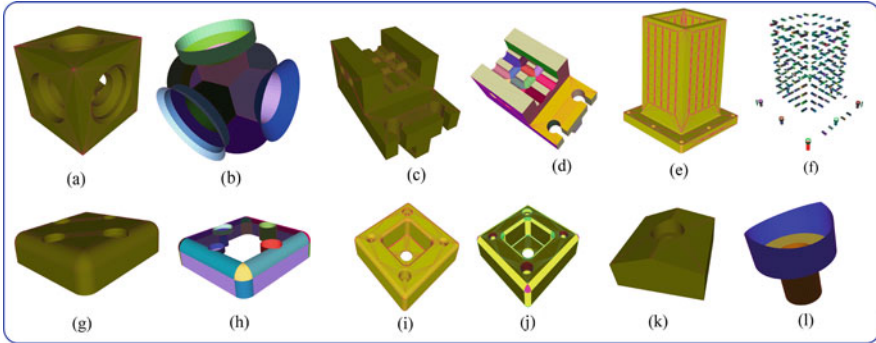
## 11.4 Volumetric Feature Recognition

The volumetric features like holes and blends are detected by applying a set of rules based on adjacency information of the primitives noticed in the previous step. Most of the existing approaches evaluate pockets, slots, etc. However, 60% of the average portion of the total facets in CAD mesh model are blended features [25], and holes constitute a significant percentage of features in mechanical engineering parts. Hence, we considered blends and hole recognition.

To test the efficacy of propose algorithm to recognize volumetric features, the benchmark test cases from repository have been used. These test cases have either complex interacting features or the freatures are in large in number. Using random color for different primitives, features can be interpreted.

Table 11.1 summarizes the performance measure for a proposed algorithm for the test cases shown in Fig. 11.5a, c, e, g, i, k. We used percentage coverage as a





**Fig. 11.5** Illustrates the interacting feature recognition of a model. (a) Test case 1. (b) Output of test case 1. (c) Good die. (d) Output of good die. (e) Tooling block. (f) Features of tooling block. (g) Text case 2. (h) Output for test case 2. (i) Test case 3. (j) Output of test case 3. (k) Test case 4. (l) Features of test case 4

measure of an indicator for successful segmentation. It is a ratio of a number of features recognized to actual the number of features present in a CAD mesh model.

## 11.5 Results and Discussion

### 11.5.1 Comparison with a Recently Developed Algorithm

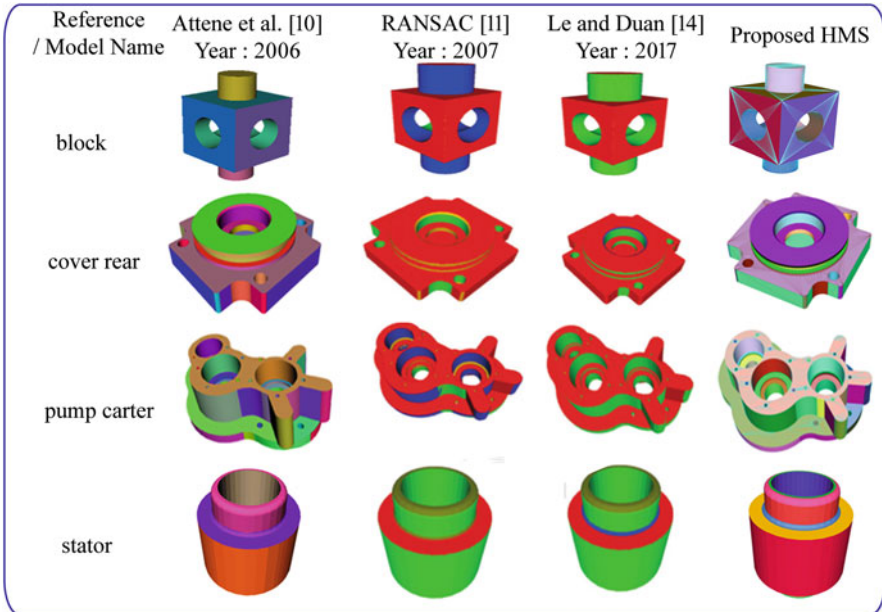
The comparison of the propose technique is made with existing state-of-the-art approaches like RANSAC [11], Attene et al. [10], and Li et al. [12] where source code is publicly available. The results for Le and Duan [14] are taken from [14] as the source code was not available. The proposed approach does not depend on attributes like curvature, minimum feature dimension, number of clusters, number of cutting planes, the orientation of model, and thickness of the slice to extract volumetric features.

Table 11.2 summarizes the quantitative comparison for a proposed algorithm for the benchmark test cases. Quantitative evaluation has been carried out using a number of primitives, the coverage percentage, and the distance error. As noted in Fig. 11.6, the proposed algorithm yields better results than RANSAC [11] and Attene et al. [10]. The results revealed that the proposed technique is comparable to Le and Duan [14].

**Table 11.2** Quantitative evaluation of primitive quality test cases shown in Fig. 11.6

Model name	Number of primitives					Coverage (%)					Distance error ( $\times 10^{-3}$ )				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Block	<b>14</b>	14	14	9	14	<b>100</b>	100	100	64.3	99	<b>0.04</b>	0.37	0.08	n/a	0.69
Cover rear	<b>45</b>	28	45	45	28	<b>100</b>	87.8	100	100	87.8	<b>0.02</b>	0.11	0.04	n/a	0.15
Pump carter	<b>83</b>	57	63	76	57	<b>99.5</b>	92.9	98.6	99.2	92.9	<b>0.03</b>	0.16	0.3	n/a	2.3
Stator	<b>12</b>	12	12	6	n/a	<b>100</b>	100	100	50	n/a	<b>0.01</b>	0.8	0.47	n/a	n/a

I: Proposed algorithm, II: RANSAC [11], III: Le and Duan [14], IV: Attene et al. [10], V: GlobFit [12]



**Fig. 11.6** Comparison with the existing algorithm

### 11.6 Conclusion

In this research, an elegant method has been proposed and implemented for extracting volumetric features from CMM using a hybrid region growing approach. The rule-based reasoning approach for feature recognition has been used. The proposed algorithm captures and separates intersecting features as well.

Comparing with existing recent approaches such as Attene et al. [10], RANSAC [11], Adhikary et al. [13], Le and Duan [14], Muraleedharan et al. [20], and other benchmark test cases, the proposed technique successfully recognized the features such as blends, compound holes and their interactions and found to be robust and

consistent with coverage of more than 95% in addressing volumetric features. The proposed approach is simple, general, and more reliable.

The future work could be aimed at capturing the parent-child relationship of extracted features and threshold prediction using various methods such as deep learning, machine learning for automatic segmentation.

**Acknowledgments** This work is supported by Centre for Computational Technologies, Pune, India. We also appreciate the authors of the Attene et al. [10], Schnabel et al. [11], and Li et al. [12] for making their code publicly available. Authors are also grateful to Dr. Truc Le, Dr. Ye Duan, and Dr. V.S. Gadakh for helping us to compute percentage coverage.

## References

1. D. Yan, W. Wang, Y. Liu, Z. Yang, Variational mesh segmentation via quadric surface fitting. *Comput. Aided Des.* **44**(11), 1072–1082 (2012)
2. S. Xú, N. Anwer, C. Mehdi-Souzani, R. Harik, L. Qiao, STEP-NC based reverse engineering of in-process model of NC simulation. *Int. J. Adv. Manuf. Technol.* **86**(9–12), 3267–3288 (2016)
3. T. Várady, M. Facello, Z. Terék, Automatic extraction of surface structures in digital shape reconstruction. *Comput. Aided Des.* **39**(5), 379–388 (2007)
4. P. Benkő, T. Várady, Segmentation methods for smooth point regions of conventional engineering objects. *Comput. Aided Des.* **36**(6), 511–523 (2004)
5. D. Tang, L. Zheng, Z. Li, An intelligent feature-based design for stamping system. *Int. J. Adv. Manuf. Technol.* **18**(3), 193–200 (2001)
6. J. Corney, C. Hayes, V. Sundararajan, P. Wright, The CAD/CAM interface: a 25-year retrospective. *J. Comput. Inf. Sci. Eng.* **5**(3), 188–197 (2005)
7. M. Hayasi, B. Asiabanpour, Extraction of manufacturing information from design-by-feature solid model through feature recognition. *Int. J. Adv. Manuf. Technol.* **44**(11–12), 1191–1203 (2009)
8. F. Bianconi, Bridging the gap between CAD and CAE using STL files. *Int. J. CAD/CAM* **2**(1), 55–67 (2002)
9. V. Sunil, S. Pande, Automatic recognition of features from freeform surface CAD models. *Comput. Aided Des.* **40**(4), 502–517 (2008)
10. M. Attene, B. Falcidieno, M. Spagnuolo, Hierarchical mesh segmentation based on fitting primitives. *Vis. Comput.* **22**(3), 181–193 (2006)
11. R. Schnabel, R. Wahl, R. Klein, Efficient RANSAC for point-cloud shape detection. *Comput. Graph. Forum* **26**(2), 214–226 (2007)
12. Y. Li, X. Wu, Y. Chrysathou, A. Sharf, D. Cohen-Or, N. Mitra, GlobFit: consistently fitting primitives by discovering global relations. *ACM Trans. Graph.* **30**(4), 52, 12 (2011)
13. N. Adhikary, B. Gurumoorthy, A slice based approach to recognize and extract free-form volumetric features in a CAD mesh model. *Comput. Aided Des. Appl.* **13**(5), 587–599 (2016)
14. T. Le, Y. Duan, A primitive-based 3D segmentation algorithm for mechanical CAD models. *Comput. Aided Geom. Des.* **52–53**, 231–246 (2017)
15. J. Shah, D. Anderson, Y.S. Kim, S. Joshi, A discourse on geometric feature recognition from CAD models. *J. Comput. Inf. Sci. Eng.* **1**(1), 41–51 (2001)
16. B. Babic, N. Nesic, Z. Miljkovic, A review of automated feature recognition with rule-based pattern recognition. *Comput. Ind.* **59**(4), 321–337 (2008)

17. A. Verma, S. Rajotia, A review of machining feature recognition methodologies. *Int. J. Comput. Integr. Manuf.* **23**(4), 353–368 (2010)
18. R. Zbiciak, C. Grabowik, Feature recognition methods review, in *Proceedings of the 13th International Scientific Conference. RESRB 2016. Lecture Notes in Mechanical Engineering*, ed. by E. Rusiński, D. Pietrusiak, (Springer, Cham, 2017), pp. 605–615. [https://doi.org/10.1007/978-3-319-50938-9\\_63](https://doi.org/10.1007/978-3-319-50938-9_63)
19. D. Xiao, H. Lin, C. Xian, S. Gao, CAD mesh model segmentation by clustering. *Comput. Graph.* **35**(3), 685–691 (2011)
20. L. Muraleedharan, S. Kannan, A. Karve, R. Muthuganapathy, Random cutting plane approach for identifying volumetric features in a CAD mesh model. *Comput. Graph.* **70**, 51–61 (2018)
21. A. Agathos, I. Pratikakis, S. Perantonis, N. Sapidis, P. Azariadis, 3D mesh segmentation methodologies for CAD applications. *Comput. Aided Des. Appl.* **4**(6), 827–841 (2007)
22. V. Hase, Y. Bhalerao, G.J. Vikhe Patil, M.P. Nagarkar, in *Proceedings of the ICCET 2019 4th International Conference on Computing in Engineering and Technology. ICCET 2019*, (AISC Series of Springer, 2019), ed. by B. Iyer, P. S. Deshpande, S. C. Sharma, U. Shiurkar, (2019)
23. V. Hase, Y. Bhalerao, S. Verma, S. Jadhav, G. Vikhe Patil, Complex hole recognition from CAD mesh models. *Int. J. Manage. Technol. Eng.* **8**(IX), 1102–1119 (2018)
24. H. Kim, H. Choi, K. Lee, Feature detection of triangular meshes based on tensor voting theory. *Comput. Aided Des.* **41**(1), 47–58 (2009)
25. N. Rafibakhsh, M. Campbell, Hierarchical fuzzy primitive surface classification from tessellated solids for defining part-to-part removal directions. *J. Comput. Inf. Sci. Eng.* **18**(1), 011006 (2017)

# Chapter 12

## Factors Affecting a Mobile Learning System: A Case Study



Sudhindra B. Deshpande, Shrinivas R. Mngalwede, and Padma Dandannavar

### 12.1 Introduction

Education is an idea of lifelong learning, initiative learning. In today's world one of the most interesting domains available through the Internet is distance learning [1]. With the increase of networks and mobile computing, people are more interested in distance learning [2]. The m-learning focuses majorly on the student versatility with convenient gadgets [3], and discovering that how society and its organizations can oblige and bolster with an increasingly portable mobile population.

With technology advancements in smart devices like mobile devices, ipads, and tabs, people are more interested in distance and mobile learning as the abilities of these pervasive gadgets are expanding at an unfaltering rate [4, 5]. The students are distinctive in age level, sex, social job, their way of life, training foundation, consideration, and premiums; pastimes have an extraordinary effect in their learning conduct [6].

Providing corresponding learning content and strategies to acknowledge instructing as indicated by students' needs is an exceptionally difficult and a very challenging task [7]. With emerging mobile devices teaching/learning has to change in its entirety to adapt to this new mobile education [8]. m-Learning, which gives a consistent figuring out how students, can conquer any hindrance of portable mobile

---

S. B. Deshpande (✉)

Department of Information Science and Engineering, Gogte Institute of Technology, Belagavi, Karnataka, India

e-mail: [sbdeshpande@git.edu](mailto:sbdeshpande@git.edu)

S. R. Mngalwede · P. Dandannavar

Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi, Karnataka, India

e-mail: [mangalwede@git.edu](mailto:mangalwede@git.edu); [padmad@git.edu](mailto:padmad@git.edu)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_12](https://doi.org/10.1007/978-3-030-19562-5_12)

121

and desktop computing [9]. A large gap can be found between learners' expectation and in the actual m-learning.

Therefore, it is the need of hour to study and find out the factors that affect m-learning [10, 11]. Learner's context, various types of smart devices [12], speed of the network, and hardware and software of smart devices are the major factors that affect internal quality of m-learning [13]. Therefore, it is very meaningful to study and find the learner factor that impacts on m-learning.

## 12.2 Factors Affecting the m-Learning

The m-learning system should be adaptive to the needs of different learners, who have different individual mobile learning preferences. The m-learning preferences of a learner can be used as a basis for providing personalized learning platforms catered to the individual needs of learners. The learning styles are the major factors that impact on m-learning. Learners have different styles, preferences of learning, and tools which assist them in effective learning. Understanding how a learner learns is called meta-cognition, means thinking about thinking. Meta-cognition is about the perceptive how learner, as an individual, learns the best.

A learning style portrays the manner in which that learner wants to learn; learner may utilize certain procedures or like accepting data in a specific way. The learning styles can be affected by the manner in which learner think, feel, and behave. The learner factors can be influenced by personalization, the context of a learner. The two important factors are: (1) learner analysis and (2) context analysis.

### 12.2.1 Learner Analysis

Learner analysis includes analysis of learning behavior, styles, type of learning, and brain dominance. Each learner carries different characteristics of each of them. Attributes of the learners impact learning objectives and effect the way in which learning happens. Understanding and taking into contemplations the qualities of the learners can decide if the learning knowledge is significant.

Creating instruction that suits to every type of learning style for learners is not easily possible. Understanding the various learners learning styles can provide alternatives. Types of learning styles of learners are listed below in Table 12.1.

Based on the above learning styles, four learner categories are listed in Table 12.2.

**Table 12.1** Various learning styles

Style	Description
Tactile/kinesthetic	Learners prefer physical engagement, i.e., “Hands on” activity. Prefer performing/doing practices rather just reading.
Visual/perceptual	Learners prefer looking. Demonstrations, for example, charts, writing on a blackboard, diagrams, and graphs are of interesting to them. Visual learners recall best what they see—pictures, outlines, flowcharts, courses of events, movies, and exhibits.
Auditory	Learners prefer information presented in an oral way. For example, classroom; listening to lectures; participating in group discussions.

**Table 12.2** Lerner categories

Type	Description
Active	Learners comprehend the data best by effectively accomplishing something with it. [Discussions/applying/explaining to others.]
Reflective	Learners desire to think. [Think about the information is reflective learner’s response.]
Sequential	Learners are preferred to learn slowly. [Step-by-step explanation, in an orderly process, up to the end result.]
Global	Learners like to have examples so that they know where they are headed and what they are working toward. Before learning a complex process learners first prefer an overview of what and how they are going to do.

### 12.2.2 Context Analysis

In addition to analyzing the learners, the learning context should also have to be analyzed. For, understanding the setting in which new abilities, information, or state of mind will be utilized can advise the arranging of instructional exercises that will estimate what learners will look in reality. Also, a comprehension of the learning context encourages in recognizing obstructions in the setting and best utilizes the instructional condition. It additionally includes depicting the idea of the learning context and compatibility and requirements of the environment for the learners and instructional objectives. Personalization provides personalized learning depending on the learners’ profile; profiles are constructed based on the various factors of students’ characteristics, like:

- In which location student prefers to study—home, college, laboratory, library, lounge park, office, etc.
- Preferences for sensed distractions within locations—noise, activities in surrounding, environmental factors, light, temperature, room layout, near-by attractions, seating.

- Personal factors like—friends, working culture, food, drink, time of the day, likes to be alone, in a group.
- Format of content—learner prefers audio, video, or animations in learning.
- Preferred time of study—daytime, morning, afternoon, evening, night, or midnight.
- Type of smart devices, network type and other features like screen size, RAM, etc.

### 12.3 A Case Study Based Results and Discussions

For the case study, Java Programming for the students has been considered. We have captured individual learning preferences of Java Programming. 240 students of various branches of engineering participated in this study. Table 12.3 provides the details of students of various branch of engineering.

Table 12.4 furnishes the figures of the 240 students’ interests in studying Java Programming.

We have collected data with respect to context that we have assumed. We have used Google form to collect the data. Table 12.5 will give details about the characteristics that are assumed for our circumstances and assumptions.

These factors help to identify the context of a student. Figure 12.1 depicts an example of context scenario of student 1.

Another similar scenario of student 2 is represented in Fig. 12.2.

**Table 12.3** Number of students participated

Branch	Number
Information Science	50
Computer Science	60
Mechanical	40
Civil	40
Electronics	50

**Table 12.4** Opinion about Java Programming

Questions	No (%)	Yes (%)
I enjoy studying Java in any noisy situation	76	34
I can study Java in any location with full concentration.	77	33
Studying Java, whatever the mood, time may be, makes me joyous/happy.	44	56
Studying Java bores me	84	16
Java motivates me to learn to programming	15	85



**Table 12.5** Context characteristics and possible values

Characteristics	Possible values
Place	House—Room or hall
	College—Classroom or library
Time	Morning
	Afternoon
	Evening
	Night
Posture	Laying on
	Table and chair
Noise	Alone
	One friend
	Two or more friends
Smart device	Smart phone
	iPad
	Tab
Network	4g
	3g
	2g
RAM	6 GB
	4 GB
	2 GB
	1 GB
Battery	Full
	Average
	Low
Screen size	6"
	5"
	4"
	10"
Content	Video (demonstrations, films, animations)
	Audio (mp3)
	Text/pdf (hands on, tutorials, pictures/images)

Similarly we have captured the context of all the students, and content format they are interested in. Students are interested in various formats of content relative to their day-to-day contexts. Few sets of students are interested in fixed content formats and other showed interest in mixed kind of content formats.

The graph in Fig. 12.3 depicts the formats of content accessed by different number of students. Content formats change with varying contexts of students in routine life.

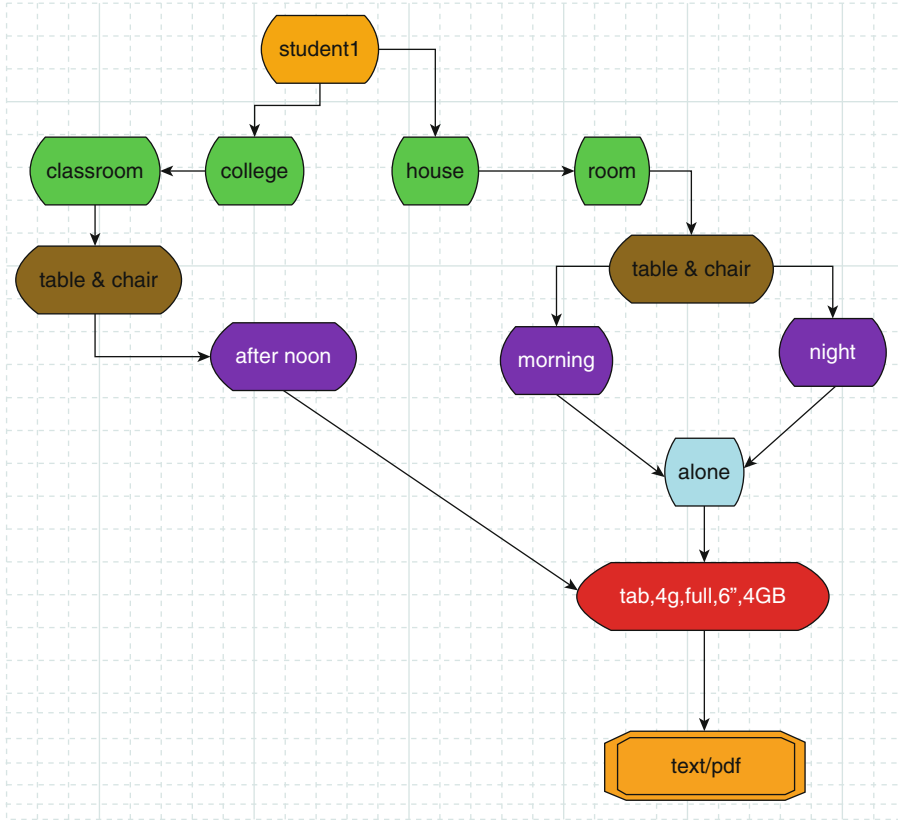


Fig. 12.1 Scenario of student 1

The learning styles are of three types: visual, auditory, and tactile. Students are interested to study various content formats: only video or only audio or only text; in any context, college or house. Students have shown interest in studying mixed content formats also: text and audio, text and video, text and animations. Figure 12.4 shows the mapping of learning styles mapped to the students.

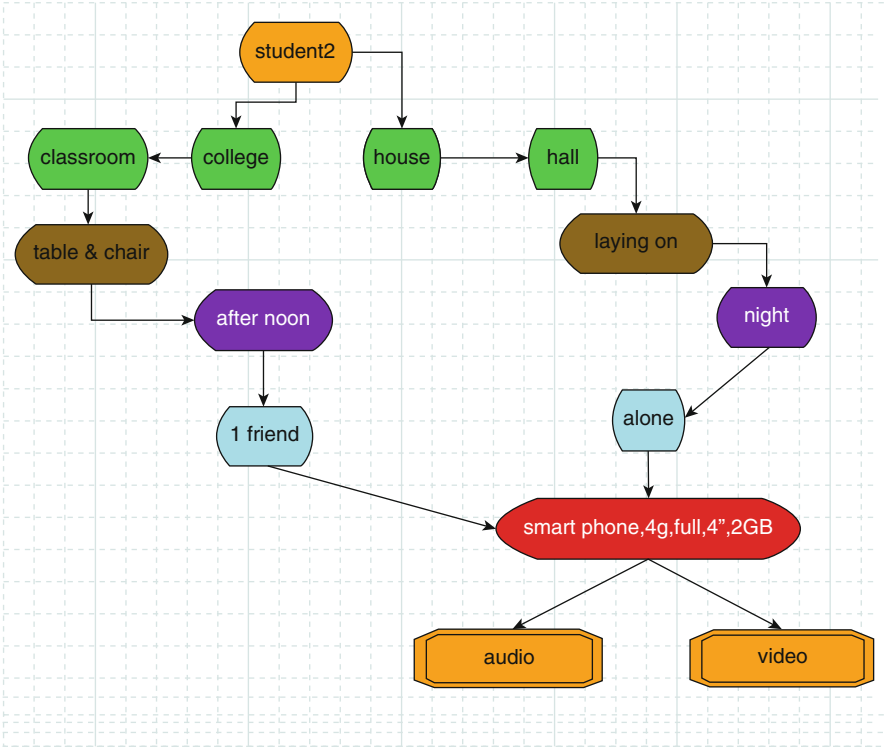


Fig. 12.2 Scenario of student 2

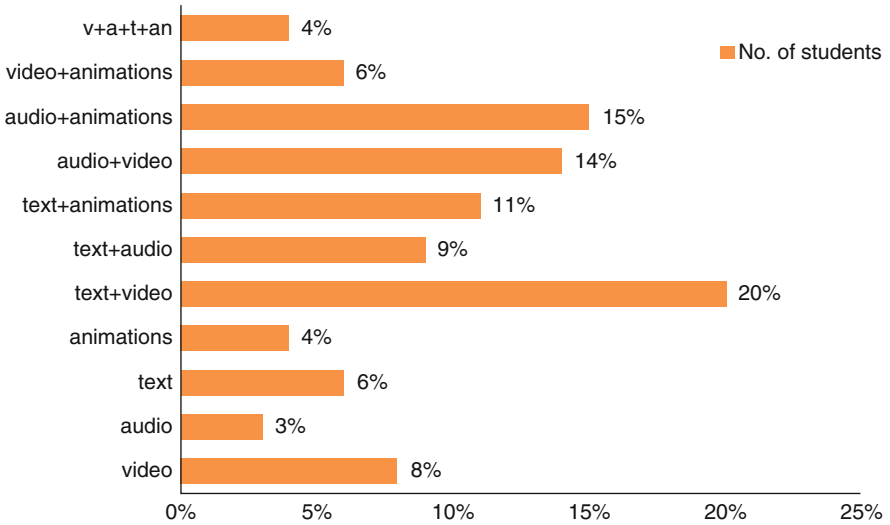
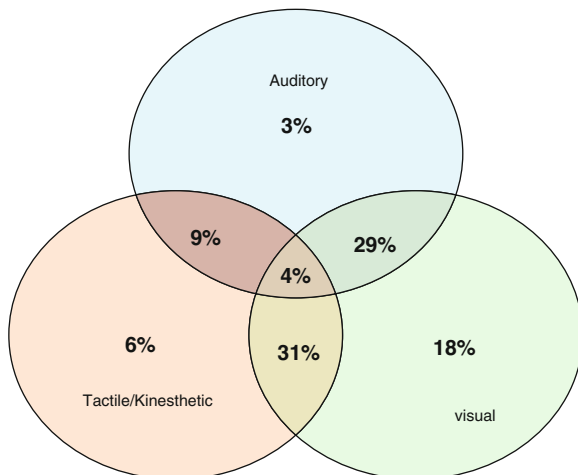


Fig. 12.3 Content formats interested against the number of students

**Fig. 12.4** Mapping of the students to learning styles



## 12.4 Conclusions

In this chapter an attempt has been made to study and analyze the factors that affect m-learning system. The chapter focuses on two factors: (1) Learner analysis, which is identifying what type of learner he/she is, and (2) context analysis—the real-time scenario the learner involved in. As m-learning should cater the needs of learner with various learning styles and different contexts, this study helped to understand behavior of the learners with respect to their preferred contexts and also various content formats to be delivered. Understanding each learner’s context and preferences is very crucial as the individual needs and requirements are different. So it’s challenging and complex task to cater the adaptive content delivery system based on the various interests of various learners at one platform.

## References

1. M. Hamdani, Advance Principles for Visual Aesthetics in Designing the Contents of E-Learning (2012)
2. C. Hesselbarth, S. Schaltegger, Educating change agents for sustainability—learnings from the first sustainability management master of business administration. *J Cleaner Prod* **62**, 24–36 (2014)
3. W. Horton, K. Horton, E-learning Tools and Technologies [online] (2010). <http://www.egypteducation.org/moodle/file.php/22/elearningbook.pdf>
4. C.E. Ortiz, An Introduction to the Mobile Context and Mobile Social Software (2008). <http://cenriqueortiz.com/pubs/themobilecontext/themobilecontext-cenriqueortiz.pdf>

5. X. Zhao, An email-based discussion learning system with ubiquitous device support, in *2009 Fourth International Conference on Computer Science and Education*, Jul 2009
6. X. Li, Q. Luo, J. Yuan, Personalized recommendation service system in E-learning using web intelligence, in *Computational Science – ICCS 2007. ICCS 2007. Lecture Notes in Computer Science*, ed. by Y. Shi, G. D. van Albada, J. Dongarra, P. M. A. Sloot, vol. 4489, (Springer, Berlin/Heidelberg, 2007)
7. U. Farooq, W. Schafer, M.B. Rosson, J.M. Carroll, M-education: bridging the gap of mobile and desktop computing, in *IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE02)*, 2002, pp. 91–94
8. E. Riedling, K. Riedling, An E-Learning Platform with a Deliberately Simple Design [online] (2003). [http://publik.tuwien.ac.at/files/pub-et\\_11167.pdf](http://publik.tuwien.ac.at/files/pub-et_11167.pdf)
9. S. Na, The impact of learner factor on e-learning quality, in *International Conference on e-Learning e-Business Enterprise Information Systems and e-Government*, Dec 2009
10. B.D. Sudhindra, R.M. Srinivas, Context-aware personalized M-learning application using multi agents. *Int. J. Comput. Theory Appl.* **9**(10), 1–11 (2016). ISSN-0974-5572
11. S.L. Brown, K.M. Eisenhardt, Product development: past research, present findings, and future directions. *Acad. Manag. Rev.* **20**(2), 343–378 (1995)
12. A. D’Andrea, F. Ferri, L. Fortunati De Luca, T. Guzzo, Mobile devices to support advanced forms of e-learning, in *Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility*, ed. by P. Grifoni, (IGI Global, Hershey, PA, 2009)
13. K. Friedman, Theory construction in design research: criteria, approaches, and methods. *Des. Stud.* **24**, 507–522 (2003)

# Chapter 13

## Document Similarity Approach Using Grammatical Linkages with Graph Databases



V. Priya and K. Umamaheswari

### 13.1 Introduction

Document similarity assessment is helpful in exploring linked documents based on a source text. It is a useful mechanism for many fields dependent on text processing. Most systems measure text similarity [1] depending on the word distribution statistics. They measure the similarity assuming that the words have analogous meaning when they appear in the identical environment. These word-based techniques become vulnerable to many complications when they deduce some inference from text without using clear knowledge. When utilizing distributional measures, word-based approaches become inefficient for comparisons. Some of the reasons include document heterogeneity, varied vocabulary, text length, and languages.

In traditional way the semantic similarities between the documents cannot be found accurately because the semantic considerations are not used. So the document can be easily modified and used for other purposes. To overcome this problem introducing the semantic meaning from Word Net has been widely used to find the similarities between the documents. Conventional approaches use hypothesis based on text distribution statistics. They assume that two words might have same meaning when found in the similar situation. When documents are heterogenic in nature, concluding interpretations in the text with no clear knowledge might rise to have problems.

---

V. Priya (✉)

Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

K. Umamaheswari

PSG College of Technology, Coimbatore, Tamil Nadu, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_13](https://doi.org/10.1007/978-3-030-19562-5_13)

131

Word distributional metrics are not feasible since documents with different vocabularies and length follow a diverse distribution of words. It becomes complex to compare these textual units. Conventional semantic similarity approaches utilizing semantic graphs are infeasible because of expensive operations. So a new approach involving grammatical linkages using verbal intent technique is proposed and using this document similarity can be computed efficiently in comparison with other state-of-the-art graph-based mechanisms.

Section 13.2 of the chapter details several techniques used for text similarity using graph-based approaches. Section 13.3 presents the work on the verbal intent-based document similarity identifier. Finally, Section 13.4 presents conclusion.

## 13.2 Literature Survey

This section presents a study on various techniques used in detection and computation of document similarity. Researches explored a variety of measures depending on textual units, graphical units, and semantic units.

Some of the commonly used text-based techniques in document similarity identification and computation are text-based and semantic-relation-based approaches. The authors [2] have utilized two measures which rely on character and term-based algorithms for computing the similarity of two documents. In the first method, n-gram is utilized to identify fingerprint using winnowing algorithms. Then Dice coefficient is adopted to find similarity in the two fingerprints identified. They have employed an algorithm to link noun expressions using an affiliated multi-lingual corpus.

Christian et al. [3] have analyzed that document similarity could be computed effectively compared to graphical unit-based methods by using similarity measure. The measure should provide a significantly higher connection with human notions of document similarity. The authors in [4] have employed random graphical approach for computing comparative prominence of textual units and a detailed analysis of Lex rank mechanism and applied it to a huge data set. They deliberate the helpfulness of applying random walks to sentence-based graphs would improve in text summarization. They also briefly explain the possibility of deploying such methods in NLP tasks such as classifying named entities, attaching prepositional phrases, and text categorization. Graph-based centrality has quite a few advantages over Centroid method. In [5] document similarity has been studied using semantic similarity. Semantic similarity [6–10] and their measures had been explored in literature extensively. In these schemes semantic similarity among the concepts found in Knowledge Graphs such as WordNet and DBpedia are measured using wpath metric. This combines information content to identify the length of the shortest path between any two concepts.

### 13.3 Proposed Work

This section presents the work on text similarity detection using verbal intent and graph databases. The design of the proposed system is shown in Fig. 13.1. This consists of a Data preprocessing module which tokenizes and tags the entities from the sentences. Tokenization is done with Part-of-Speech (POS) tagging. Identification of entities is done with POS tags [11]. They are used for generating knowledge graph with links. Now weights are assigned using verbal intent technique along with graph database generation. Lastly, similarity is calculated based on the weights from the verbal intents and links in the knowledge graph.

#### 13.3.1 Architecture

The complete design of the system is shown (in Fig. 13.1).

##### 13.3.1.1 Tokenization

In the given input documents, tokenization is done. Tokenization is performed by splitting the sequence of strings. This splitting results in individual elements called as tokens which are keywords, phrases, symbols, and other special elements. Spaces, tags, and special characters which are considered to be irrelevant are removed.

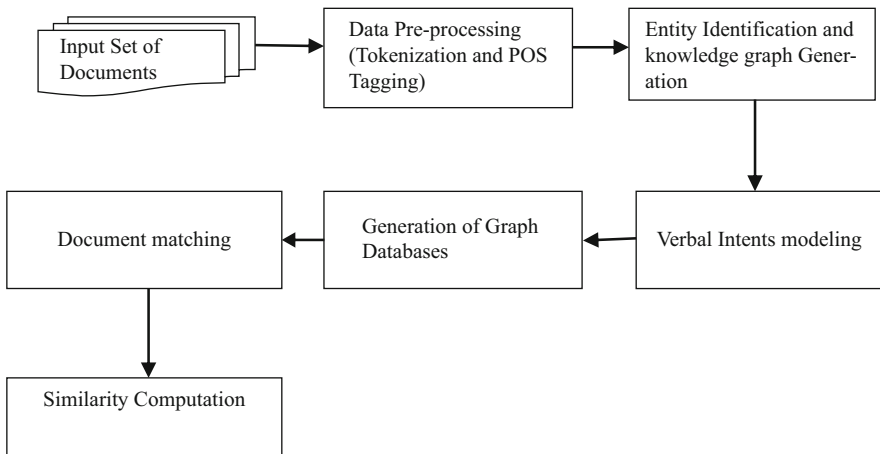


Fig. 13.1 System architecture



### 13.3.1.2 Verbal Intent Modelling

The verbal intent of any document is computed as follows. A verbal intent describes the author's action perspective, based on the subject of the document. Given document and its noun terms, a verbal intent is modelled as follows: First verbal feature is defined and represented as a vector of weighted values as shown in Eq. (13.1):

$$V_n = W_s + W_a \quad (13.1)$$

where every facet of a vector of verb  $V_n$  matches conventional relevant verbs intended for the noun associated,  $W_s$  and  $W_a$  represents the weight from synonyms and adverbs of the referenced document, respectively. Given a document, computation of verbal weights is performed using cosine similarity between each term vector of document and  $d_n$  is done as given in Eq. (13.2).

$$W_n = \text{cosine}(V_n, d_n) \quad (13.2)$$

where the weight of every term in the term vector is calculated using the usual tf-idf scheme adopted from the Vector Space prototype [7] of the document.

The partial results are obtained with cosine similarity metric with summarized data. It is likely that the performance improvement could be observed using semantic knowledge bases and graph databases.

## 13.4 Conclusion

Document similarity systems are found to have enormous usage for many applications like plagiarism detection, template matching, and so on. The approach using verbal intents for document similarity computation was deliberated. The system is expected to generate feasible results for small documents such as short summaries as well as large size documents in the corpus. Further the system could be improved in introducing intents for all entities to improve semantic similarity. Other promising future directions include extending the system for online documents and template verification in IT contract services which can improve customer relationship.

## References

1. R. Anna, Z. Silvia, Assessing semantic similarity of texts—methods and algorithms, in *Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and Economics*, 2010, pp. 1–8

2. K. Julian, An algorithm for finding noun phrase correspondences in bilingual corpora, in *Proceedings of the 31st Annual Meeting on Association of Computational Linguistics*, 2012, pp. 17–22
3. P. Christian, R. Achim, M. Aditya, Efficient graph-based document similarity, in *Proceedings of the 13th International Conference on the Semantic Web. Latest Advances and New Domains*, vol 9678, 2016, pp. 334–349
4. E. Gunes, R. Dragomir, LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 457–479 (2015)
5. Z. Ganggao, A. Carlos, Computing semantic similarity of concepts in knowledge graphs. *IEEE Trans. Knowl. Data Eng.* **29**(1), 72–85 (2017)
6. R. Philip, Using information content to evaluate semantic similarity in a taxonomy, *ACM Digital Library*, 1995, pp. 448–453
7. E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and WordNet-based approaches, in *Proceedings of Human Language Technology Annual Conference North American Chapter Association of Computational Linguistics*, 2009, pp. 19–27
8. A. Broder et al., A semantic approach to contextual advertising, in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 559–566
9. J.-H. Lee et al., Semantic contextual advertising based on the open directory project. *ACM Trans. Web* **7**(4), 1–24 (2013)
10. N. Takagi, M. Tomohiro, Wsl: sentence similarity using semantic distance between words, in *Proceedings of the Ninth International Workshop on Semantic Evaluation*, 2015, pp. 128–131
11. A. Gupta, D.K. Yadav, Semantic similarity measure using information content approach with depth for similarity calculation. *Int. J. Sci. Technol. Res.* **3**(2), 165–169 (2014)

# Chapter 14

## Missing Data Handling by Mean Imputation Method and Statistical Analysis of Classification Algorithm



K. Maheswari, P. Packia Amutha Priya, S. Ramkumar, and M. Arun

### 14.1 Introduction

Nowadays data mining tools are used for analysis of customer behavior and their relationship to increase the profit by several companies such as retail marketing, insurance, banking, telecommunications and product sales of consumer, finance and health care. To understand the business aspects, data mining helps the organization to provide betterment of customer satisfaction and serve in order to increase the growth of the organization in the future.

To increase the market space among other competitors, the retailers can know all the information related to the customer based on why, what and they buy the products and who the customer are. By applying the benefits of data mining techniques to various sectors, the huge quantities of data related to customer behavior, product supplier, list of products, and their sales of product were analyzed.

In any marketing business, data mining plays an important role not only in paying more attention on customer but also in maintaining the existing customers without leaving the competitors. In the retail marketing business, information gained by data mining techniques can be useful in various ways:

- Profit of the business can be increased by reducing the cost of the product.
- Stock price can be increased in order to maintain the future plan to raise the profit.

If a marketing business fails to retain the existing customers, then the company will not be able to retain its position in the market, their shares will go down and the profit of the company goes down slowly and it finally disappears. This chapter

---

K. Maheswari (✉) · P. Packia Amutha Priya · S. Ramkumar · M. Arun  
School of Computing, Kalasalingam Academy of Research and Education, Virudhunagar, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_14](https://doi.org/10.1007/978-3-030-19562-5_14)

137

is organized as follows: Section 14.2 deals with review of literature, Sect. 14.3 describes the methodology, Sect. 14.4 presents the results of this work, and Sect. 14.5 concludes the research work.

## 14.2 Review of Literature

Pratama et al. [1] discussed about the reasonable option used for guessing out the missing values used by other researchers in various research work and suggested that mean, mode, median, deletion, and other imputation methods are methods used to handle missing values in various time series dataset. The imputation method used in the research work will produce a solution to minimize the preference outcome of the predictable technique.

Moore and Carpenter [2] suggested decision tree method to analyze the behavioral and demographic issues on private label attire from the top five private label attire merchant from the USA. The author pointed out observable drivers is more frequent among customers of vendors that are distinguished based on brand or service provided to the customer.

Maheswari and Packia Amutha Priya [3] analyzed the purchase behavior of the customer using SVM algorithms and the experimental result is carried out using the inventory dataset and sales dataset for analyzing the customer purchase behavior. Maheswari and Packia Amutha Priya [4] experimented the research work using text documents. After preprocessing, the missing values in the documents were found out. The frequency of words present in the documents is analyzed and visually represented. The twitter dataset [5] was used to carry out classification using SVM and KSVM. This work focuses on categorizing the sentiments or emotions from various age groups of people. The author [6] experiments the twitter dataset using Knn to identify the human behavior by improving the accuracy result in the sentiment classification analysis.

Tinabo [7] described four possible data mining methods to the problem of customer retention in the retail sector and suggested decision tree to be the most effective technique. The decision is made based on the features of the retail datasets such as size of records. Islam and Habib [8] proposed a prediction model for analyzing the business region in retail commercial banking. Business customer records of both rural and urban fields from Bangladesh dataset is taken for the experiment by applying decision tree method in weka tool to analyze its performance.

Patel et al. [9] suggested building a decision tree classification model to categorize the training dataset based on the rules. Result of the proposed classification model strength was calculated based on their performance. Li and Zhang [10] surveyed the solution for handling empty value property, selecting more than one value property and condition based selection property problems. Simplified and weighted Entropy in decision tree algorithm performs better results when compared to ID3 algorithm during the experimental process.

Senapati et al. [11] mentioned a principal component analysis model to clear up the missing values present in the microarray dataset. The result shows that the accuracy produced by the proposed model was good when compared with other imputation methods. Linear Discriminant Analysis (LDA) was used to validate the proposed model.

Houard et al. [12] recommended the new sampling methodology to handle missing values during preprocessing process. The main aim was to produce the samples of good trait and the extracted information should be highly secure and dependable. Houcka et al. [13] inspects about the way to categorize the missing values depending on the machine learning and statistical techniques. In this work, the classification of the missing values by imputation, model-based, and ignoring value methods was performed. Each of the methods has integrated with their own merits and demerits.

Song and Lu [14] suggested visualizing the training dataset results in tree structure in order to characterize the SAS and SPSS programs by applying various algorithms namely QUEST, CHAID, CART, and C4. 5. Validation dataset is used to determine the correct tree size and attain the excellent model validation dataset which was used for their analysis.

Agarwal et al. [15] aimed to categorize the society college dataset of the student using support vector machines. Various classification methods are compared for their research study and find that SVM produces more accuracy and less root mean square error. Decision tree method may be used to find the course selection of the students during the program. Kishor Kumar Reddy et al. [16] attempted to summarize the proposed approaches, tools, etc. for decision tree learning with emphasis on optimization of constructed trees and handling large datasets. Further, we also discussed and summarized various non-decision tree approaches like Neural Networks, Support Vector Machines, and Naive Bayes.

## 14.3 Methodology

It is a diagrammatic flow structure where each node represents an attribute; each branch is a result of test condition. The leaf or terminal nodes represent a class label. The top node in a decision tree is designated as root node [17]. The decision tree is useful for constructing decision tree classifiers without any basic knowledge. It can be able to handle higher dimensional data in a database. Good accuracy can be given by decision tree classification technique [18].

### 14.3.1 *Methods for Handling Missing Data*

Missing data or values become one of the major problems that occur frequently during the data collection process. Missing data reduces the sample representation

and there was a alter presumption in population. It is important to know the reason why data values are missing in order to correct the remaining data. Missing data can be handled in different ways:

- Missing at Random (MAR)
- Missing completely at Random (MCAR)
- Missing not a Random (MNAR)

#### **14.3.1.1 Missing at Random**

It happens when the absence of data was not random or unplanned whereas the absence of data can be entirely computed for some variable where the information are complete.

#### **14.3.1.2 Missing Completely at Random (MCAR)**

Missing Completely at Random (MCAR) means any particular data values being missed are separate for unobserved variables and observed variables of important and occur fully at random choice.

#### **14.3.1.3 Missing Not a Random (MNAR)**

Non-ignorable non-response is also known as Missing not a Random (MNAR) is data point or values that is neither MCAR nor MAR. The missing value depends on the two probable reasons such as missing data is dependent on few other variable values or hypothetical value.

If an observation has one or more missing values, then all data value can be removed. This is known as list wise deletion. To a fewer number of observation, if the missing value is finite, then it is preferred to ignore these cases from the analysis process. Imputation is the process of replacing the missing values by some other predictable values in the entire dataset. It is the important step in the machine learning and data mining process whenever the definite values are missed in the Dataset. There are various ways to handle missing data in dataset:

- Removing the entire row in the dataset.
- Replacing the missing value with mean, median, and mode in the dataset.
- Assigning a unique categorical value.
- Missing values can be predicted.

Dataset is divided into two sets. They are:

- Without missing values as a one set of data for the training set.
- Dataset with missing values for the testing set has been used the work.

Using decision tree, logistic regression and ANOVA method for prediction process in retail marketing has been used in our proposed work and it is implemented by R programming.

### 14.4 Methodology Experimental Results

Missing data is an important issue in datasets which takes main role in statistical processing of machine learning algorithms. The data missed is not by the mistake of data collector. There are so many reasons to miss the data in the dataset. During data collection, the respondent may not respond for some questions. Sometimes the data will not be available, not applicable, and not possible to provide. The lack of non-responded answers is treated as missing values. The proper handling of the missing values, empty values, NA values, not relevant values, and improper values draws inaccurate conclusion about the data set. One of the most important techniques for handling missing data is imputation method.

It can be seen that the variables shown in the above chart have missing values from 30% to 40%. The margin plot is shown in Fig. 14.1.

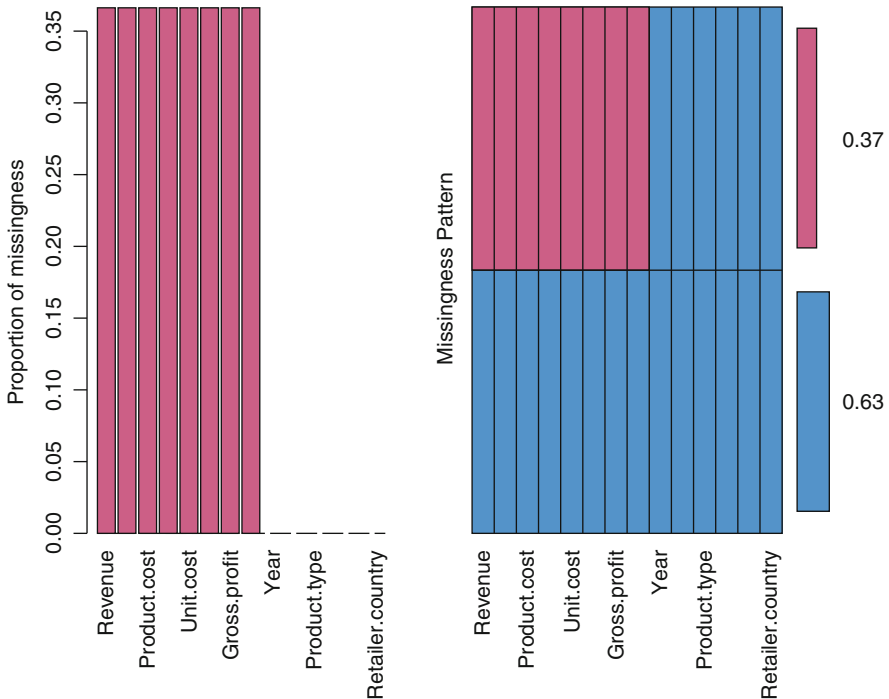


Fig. 14.1 Dataset with missing values

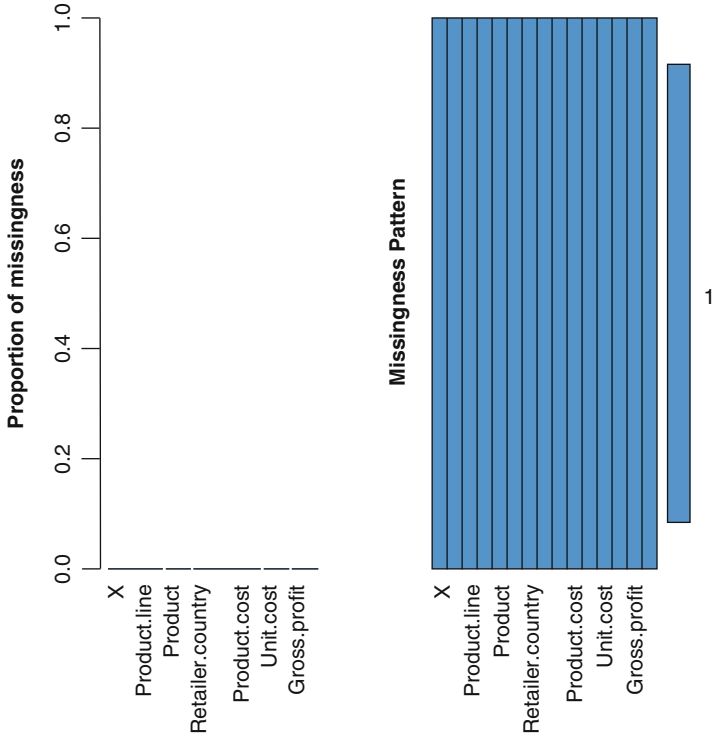


Fig. 14.2 Missing pattern

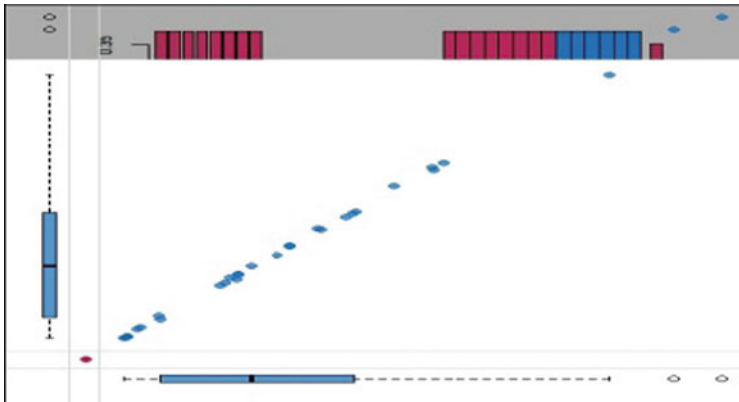


Fig. 14.3 Margin plot of dataset

The missing pattern of dataset is presented in Fig. 14.2. The margin plot of data set is shown in Fig. 14.3. The plotting of two attributes, revenue and product cost, is shown in Fig. 14.4.



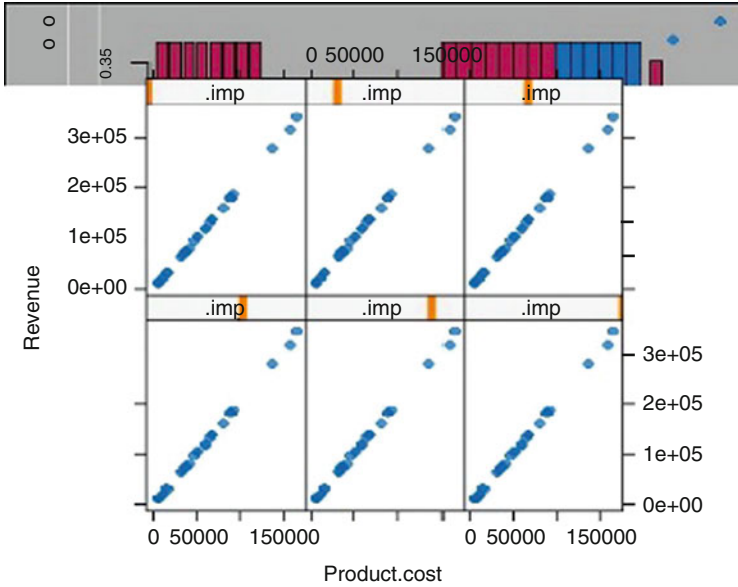


Fig. 14.4 Plot with two attributes

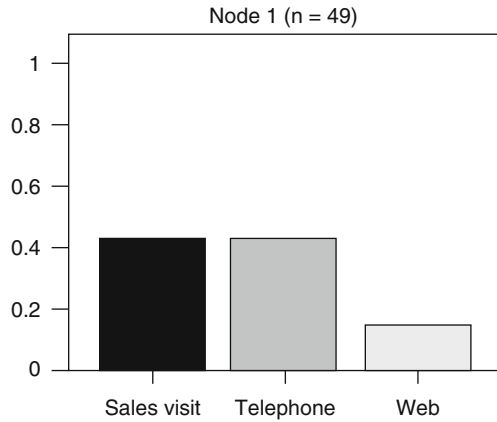
The margin plot indicates the plotting of two attributes, revenue and product cost, at a time. The blue color indicates the experiential data. The red one represents the mean imputed data. The red plot indicates missing distribution of one feature. The blue box is the distribution of attributes which is present. The red box plot and blue box plot are equal, the missing values in the data set is MCAR (Missing Continuously At Random). The above chart represents that the values are not missing Continuously At Random. Hence mean imputation is performed for missing values. The next thing is, measuring whether the imputed value is good or bad. The `xyplot()` and `densityplot()` functions are used to plot, compare, and verify our imputations. The sales order method type is plotted in Fig. 14.5.

The number of samples used in this work is 49. There are three types of order method type used in this sample.

- Sales visit to a particular site
- Through telephone
- Through internet or web

From these observations, it is found that the sales order method type 1 and 2 were used equally which is 40% whereas type 3 (Through Internet) was used by people which are less than 20%. The retail marketing data set is downloaded from the internet and 50 samples were taken into consideration. The data present in the dataset is both numeric and nonnumeric. The classification algorithm `rpart` is used for classification with order method type and unit sale price of mean imputed data and is shown in Fig. 14.6. The retailer country is a feature in this data set and is

**Fig. 14.5** Sales order method type plot



**Fig. 14.6** Decision tree for order method type

used for classification and is shown in Fig. 14.7. The plot of unit sale price among countries is shown in Fig. 14.8.

GLMs are most commonly used to model binary data or countable data to predict a class accurately. The output of the function lies between 0 and 1. The glm model is built with two attributes, namely order method type and retailer country of mean imputed data with binomial family link logit.

The general linear model of glm is

$$\hat{Y} = \beta_0 + \beta_1 X \tag{14.1}$$

where

$\hat{Y}$  is predicted variable,  
 $\beta_0$  constant value,

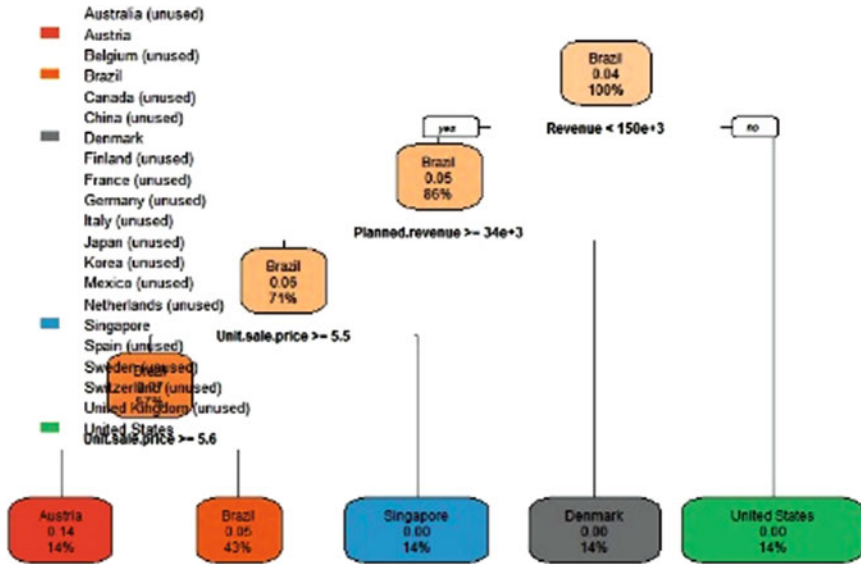
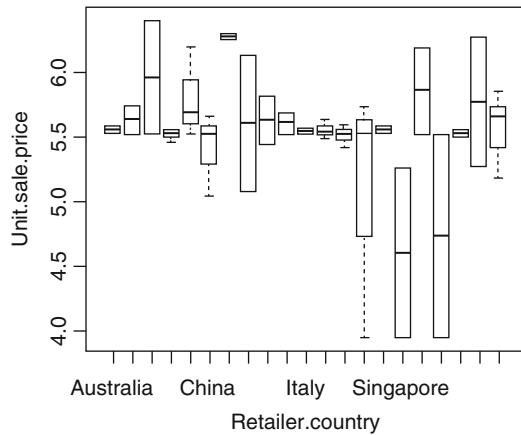


Fig. 14.7 Retailer country tree

Fig. 14.8 Plot of unit sale price among countries



$\beta_1$  coefficient weight, and  $X$  is variable

The glm procedure can deal with a large number of variables, including a numeric and nonnumeric. The nonnumerical values are converted into numerical value during the processing. The glm's are standardized with a mean value of 0 and standard deviation of 1. The GLM equation with standardized  $\beta$ s is:

$$Z_Y = \beta_0 + \beta_1 Z_{x1} + \beta_2 Z_{x2} + \dots + \beta_k Z_{xk} \quad (14.2)$$

Z is calculated by dividing the regression coefficient with standard error. If the z-value is big, the true regression coefficient is not 0 and the X-variable changes.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{14.3}$$

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as taking the square of residuals and summing all observations. The deviance is used to assess the goodness of fit. The higher number of deviance says that there is a bad fit. The response variable is predicted by the null deviance values, by a model that includes only the intercept (grand mean) where as residual with inclusion of independent variables. The data model is pretty only when the null deviance and residual deviance is small. The model is fitted with `glm(formula = Order.method.type ~ Retailer.country, family = binomial(link = "logit"), data = meandata)`. The Deviance Residuals with `Order.method.type ~ Retailer.country` is shown in Table 14.1.

The final output for a GLM models displays.

- Call
- Residual
- Coefficient
- Dispersion parameter
- Deviance values

The dispersion parameter for binomial family is taken to be 1, the null deviance is 66.925 on 48 degrees of freedom, and residual deviance is 65.550 on 28 degrees of freedom with AIC as 107.55. The Number of Fisher Scoring iterations is 4. The Fisher’s scoring algorithm is a method for providing solution to maximum likelihood problems on numerical values. This is derivative of newton’s method to perform fit. The fit is performed with `glm(formula = Order.method.type ~ Retailer.country + Revenue + Planned.revenue + Product.cost + Quantity, family = binomial(link = "logit"), data = meandata)` and is shown in Table 14.2.

**Table 14.1** Deviance residuals with `Order.method.type ~ Retailer.country`

Min	1Q	Median	3Q	Max
-1.4826	-1.1774	0.9005	1.1774	1.1774

**Table 14.2** Deviance residuals

Min	1Q	Median	3Q	Max
-1.8122	-1.1774	0.4454	1.0957	1.4579

The ANOVA is an analysis of variance which is performed for comparing three or more mean value. It is used to determine if there exist any relationships among variables or any difference between groups in sample data. ANOVA can be used with *t*-tests, Regression, and Chi Square, whether the mean of a variable is less than, greater than, or equal to a specific value. Usually, the known value which is present in the database is a population mean. The Null hypothesis says there is no significant difference between the sample mean and the population mean.

The ANOVA is performed for glm mean model with Chi square test and binomial model link logit. The Analysis of Deviance is shown in Table 14.3. The response attribute used is order method type.

The ANOVA is performed for linear regression model with unit sale price and unitprice as features. The Analysis of Variance is shown in Table 14.4. The response variable is unitsaleprice.

The null hypothesis for ANOVA states that all population means are exactly equal. The significant level of 0.05 shows the mean population or sample used in this work is good. A significance level of 0.05 indicates a 5% risk. This means that a difference exists in the sample mean when there is no actual difference found. This shows the sample means will differ a bit.

The dispersion parameter for binomial family is taken to be 1, the null deviance is 66.925 on 48 degrees of freedom, and residual deviance is 57.901 on 24 degrees of freedom with AIC as 107.9. The Number of Fisher scoring iterations is 17. It is shown that the null variance in the first model and second model is same whereas the residual deviance of second model is decreased with decreased degrees of freedom. The significant reduction in residual deviance shows the improvement of goodness of fit and is shown in Table 14.5. The comparison results are shown in Fig. 14.9.

**Table 14.3** Response-Order.method.type

Response	Deg of freedom	Deviance	Resid.Df	Resid.Dev	Pr(>Chi)
NULL		48	66.925		
Retailer.country	20	1.3752	28	65.550	0.711

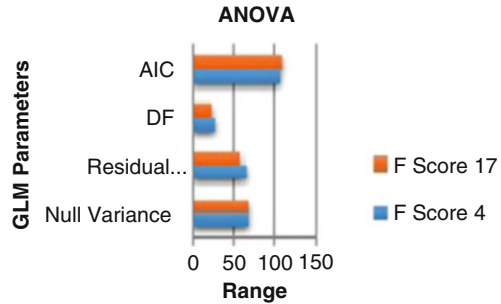
**Table 14.4** Response: Unit.sale.price

Response	Deg of freedom	Sum square	Mean square	F score value Pr(>F)
Residuals	48	12.181	0.25377	0.005

**Table 14.5** Comparison of GLM model

F score	Null variance	Residual deviance	AIC
4	66.295 (48 DF)	65.550 (28 DF)	107.55
17	66.925 (48 DF)	57.901 (24 DF)	107.9

**Fig. 14.9** Comparison of GLM Model



## 14.5 Conclusion

The important outcome of this work is choosing most appropriate missing value handling method. Mean imputation method was proposed to overcome the missing value. The Decision tree classification algorithm was performed. The experimental results show that there is a good improvement in the accuracy of classification algorithm after imputing mean value in the dataset. The goodness of fit of mean imputed value is also analyzed. The future work focuses on implementing other machine learning algorithms with increased results.

## References

1. I. Pratama, A. Erna Permanasari, I. Ardiyanto, R. Indrayani, A review of missing values handling methods on time-series data, in *International Conference on Information Technology Systems and Innovation (ICITSI)* (IEEE, Piscataway, NJ, 2016), INSPEC Accession Number: 16675571
2. M. Moore, J.M. Carpenter, A decision tree approach to modeling the private label apparel consumer. *Mark. Intell. Plan.* **28**(1), 59–69 (2010)
3. K. Maheswari, P. Packia Amutha Priya, Predicting customer behavior in online shopping using SVM classifier, in *IEEE International Conference on Intelligent Techniques in Control, Optimization, Signal Processing, INCOS'17*, 1 Mar 2018
4. K. Maheswari, P. Packia Amutha Priya, Analysis and implementation of text mining for different documents. *Int. J. Scient. Res. Sci. Technol.* **3**(5), 109–113 (2017). ISSN: 2395-6011
5. K. Maheswari, P. Packia Amutha Priya, Classification of twitter data set using SVM and KSVM. *Int. J. Pure Appl. Math.* **118**(7), 675–680 (2018). ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version), Scopus Indexed
6. K. Maheswari, Improving accuracy of sentiment classification analysis in twitter data set using knn. *Int. J. Res. Anal. Rev.* **5**(1), 422–425 (2018). E ISSN: 2348-1269, Print ISSN: 2349-5138, UGC Approved Journal
7. R. Tinabo, Decision tree technique for customer retention in retail sector, in *International Conference on Integrated Computing Technology INTECH 2011* (Springer, Berlin, 2011), pp. 123-131
8. M. Rafiqul Islam, M. Ahsan Habib, A data mining approach to predict prospective business sectors for lending in retail banking using decision tree. *Int. J. Data Min. Knowl. Manag. Process (IJDKP)* **5**(2), 13–22 (2015)

9. B.N. Patel, S.G. Prajapati, K.I. Lakhtaria, Efficient classification of data using decision tree. *Bonfring Int. J. Data Min.* **2**(1), 6–12 (2012)
10. L. Li, X. Zhang, Study of data mining algorithm based on decision tree, in *2010 International Conference on Computer Design and Applications*, 5 Aug 2010, INSPEC Accession Number: 11523965
11. R. Senapati, K. Shaw, S. Mishra, D. Mishra, A novel approach for missing value imputation and classification of microarray dataset. *Proc. Eng.* **38**, 1067–1071 (2012)
12. R. Houari, A. Bounceur, T. Kechadi, T. Abdelkamel, R. Euler, A new method for estimation of missing data based on sampling methods for data mining, in *Advances in Intelligent Systems and Computing (AISC)*, vol. 225, (Springer, Cham, 2012), pp. 89–100
13. P.R. Houcka, S. Mazumdarb, E. Hartin, Classification of missing values handling method during data mining: review. *Sigma Epsilon* **21**(2), 49–60 (2017). ISSN: 0853-9103
14. Y.-Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**(2), 130–135 (2015)
15. S. Agarwal, G.N. Pandey, M.D. Tiwari, Data mining in education: data classification and decision tree approach. *Int. J. e-Education, e-Business, e-Management and e-Learning* **2**(2), 140–144 (2012)
16. C. Kishor Kumar Reddy, B. Vijaya Babu, A survey on issues of decision tree and non-decision tree algorithms. *Int. J. Artif. Intell. Appl. Smart Dev.* **4**(1), 9–32 (2016)
17. X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, et al., Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**(1), 1–37
18. H. Yang, S. Fong, Optimized very fast decision tree with balanced classification accuracy and compact tree size, in *2011 Third International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA)*, 24–26 Oct 2011, pp. 57–64

# Chapter 15

## Task Identification System for Elderly Paralyzed Patients Using Electrooculography and Neural Networks



S. Ramkumar, G. Emayavaramban, K. Sathesh Kumar,  
J. Macklin Abraham Navamani, K. Maheswari, and P. Packia Amutha Priya

### 15.1 Introduction

Communication based on eye movements was playing a vital role in developing communication devices for the patients with amyotrophic lateral sclerosis and other motor neuron degenerative diseases like quadriplegia, Guillain-Barre syndrome, spinal cord injury, and hemiparesis. Such types of diseases attack all the controllable movements including the speech, writing, walk, etc., except the eye movement activities. They need some assist to move from one place to other. Statistical survey showed that motor neuron diseases were increased day by day and reached 15–18% of the population. People affected by such diseases are also in progress. To avoid the condition, we need help from rehabilitation device to overcome the biological channel in natural way. Recently many works based on EOG-based HCI have been taken place to reestablish the communication channels in the absence of biological channels for people with severe motor disabilities. Currently some of the input devices used for communication were mouse, computer, keyboard, touch screen, touchpad, and track ball. The following devices need manual control and cannot be controlled by the people with disability. So the need of alternative method of communication between man and machines to communicate with caretakers. One

---

S. Ramkumar (✉) · K. Sathesh Kumar · K. Maheswari · P. Packia Amutha Priya  
School of Computing, Kalasalingam Academy of Research and Education, Virudhunagar, India  
e-mail: [s.ramkumar@klu.ac.in](mailto:s.ramkumar@klu.ac.in)

G. Emayavaramban  
Department of Electric and Electronic Engineering, Karpagam Academy of Higher Education,  
Coimbatore, India

J. Macklin Abraham Navamani  
Department of Computer Applications, Karunya Institute of Technology and Sciences,  
Coimbatore, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_15](https://doi.org/10.1007/978-3-030-19562-5_15)

151



such technology was developing rehabilitation device that helps the disabled person to control the ecosystems and exchange the thoughts more efficiently. These devices were encouraging the persons to perform tasks normally by associating help from technology. A small potential that appear in between the front and back of the eye is called electrooculogram. The technique of making communication between man and machine is called HCI. By combining these two methods it created new pathway for the people with elderly disabled by creating the rehabilitative aids. By making this combination EOG-based HCI plays a vital role for developing assistive device for the people with motor neuron disease [1–5].

Some of the EOG-based interfaces created by several researchers are eye reading system [6], EOG speller [7], security system [8], speech interaction system [9], multimedia control system [10], robotic wheelchair [11], indoor positioning [12], cursor controller [13], DC motor controller [14], eye gestures [15], virtual keyboard [16], and hospital alarm system [17]. In this experimentation we compared right-handers performance with left-handers to analyze the chances of creating nine states HCI for disabled individuals to convey their thoughts without some others help.

## 15.2 Previous Work Done

Many techniques have been already available to execute the rehabilitation devices with the help of eye movements. Some of the necessary studies were explained below. Tanguksant et al. (2012) designed virtual keyboard using voltage threshold algorithm. Signals were collected in both horizontal and vertical eye movement tasks by placing six electrodes nearby eyes. Collected signals were classified using voltage threshold algorithm. The proposed methodology shows an average accuracy of 95.2% with a speed of 25.94 s/letter [18]. Swami Nathan et al. (2012) designed virtual keyboard for motor-disabled people using signal-processing algorithm for both vertical and horizontal eye movements and acquired an improved accuracy of 95.2% with typing speed of 15 letters/min [19]. Souza and Natarajan (2014) designed EOG-based interface for elderly disabled using nonparametric method. Data were collected from 40 subjects and applied to nonparametric method and classified with Elman network and acquired the mean classification accuracy of 99.95% for both vertical and horizontal tasks [20]. Pratim Banik et al. (2015) created virtual keyboard for physically disabled persons by selecting the key using EOG signals collected from five subjects. Collected signals were sending to microcontroller through serial communication. Graphical user interface was designed to collect the output. The system obtained an average classification accuracy of 95%, with average button selection time of 4.27 and 4.11 s for selecting ten buttons respectively [21]. Rakshit et al. (2016) developed assistive device for speech disabled due to brain stroke or spinal cord injury. Twelve subjects were participated in the study. Collected signals were applied with power spectral density to extract the features. Features were classified using support vector machine with multilayer perceptron kernel function with an average accuracy of 90% [22]. Hossain et al. (2017) developed

cursor controller for disabled using vertical and horizontal eye movement tasks by using instrumental amplifier AD620 and operational amplifier LM741 for four tasks. SVM and LDA classifiers were used to classify the online data [23]. Ramkumar et al. (2017) developed EOG-based HCI for ALS patients using nine movements from six subject. Features were extracted by using Euclidean norm and trained with neural network technique to categorize the signals. The system shows an average classification accuracy of 87.72% using dynamic network [24]. The literature survey shows that parametric and nonparametric methods were more suitable for obtaining the features using neural network classifier. Through this survey we concluded that neural classifier outperforms other classifiers used in the previous study. So we planned to conduct our study by using parametric method using neural network.

## 15.3 Methods

### 15.3.1 Experimental Protocol

Master data set were acquired from ten healthy participants for 2 s with five electrodes system and ADT26 bio amplifier. Signals were divided into two Hz from 0.1 to 16 Hz and sampled at 100 Hz. The Signal acquisition and preprocessing were previously enlightened by same authors in his earlier work [25]. Raw signal acquired from subject S4 is shown in Fig. 15.1.

### 15.3.2 Feature Extraction

Features were extracted from cleaned signal using periodogram method. Periodogram states that it is a mathematical tool to calculate the differences in the periodic signals. It calculates the significance of different frequencies in time-series data to recognize any essential periodic signals. The feature extraction method contains the following steps.

*Step 1:* S = Sample data of two channel EOG signal for 2 s.

*Step 2:* S was partitioned into 0.1 s windows.

*Step 3:* Band-pass filters were applied to extract eight frequency bands from S.

*Step 4:* Apply Fourier Transform to the frequency band signal to extract the features.

*Step 5:* Extract the absolute values and sum of the power values is extracted.

*Step 6:* Take the average values from each frequency band.

*Step 7:* Repeat steps 1–6 for each trial for all tasks and for ten subjects.

*Step 8:* Sixteen features were obtained for each one tasks per trial shown in Fig. 15.2 and repeat for ten such trials for 11 tasks.

*Step 9:* 110 data samples were obtained from each subject individually to train and test the neural network.

*Step 10:* Repeat steps 1–9 for ten subjects to collect master dataset.

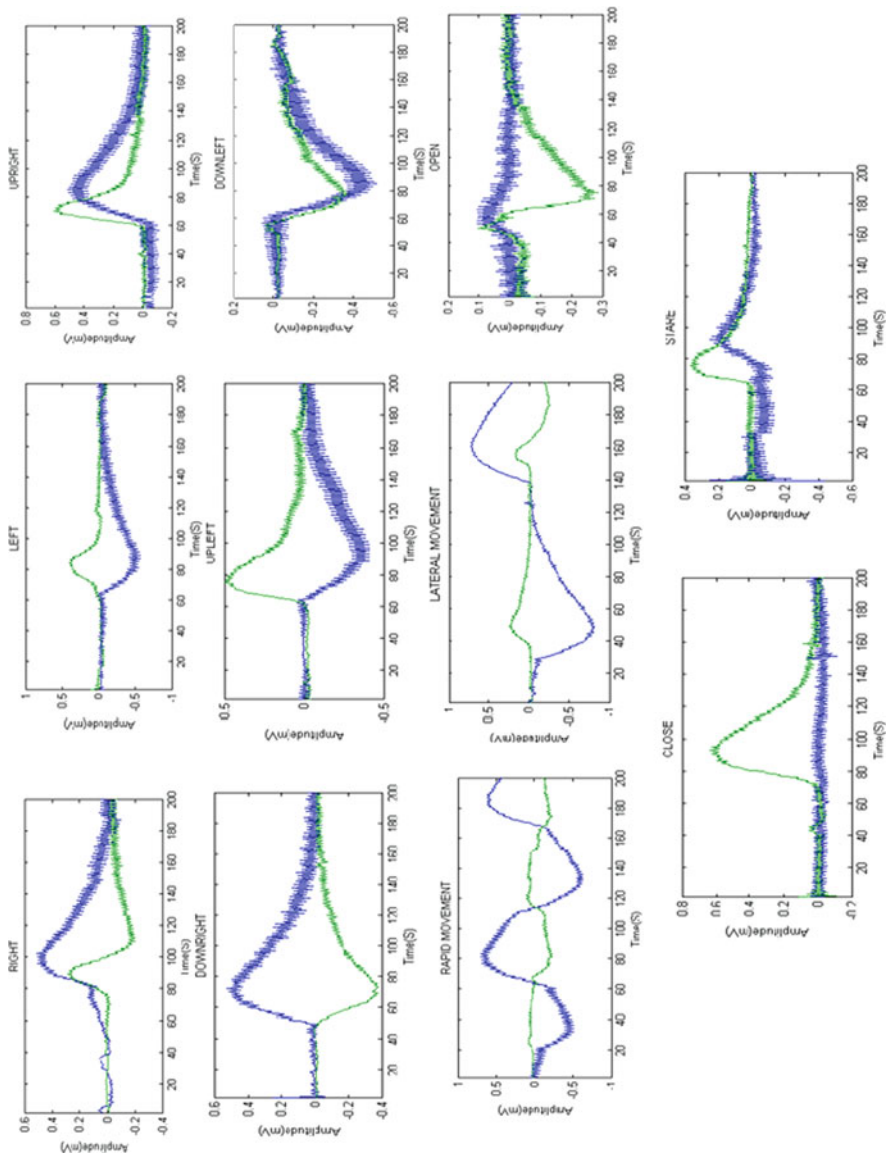


Fig. 15.1 Raw EOG signal acquired from subject S4 for 11 tasks

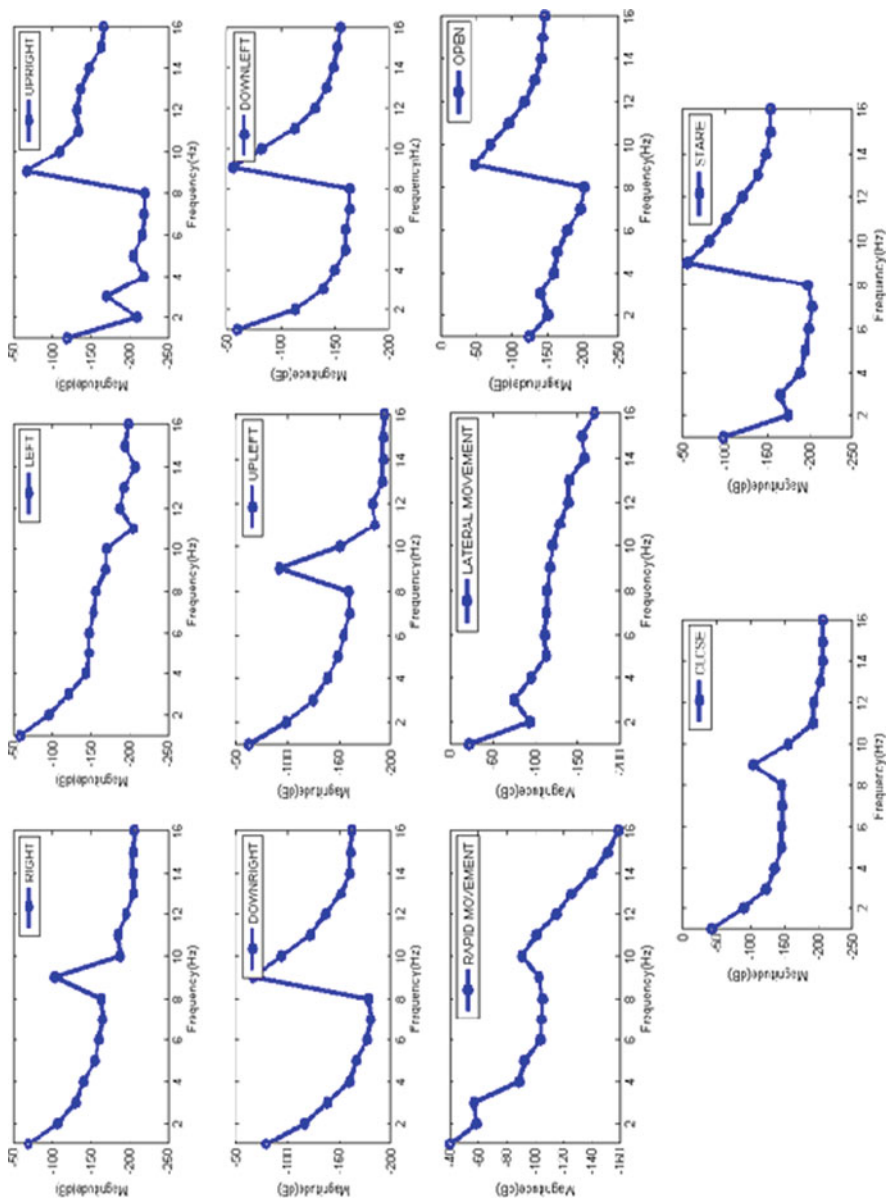


Fig. 15.2 Feature extracted signal for 11 different eye movements for subject S4 using periodogram

## 15.4 Classification Method

Prominent features selected from abovementioned steps were applied to neural network to categorize the signals. In this study we focused on probabilistic neural network (PNN). PNN was based on statistical principles derived from Bayes decision strategy and Gaussian kernel-based estimators of probability density function (PDF). Every input neuron communicates to an element of an input vector and is fully coupled to the  $n$  hidden layer neurons. Again, each of the hidden neuron is fully connected to the output neurons. Input layer simply feed the inputs into the classifier. Hidden layer will be able to compute the distance between the input vector and the training input vectors using Bayes decision strategy and Gaussian kernel function, and produce PDF features whose elements show the closeness between the input data points and the training vector points. As the last step, output layer will pick the summing output of the hidden layer with weight and apply Bayes decision learning; it will select the most of the probabilities on the hidden layer and also supply a one for that class and a zero for the other classes [26–30]. The network design used during this experiment is shown in Fig. 15.3.

## 15.5 Outcome of the Study

To evaluate the proposed method finally we designed ten network models to categorize the signal acquired for five left-handers and five right-handers through AD Instrument. T26 labchart was connected to a computer with the help of wires. Five electrodes were placed near to the eye to measure the horizontal and vertical movements. List of different subjects participated in the study is shown in Table 15.1 and their age was between 20 and 36. We analyzed the individual subjects performance throughout the study to analyze the performance.

Table 15.2 shows the average classification performance of left-hander subjects using periodogram features with PNN model. Table 15.2 particularly shows the result of mean, minimum, maximum, testing time and training time to determine

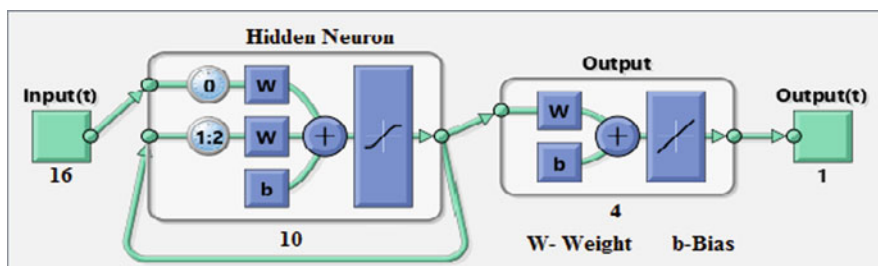


Fig. 15.3 Probabilistic neural network

**Table 15.1** List of different subjects participated in the study

Subjects	Left-handers	Right-handers
	S1, S2, S3, S4, and S5	S6, S7, S8, S9, and S10

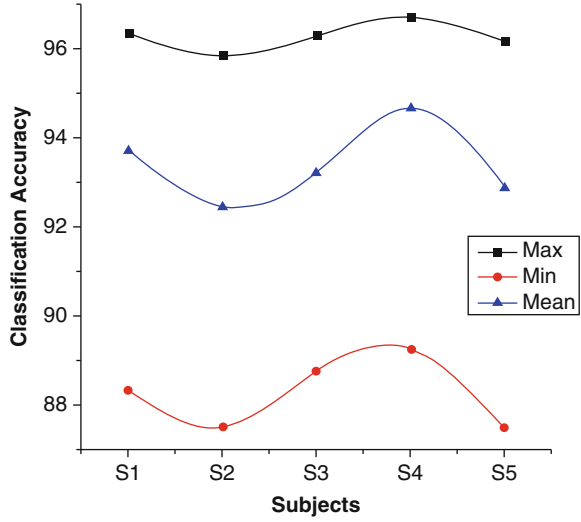
**Table 15.2** Average performance accuracy for left-handers using periodogram and PNN

Subjects	Mean training time (s)	Mean testing time (s)	Average performance (%)			
			SD	Max	Min	Mean
S1	13.26	0.68	1.38	96.35	88.33	93.70
S2	13.42	0.64	1.76	95.83	87.50	92.45
S3	13.48	0.63	1.67	96.30	88.76	93.21
S4	13.59	0.61	1.85	96.70	89.24	94.67
S5	13.37	0.62	1.98	96.17	87.50	92.88

the performance of the left-handers participated in this study. From Table 15.2 we identify that subject S4 performance was marginally high compared with other left-handed subjects participated in this experiment with an maximum classification accuracy of 96.70% and minimum classification accuracy of 89.24% and average classification accuracy of 94.67% with a training and testing time of 13.59 and 0.61 s for ten trials per tasks. Next maximum classification accuracy was obtained for Subject S1 with maximum classification accuracy of 96.35% and minimum classification accuracy of 88.33% and average classification accuracy of 93.70% with a training and testing time of 13.26 and 0.68 s for ten trials per tasks. The minimum classification accuracy of left-handed subject was obtained for Subject S2 with maximum classification accuracy of 95.83% and minimum classification accuracy of 87.50% and average classification accuracy of 92.45% with a training and testing time of 13.42 and 0.64 s for ten trials per tasks. From the individual performance stated in Table 15.2 we found that Subject S4 performance was appreciable compared with other left-handers participated in this study which is shown in Fig. 15.4.

Table 15.3 express the average classification performance of right-hander subjects using periodogram features with PNN model. Table 15.3 particularly shows the result of mean, minimum, maximum, testing time and training time to determine the performance of the right-handers participated in this study. From Table 15.2 we identify that subject S9 performance was marginally high compared with other right-handers participated in this experiment with a maximum classification accuracy of 95.30% and minimum classification accuracy of 90.00% and average classification accuracy of 92.10% with a training and testing time of 13.88 and 0.72 s for ten trials per tasks. Next maximum classification accuracy for right-hander was obtained for Subject S8 with maximum classification accuracy of 94.16% and minimum classification accuracy of 86.80% and average classification accuracy of 91.78% with a training and testing time of 13.76 and 0.71 s for ten trials per tasks. The minimum classification accuracy of right-handed subject was obtained for Subject S7 with maximum classification accuracy of 91.64% and minimum classification accuracy of 86.67% and average classification accuracy of 90.39% with a training

**Fig. 15.4** Performance of left-hander using periodogram with PNN



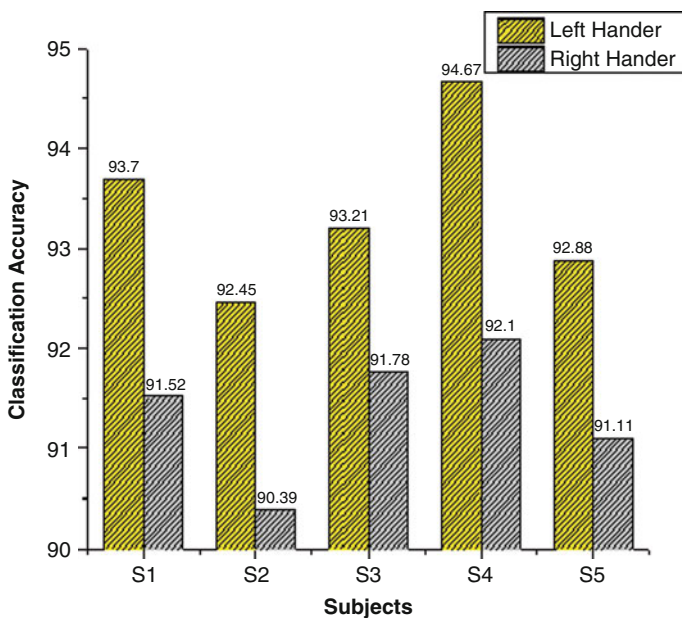
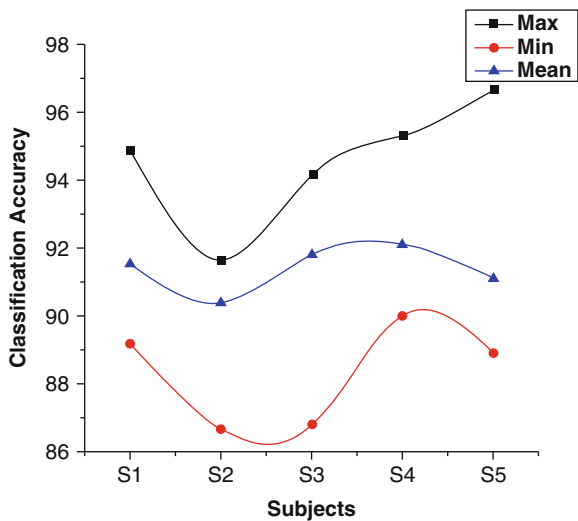
**Table 15.3** Average performance accuracy for right-handers using periodogram and PNN

Subjects	Mean training time (s)	Mean testing time (s)	Average performance (%)			
			SD	Max	Min	Mean
S6	13.84	0.69	2.19	94.89	89.17	91.52
S7	13.92	0.70	1.98	91.64	86.67	90.39
S8	13.76	0.71	1.91	94.16	86.80	91.78
S9	13.88	0.72	2.03	95.30	90.00	92.10
S10	13.76	0.74	2.22	96.67	88.90	91.11

and testing time of 13.92 and 0.70 s for ten trials per tasks. From the individual performance stated in Table 15.3 we found that Subject S9 performance was appreciable compared with other right-handers participated in this study which is shown in Fig. 15.5.

From Tables 15.2 and 15.3, we concluded that Subject S4 and S9 performances were high compared with all the left-handers and right-handers who took part in this study using periodogram features with PNN model as is illustrated in Fig. 15.6. After individual comparison between left-hander performance and right-hander performance, we concluded that left-handed subjects performance were marginally greater than right-handed subjects participated in this setup and also we found that left-handed subjects took less time during the training period than that of right-handed subjects participated in this study. The average results obtained from this experiment was exceed compared with our previous study [25] in terms of classification accuracy. The reason we identified was, we divide the ten subjects (five left-handers, five right-handers) equally and also we changed the feature extraction and classification techniques to analyze the performance. Through this experiment we found that making the HCI was possible using the left-handed subjects and also right-handed subject need some training to achieve this event.

**Fig. 15.5** Performance of right-hander using periodogram with PNN



**Fig. 15.6** Overall task classification for left-handers and right-handers using periodogram with PNN



## 15.6 Conclusion

The experiment was conducted with ten subjects (five left, five right) using bio amplifier by placing five electrodes to identify the task performed by the different subjects using periodogram with probabilistic neural network. From the study we found that average performance of left-hander subjects was appreciable compared to the right-hander subjects with an mean accuracy of 93.38% and 91.38%. Throughout the study we analyzed that all the left-handed subjects average performance were greater than that of the right-handed subjects participated in this experiment. From these analysis we concluded that creating HCI is possible by using abovementioned technique. In future we are planned to conduct this experiment in online phase to check the possibility of designing HCI.

## References

1. F. Fang, T. Shinozaki, Electrooculography-based continuous eyewriting recognition system for efficient assistive communication systems. *PLoS One* **13**(2), 1–20 (2018)
2. C. Mondali, Md. Kawsar Azami, M. Ahmadi, S.M. Kamrul Hasani, Md. Rabiul Islam, Design and implementation of a prototype electrooculography based data acquisition system, in *International Conference on Electrical Engineering and Information & Communication Technology*, 2015, pp. 1–6
3. A. Krolak, P. Strumiłło, Eye-blink controlled human-computer interface for the disabled. *Hum. Comput. Syst. Interact.* **60**, 123–133 (2009)
4. R. Hajare, M. Gowda, S. Jain, P. Rudraraju, A. Bhat, Design and development of voice activated intelligent system for elderly and physically challenged, in *International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques*, 2016, pp. 372–346
5. Y.M. Nolan, Control and communication for physically disabled people, based on vestigial signals from the body. PhD thesis, Paper submitted to Natl. Univ. Ireland, Dublin, 2005, pp. 7–18
6. A. Banerjee, A. Rakshit, D.N. Tibarewala, Application of electrooculography to estimate word count while reading text, in *International Conference on Systems in Medicine and Biology*, 2016, pp. 174–177
7. N. Barbara, T.A. Camilleri, Interfacing with a speller using EOG glasses, in: *International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 001069–001074
8. Md. Shazzad Hossain, K. Huda, S.M. Sadman Rahman, M. Ahmad, Implementation of an EOG based security system by analyzing eye movement patterns, in *International Conference on Advances in Electrical Engineering (ICAEE)*, 2015, pp. 149–152
9. M. Katore, M.R. Bachute, Speech based human machine interaction system for home automation, in *IEEE Bombay Section Symposium (IBSS)*, 2015, pp. 1–6
10. L. Li, X. Wu, Design and implementation of multimedia control system based on bluetooth and electrooculogram (EOG), in *International Conference on Bioinformatics and Biomedical Engineering*, 2011, pp. 1–4
11. A. Banerjee, S. Datta, P. Das, A. Konar, D.N. Tibarewala, R. Janarthanam, Electrooculogram based online control signal generation for wheelchair, in *International Symposium on Electronic System Design*, 2012, pp. 251–255

12. X. Li, D. Luo, F. Zhao, Y. Li, H. Luo, Sensor fusion-based infrastructure independent and agile real-time indoor positioning technology for disabled and elderly people, in *International Symposium on Future Information and Communication Technologies for Ubiquitous HealthCare (Ubi-HealthTech)*, 2015, pp. 1–5
13. B. Akan, A.O. Argunsah, A human-computer interface (HCI) based on electrooculogram (EOG) for handicapped, in *International Conference on Signal Processing and Communications Applications*, 2007, pp. 1–3
14. Z.-H. Wang, Hendrick, Y.-F. Kung, C.-T. Chan, S.-H. Lin, G.-J. Jong, Controlling DC motor using eye blink signals based on LabVIEW, in *International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 2017, pp. 61–65
15. M. Lin, G. Mo, Eye gestures recognition technology in human-computer interaction, in *International Conference on Biomedical Engineering and Informatics (BMEI)*, 2011, pp. 1316–1318
16. D.R. Lingegowda, K. Amrutesh, S. Ramanujam, Electrooculography based assistive technology for ALS patients, in *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2017, pp. 36–40
17. S. Venkataramanan, P. Prabhat, S.R. Choudhury, H.B. Nemade, J.S. Sahambi, Biomedical instrumentation based on electrooculogram (EOG) signal processing and application to a hospital alarm system, in *Proceedings of the Second International Conference on Intelligent Sensing and Information Processing*, 2005, pp. 535–540
18. W. Tangsuksant, C. Aekmunkhongpaisal, P. Cambua, T. Charoenpong, T. Chanwimalueang, Directional eye movement detection system for virtual keyboard controller, in *International Conference on Biomedical Engineering*, 2012, 1–5
19. D. Swami Nathan, A.P. Vinod, K.P. Thomas, An electrooculogram based assistive communication system with improved speed and accuracy using multi-directional eye movements, in *International Conference on Telecommunications and Signal Processing*, 2012, pp. 18–21
20. S.D. Souza, S. Natarajan, Recognition of EOG based reading task using AR features, in *International Conference on Circuits, Communication, Control and Computing*, 2014, pp. 113–117
21. P.P. Banik, Md. Kawsar Azam, C. Mondal, Md. Asadur Rahman, Single channel electrooculography based human-computer interface for physically disabled persons, in *International Conference on Electrical Engineering and Information Communication Technology (ICEE-ICT)*, 2015, pp. 1–6
22. A. Rakshit, A. Banerjee, D.N. Tibarewala, Electro-oculogram based digit recognition to design assistive communication system for speech disabled patients, in *International Conference on Microelectronics, Computing and Communications*, 2016, pp. 1–5
23. Z. Hossain, Md. Maruf Hossain Shuvo, P. Sarker, Hardware and software implementation of real time electrooculogram (EOG) acquisition system to control computer cursor with eyeball movement, in *International Conference on Advances in Electrical Engineering*, 2017, pp. 32–37
24. S. Ramkumar, K. Sathesh Kumar, G. Emayavaramban, A feasibility study on eye movements using electrooculogram based HCI, in *International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 380–383
25. S. Ramkumar, K. Sathesh Kumar, G. Emayavaramban, EOG signal classification using neural network for human computer interaction. *Int. J. Comput. Theory Appl.* **9**(24), 223–231 (2016)
26. T. Gandhi, B.K. Panigrahi, S. Anand, A comparative study of wavelet families for EEG signal classification. *Neurocomputing* **74**, 3051–3057 (2011)
27. M. Hariharan, M.P. Paulraj, S. Yacob, Time-domain features and probabilistic neural network for the detection of vocal fold pathology. *Malaysian J. Comput. Sci.* **23**(1), 60–67 (2010)
28. M. Hariharan, M.P. Paulraj, S. Yacob, Detection of vocal fold paralysis and oedema using time-domain features and probabilistic neural network. *Int. J. Biomed. Eng. Technol.* **6**(1), 46–57 (2011)
29. D.F. Specht, Probabilistic neural networks. *Neural Networks* **3**(1), 109–118 (1990)
30. T. Sitamahalakshmi, A. Vinay Babu, M. Lagadesh, K.V.V. Chandra Mouli, Performance of radial basis function networks and probabilistic neural networks for Telugu character recognition. *Global J. Comput. Sci. Technol.* **11**, 9–16 (2011)

# Chapter 16

## A Software-Defined Networking (SDN) Architecture for Smart Trash Can Using IoT



T. Vairam, S. Sarathambekai, and D. Vigneshwaran

### 16.1 Introduction

Networking as we know it today has been started in the late 1960s and early 1970s. From then, the evolution of network achieved a great height starting from LAN, WAN, cellular network, WLAN to VAN, Adhoc Network, Mobile network, Wireless Sensor Network. There are lot more communication technologies which are added in the queue. Now we are in the era of Internet of Things where any object in the world can act smart according to its purpose on which it has been made and make that object to collect information and transform to the other side of the network through internet. The thought of IoT is simple, the requirement of IoT is not restricted to the frontier and it becomes obligatory in day-to-day living and changes the whole prototype of heritage technology [1]. The object is embedded with the network interface and enables them to communicate with the users. Each object is identified through its unique identification number or IP address. Prior to the IoT era the user can obtain their information only through service provider but now they can obtain the information from any object which is provided with computing and internet facility. The architecture of IoT should be properly designed so that the IoT application will function efficiently. Long-Term Evolution (LTE), ZigBee, Wi-Fi, Z-wave, and Bluetooth are the protocols through which the communication is being accomplished among IoT devices. IoT devices are heterogeneous in nature. Each and every device in an IoT application has their own procedure and rules. The process, services, and hardware everything is predetermined according to its requirement of its respective application. The infrastructure of the IoT environment is fixed in panorama. Further modification of the process or replacement of any

---

T. Vairam (✉) · S. Sarathambekai · D. Vigneshwaran  
Department of Information Technology, PSG College of Technology, Coimbatore, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_16](https://doi.org/10.1007/978-3-030-19562-5_16)

163

device's function related to the application will not be performed easily since, this will influence the complete network infrastructure.

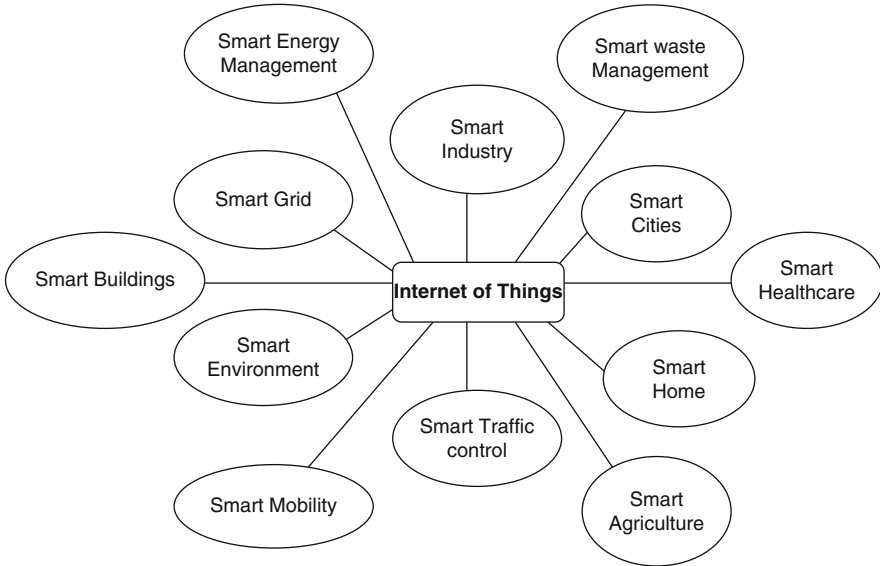
It also requires dynamic updation of the network and it has to be done as quick as possible without spending the amount of cost. To conquer this challenge, the design of the programmable network has been initiated which is named as software-defined network (SDN). Software-defined network (SDN) is a next invention of Internet technology which split the functionality of data plane and control plane [2]. The function of control plane is taken away and is placed in a centralized location by means of the server called controller. SDN is a promising technology that meets the demand of IoT as it needs to communicate with different network which is the heterogeneity in nature [3]. SDN also provides central control across the network. As every research stated in their article [3–5], the functionality of the control plane and forwarding plane decoupled in the SDN process. By having central control, SDN helps to automate the network configuration process in an efficient manner. SDN architecture is a layered architecture which includes application layer, control plane layer, and data plane layer. It also includes two interfaces called northbound interface (NI) and southbound interface (SI) through which the layers are communicated. NI is responsible for providing an interaction between application plane and control plane, whereas SI is responsible for providing interaction between the data plane and control plane.

Designing SDN to the IoT applications will be useful thought which make the network configuration process easy. In this paper, we developed the IoT infrastructure for garbage collection and also proposed a SDN architecture for the IoT application which helps to get better performance of data promoting to the processing center in a well-organized manner. In this proposed model, the cloud-based architecture is integrated into the networks along with various software and sensors. The objective of this paper is to propose SDN architecture for smart IoT trash bin to simplify the process of data transferring and to propose a solution for garbage collection system across the city.

The organization of the paper comprises Sect. 16.2 where various related works pertaining to IoT and SDN are discussed. Sections 16.3 and 16.4 have the proposed SDN architecture for smart trash can and proposed IoT-based smart trash can system, respectively. Section 16.5 narrates about the implementation details, and the conclusion is given in Sect. 16.6.

## 16.2 Related Works

These days, every human activities are knowingly or unknowingly updated in the internet, for example, their account details, travel details, and medicine details. Many applications are built up based on IoT requirement of industry or human which compose an object to labor cleverly by adding the flavor of Internet and computing facilities. Figure 16.1 shows the various fields where the IoT plays a major role and involves everyday individual's living actions that are notified by the IoT devices. Zeinab and Elmustafa presented the diverse IoT applications which



**Fig. 16.1** IoT application

eloquent each day of human life will be enhanced and linked with internet through IoT application [6]. Shyam et al. developed the model for waste collection system using IoT infrastructure [7]. In this model the data has been collected and forwarded using internet and also some kind of intelligent algorithms. This model is the dynamic in nature where the collected data are transmitted through internet and using an optimized algorithm the forwarded data are processed.

A novel SDN architecture was introduced which comprises a component named RSU cloud which includes the features of traditional RSU (Road Side Unit), micro scale datacenters, and SDN Controller [4]. Sibylle Schaller and Hood concussed on the SDN architecture, pointed out that the ONF (Open Networking Foundation) architecture working group was the pioneer of the SDN architecture [5]. Samaresh Bera et al. recognized that the conventional networks such as WAN and enterprise network does not have the services which offer support to millions of devices to monitor the surroundings, gather data according to its application, and transmit the collected data to the processing center via internet [8].

Sahoo et al. branded some restrictions in conventional networks as follows: espousing latest protocols in an existing network is very complicated; sustaining hardware from diverse dealer is not feasible due to the closed nature of operating system; and setting up the network infrastructure is costly [2]. SDON (Software-Defined Optical Network) is used for adding optical communication [9]. Sathishkumar et al. developed the IoT-based smart alert system for garbage clearance [10]. The level of garbage bin is monitored based on which the alert signal is given to web server. Then instantaneous cleaning of dustbin will be conceded

out with proper verification. The ultrasonic sensor was used to identify the level of garbage bin. Harika et al. implemented the smart garbage system which is an expensive one as it requires many modules along with the Arduino micro controller like GSM module, GPS module, WIFI module, etc. [11].

### 16.3 Proposed SDN Architecture

In this section we describe the SDN architecture for smart trash can using IoT. The proposed architecture is given in Fig. 16.2. The main components are SDN Controller, Open Flow, Data line and forward line, Truck and Trash can. The functionality of each component in described in Table 16.1.

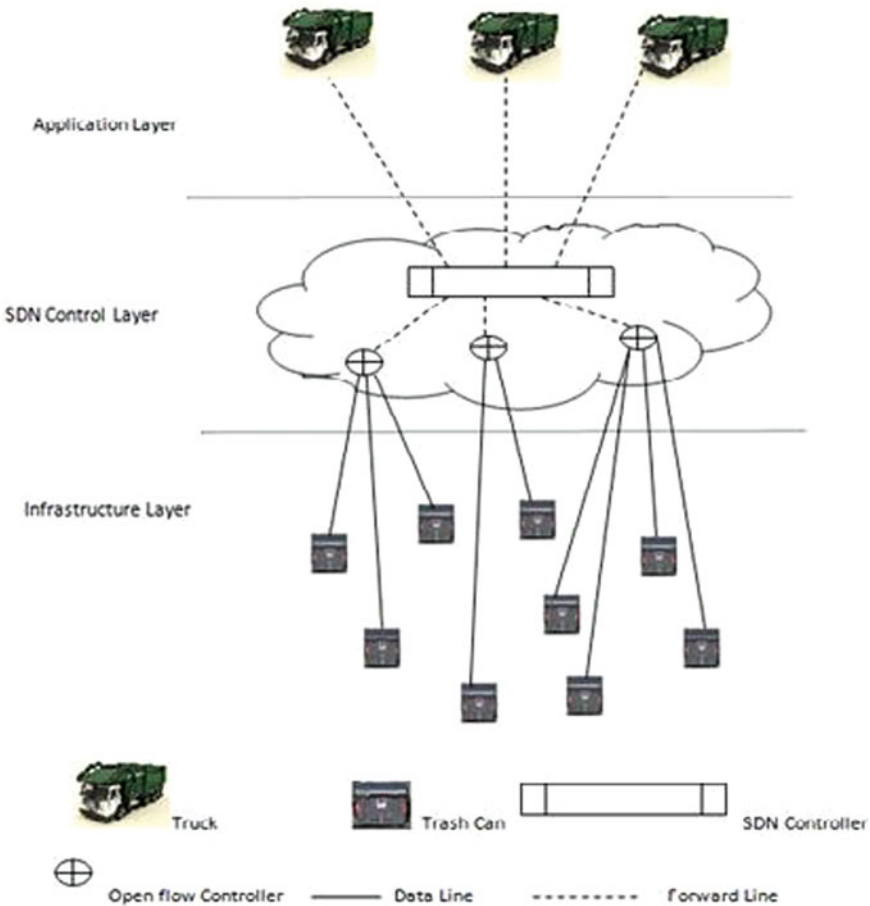


Fig. 16.2 SDN architecture for smart trash can

**Table 16.1** SDN architecture components

Component	Description
SDN Controller	SDN controller directly communicates with its network device which has been embedded into trash can through Open flow. SDN controller in turn forwards the data to the application layer. All communication between network device and the application must be done through SDN controller.
Open Flow	It is a protocol which helps SDN controller to separate the control line from the data line. Using Open Flow. SDN obtains information about the path through which the actual data has to be forwarded.
Data line and forward line	It represents the data transfer and control information transfer, respectively.
Truck	It is also considered as one of the component in SDN architecture because the application should be available in the mobile phone of the truck driver.
Trash can	This is provided with raspberry Pi, ultrasonic sensor and wifi module which will perform the tasks of identifying the trash can level and send that information to SDN controller.

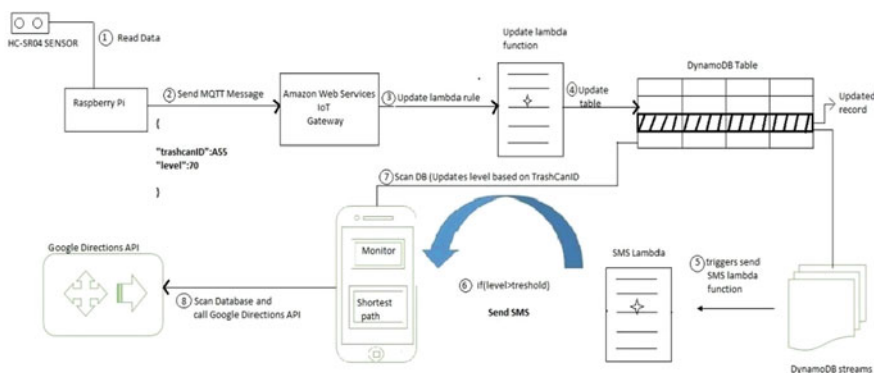
The benefits of this approach are: first, centralized management; due to this, the same truck can collect more than one trash can if they are located on the way to another trash can. Hence the time and fuel is reduced. Second, direct programmable capability, for finding shortest path, the algorithm can be changed easily based on its requirements. Third, scope for improvement is always possible without affecting the network infrastructure.

## 16.4 Proposed Smart Trash Can System

The population of metropolitan city or the urban cities is diversely spread out with some areas highly populated while others have comparatively less population. This makes the process of garbage collection to be a hectic task to the corporation people. Overflowing dustbins causes insanitary circumstance for the people and produces awful aroma around the atmosphere. This will affect the health of the human who live around the place. The current scenario for garbage collection is a very static solution for a very dynamic problem because the amount of trash we produce is not always the same. Hence we have proposed efficient way of collecting garbage's and provide the shortest path to enable the truck driver to reach the location quickly. The workflow of the smart trash can is discussed in this section. HC SR04 ultrasonic sensor and Raspberry Pi are attached to the trash can to measure the trash level. It sends the measured data as a MQTT (Message Queue Telemetry Transport) request through AWS IoT gateway. The Lambda rules are executed and values are updated in DynamoDB table. Then SMS is sent to client application using Amazon Simple Notification Service. The shortest path is calculated and Google map API (Application Programming Interface) shows the shortest path. The various

**Table 16.2** Proposed system components and its purpose

Component	Purpose
IoT module	The first step is to measure the amount of trash in the trash can. For this purpose, a Raspberry Pi board is interfaced with ultrasonic sensor (HC-SR04) [12] which is fixed to top of trash can.
AWS IoT	AWS IoT [13] receives the message transmitted by IoT module. It accesses the gateway and reads the incoming messages and forwards it to another end point.
AWS LAMBDA	AWS LAMBDA [14] is a server side event triggered function which will be executed whenever an event occurs.
AWS DynamoDB	It is a unstructured NoSQL database for efficient storage of nonstructured chunks of messages. Amazon Simple Notification Service (SNS) takes care of sending messages.
AWS SNS	Data that is received by Lambda function will be used to send a notification to the AWS SNS topic [15].
Client application	This app is used to monitor the overall status of all trash cans in the city.



**Fig. 16.3** Working process of smart trash can system

components used in the proposed system is given in Table 16.2. Figure 16.3 shows the overall working of Smart Trash Can system.

The advantages of the proposed smart trash can system are the level of the dustbin is updated then and then, the dustbin will be deployed according to the requirement of the people, cost reduction and resource optimization and improves environment quality.

## 16.5 Implementation and Results

The application is designed for android platforms. The trash level is stored in AWS cloud. The user can monitor the trash level of all trash cans across the city just by using the app and truck driver can find the shortest path using the app. Figure 16.4



Fig. 16.4 User interface

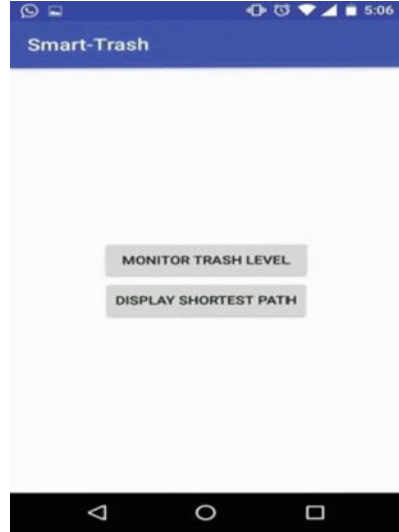
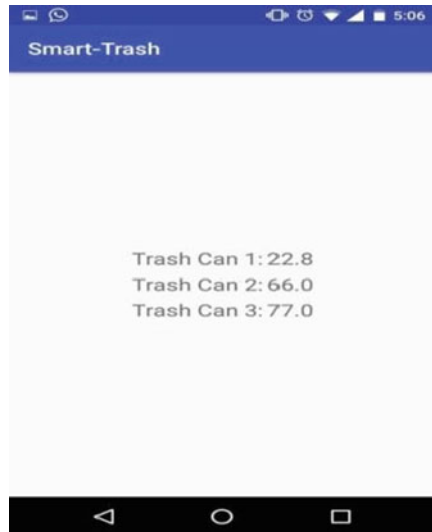


Fig. 16.5 Trash level in all trash cans



shows the user interface in client application. It has two tabs, one to monitor the trash level in different trash can and the figure shows the shortest path to the truck driver so that the filled trash can can be easily emptied.

On clicking the first tab, trash level in different trash cans is shown. With the available details truck driver can also predict when a particular trash can will get filled. The trash can level is shown in Fig. 16.5; the values represent the amount of trash filled in terms of percentage. Figure 16.6 shows the implementation of shortest

Fig. 16.6 Shortest path to reach the trash can

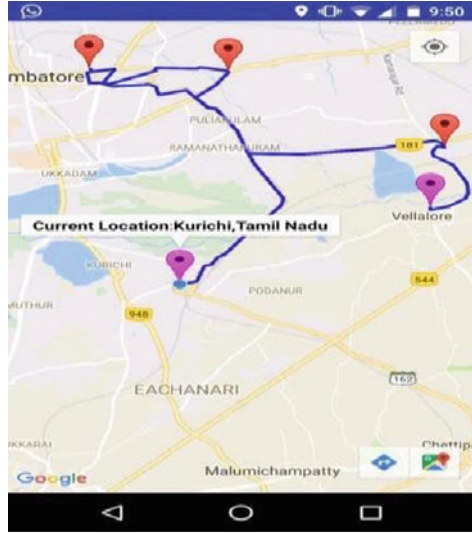


Fig. 16.7 Database structure

ID	LatLang	Location	Trashlevel
A44	11.020983,76...	Gandhipuram	90
A55	10.9572757,7...	Podanur	2
A33	10.9987351,7...	Singanalur	75
A66	10.9902127,7...	Ukkadam	66
A22	11.0200257,7...	Nava India	85

path algorithm and shows the optimized path to the user. Figure 16.6 shows the optimized path to the truck driver. Shortest path has been calculated using Dynamic Source Routing (DSR) [18]. Here the route is formed on- demand. The current location of the driver and the location of the filled trash can are given as input.

The algorithm finds the shortest path between the truck and filled trash can. The driver can also get to know if there is any other trash can that needs to be collected while reaching the destination. This can be achieved by accessing the client app. The shortest path is displayed with the help of Google app The Google Maps Directions API (2015) [17, 19].

Figure 16.7 shows the organization of database. The trash level changes dynamically in the database. Table 16.3 shows the main salient features of proposed work.

**Table 16.3** Salient feature of proposed work

Feature	Description
Cost	Very less hardware is required (Raspberry Pi board and Ultrasonic sensor). Also instead of deploying one board for each trash can, one board is enough for all trash can within a locality. This reduces the cost manifold.
Scalability	Scalability is very high. Since the data is sent to AWS dynamo DB which can be scaled up very easily.
Reliability	The reliability of AWS web services is very high. Over 2000 government web services use AWS including a lot of Defense services [16].

## 16.6 Conclusion

The novel Software-Defined Networking architecture is proposed for IoT-based smart trash can. A live tracking of trash can is monitored through android app. The app can also be used by the truck driver to identify the trash can and provide the shortest route to reach the trash can as soon as possible. The proposed SDN architecture will help to get better performance of the smart trash can system. IoT-based smart trash can helps the government in solving the critical task of maintaining the health and hygiene of citizens of nation. The proposed solution is cost effective and it is easily deployable and accurate. When deployed in large amounts the garbage collection system can reach higher levels of automation.

## References

1. S.K. Lee, M. Bae, H. Kim, Future of IoT networks: survey. *Appl. Sci.* **7**, 1072 (2017)
2. K.S. Sahoo, S.K. Mishra, S. Sahoo, B. Sahoo, Software defined network: the next generation internet technology. *Int. J. Wirel. Microw. Technol.* **2**, 13–24 (2017)
3. S.K. Tayyaba et al., Software-Defined Networks (SDNs) and Internet of Things (IoTs): a qualitative prediction for 2020. *Int. J. Adv. Comput. Sci. Appl.* **7**(11), 385–404 (2016)
4. M.A. Salahuddin, A. Al-Fuqaha, M. Guizani, Software-defined networking for RSU clouds in support of the internet of vehicles. *IEEE Internet Things J.* **2**(2), 144–197 (2015)
5. Y. Sibylle Schaller, D. Hood, Software defined networking architecture standardization. *Comput. Stand. Interfaces* **154**, 197–202 (2017)
6. K.A.M. Zeinab, S.A.A. Elmustafa, Internet of things applications, challenges and related future technologies. *World Scient. News* **67**(2), 126–148 (2017)
7. G.K. Shyam, S.S. Manvi, P. Bharti, Smart waste management using Internet-of-Things (IoT), IEEE Digital Library, in *Proceedings of International Conference on Computing and Communications Technologies (ICCCCT)*, 2017
8. Z. Samaresh Bera, S. Misra, A.V. Vasilakos, Software-defined Networking for in ternet of things: a survey. *IEEE Internet Things J.* **4**(6), 1 (2017)
9. A. Thyagaturu, A. Mercian, M.P. McGarry, M. Reisslein, W. Kellerer, Software Defined Optical Networks (SDONs): a comprehensive survey. *IEEE Commun. Surv. Tutorials* **18**(4), 2738–2786 (2016)
10. N. Sathishkumar, B. Vuayalakshmi, B. Jenifer Prarthana, A. Sankar, IOT based smart garbage alert system using arduino UNO, in *IEEE Region 10 Conference (TENCON)*, 2016

11. K. Harika, Muneerunnisa, V. Rajasekhar, P. Venkateswara Rao, L.J.N. SreeLakshmi, IoT based smart garbage monitoring and alert system using arduino UNO. *Int. J. Innov. Res. Comput. Commun. Eng.* **6**(2) (2018)
12. Interfacing HC-SR04, <https://electrosome.com/hc-sr04-ultrasonic-sensor-raspberry-pi/>, <http://www.instructables.com/id/HC-SR04-Ultrasonic-Sensor-With-RaspberryPi-2/>
13. AWS IoT, <http://docs.aws.amazon.com/IoT/latest/developerguide/IoT-sdk-setup.html>
14. AWS LAMBDA, <https://aws.amazon.com/documentation/sns/>
15. AWS SNS, <https://aws.amazon.com/documentation/lambda/>
16. AWS, <https://aws.amazon.com/government-education/government/>
17. Google Developers. Google Maps Directions API Usage Limits, <https://developers.google.com/maps/documentation/directions/usage-limits>
18. Shortest Path, <https://developers.google.com/optimization/routing/tsp#solving-tsp-with-or-tools>
19. The Google Maps Directions, <https://developers.google.com/maps/documentation/directions/>

# Chapter 17

## Modified K-Nearest Neighbor Fuzzy Classifier Using Group Prototypes and Its Application to Skin Segmentation



Priyadarshan Dhabe, Mukesh P. Chugwani, and Vaibhav B. Kahalekar

### 17.1 Introduction

Pattern recognition deals with recognition of patterns of various kinds. It is used to find regularities in input patterns, and based on similarity of patterns they are grouped into a predefined class. Pattern recognition systems are fed with the labeled set of input patterns, this set is called as training set (supervised learning), but when labels are not available other algorithms are used to discover the group created from previously unknown patterns (unsupervised learning). It has become one of the most promising fields to solve the difficult problems related with our day-to-day life. Mundane tasks like recognizing voice, face, and speech pattern are found to be very easy for humans but are equally difficult for the machines [1]. This field helps to represent a real-world pattern in computer memory and talks about using knowledge of previously known patterns to recognize class of unknown pattern. Classifier is a system that has capability to classify the patterns based on their feature vectors.

Skin segmentation which perfectly falls in the domain of pattern recognition is one of the important engineering fields. The objective of skin segmentation is to segregate skin and non-skin pixels, which further can be used to detect skin regions in an image. The detection and segmentation of skin regions in an image is widely used in many applications such as classification and retrieval of color images in multimedia applications, video surveillance, human motion detection, and gesture detection. Thus, it is an important task.

One of the simplest classifier we found in the literature is K-nearest neighbor (KNN) classifier proposed by Fix and Hodges [2] and discussed in [3, 4]. This classifier has potential drawbacks although it is simple. First drawback is that it

---

P. Dhabe · M. P. Chugwani (✉) · V. B. Kahalekar  
Vishwakarma Institute of Technology, Pune, Maharashtra, India  
e-mail: [priyadarshan.dhabe@vit.edu](mailto:priyadarshan.dhabe@vit.edu)

is using the training patterns as it is as its knowledge. On the contrary, knowledge should be represented in such a way that it can be used in great many situations [1], such representation is called generalized representation. Lack of generalized representation of knowledge reduces the scope of its applicability as well as reasoning accuracy. Second lacking we found is that such representation is equivalent to using a long hypothesis, since patterns which are present in the training set can be treated as a hypothesis learned by the system. But by Occam's razor principle [4], short hypotheses are preferred over long hypothesis. Long hypothesis, even though reasons correctly for training set, will not perform well over unseen patterns, since it can be a coincidence. As opposed to this, short hypothesis cannot be a coincidence and will definitely perform well even over the unseen patterns [4]. Thus, one can conclude that KNN cannot perform well for unseen patterns due to use of long hypothesis.

Third drawback is that the recall time per pattern for testing patterns will be very high and it is proportional to the size of the data set. This drawback reduces scope of applicability of KNN to the large data sets. Thus, we have decided to update KNN classifier for eliminating these drawbacks.

Literature survey shows that KNN is already updated by several researchers to improve its performance. In [5] KNN is updated for fuzzy reasoning using the concept of fuzzy sets and is used for recognizing driving environment. Its modified fuzzy version is used to predict the protein relative solvent accessibility in [6]. Later, the formal properties of KNN classification were established, a long line of investigation ensued including new rejection approaches [7], refinements with respect to Bayes error rate [8], distance weighted approaches [9, 10], soft computing [11] methods, and fuzzy methods [12, 13]. ITQON et al. in [14, 16] proposed a classifier, TFKNN, aiming at upgrading of distinction performance of KNN classifier and combining plural KNNs using testing characteristics.

Remaining part of this paper is organized as follows. Section 17.2 describes the modified KNN along with preparation of group prototypes and also describes about the fuzzy logic layer and fuzzy sets. Comparison of original and modified KNN is done using realistic data sets in Sect. 17.3. Experimental results are discussed in Sects. 17.4, and 17.5 concludes this work and references are cited at the end.

## 17.2 Modified Fuzzy KNN Classifier (MFKNN)

### 17.2.1 Group Prototypes

We want to modify the original KNN to overcome the drawbacks stated in Sect. 17.1 using group prototypes. A group prototype  $e$  is a prototype of a group of patterns from the same class and falling close to each other by a user-defined Euclidean distance  $d$ , where  $0 < d < 1$ . There can be multiple group prototypes from the same pattern class. Instead of using training patterns as it is to reason about the testing

patterns, these group prototypes are used. This small modification can eliminate all the drawbacks of original KNN stated earlier in this paper. The Algorithm stated below for group prototypes is taken from [15].

### 17.2.2 Algorithm to Calculate Group Prototypes

Let  $D$  be the set of  $K$ ,  $n$ -dimensional training patterns along with their class labels and belonging to  $P$  classes. Thus,  $D = \{(x_1, c_i), (x_2, c_i), (x_3, c_i), \dots, (x_k, c_i)\}$ , where  $c_i$  is the class label, where  $i = 1, 2, \dots, P$ . Each pattern  $x_q$  is a  $n$ -dimensional normalized vector as  $x_q = \{x_{q1}, x_{q2}, \dots, x_{qn}\}$

1. Initialize  $d$ , such that  $0 < d < 1$
2. Normalize the patterns  $x_q$  such that each component  $0 < x_{qj} < 1$ . For  $j = 1, 2, \dots, n$
3. while (! all patterns groups are created)
  - (a) Select any pattern  $v$  from the data set. For all the patterns belonging to the same class  $w$  of  $v$  do
    - Find Euclidian distance between  $v$  and  $w$
    - If Euclidian distance between  $v$  and  $w$  is  $\leq d$ , add the pattern  $w$  to the corresponding group of pattern  $v$ . Call this group of patterns as  $g$ .
4. Find the group prototype for each group  $g$ .

Let group  $g$  have  $m$  patterns, then the group prototype for it can be calculated as an average of the  $m$  patterns from the group  $g$ . Remove all these patterns in  $g$  from the data set and use updated data set and go to step 3.

### 17.2.3 Fuzzy Logic and Fuzzy Sets

One of the problems faced using the K-NN classifier is the time and space required for larger datasets. Since, it stores all the input patterns due to rote learning, it thus needs recall time and space proportional to the size of input data set. Due to the crisp character of the input membership given to make the classification, no information about how the data is distributed in the input space is provided. Another difficulty found is that the algorithm, once an input vector is classified, doesn't give information about the "strength" of membership to that class. These problems can be addressed using fuzzy sets theory. Modified Fuzzy KNN search is similar to simple KNN search. In simple KNN, every data point can belong to only one class which is the majority class in the K-nearest neighbor search. Whereas in fuzzy KNN, a data point can belong to multiple classes with different membership functions associated to these classes. The Fuzzy membership function used is defined as follows:

$$\mu_C(x) = 1 - \gamma * d_C$$

where  $\gamma$  is sensitivity parameter,  $\mu_C(x)$  is the fuzzy membership value of point with class “C” and  $d_C$  is the distance of point  $x$  with class “C” prototype. The sensitivity parameter introduces imprecision. According to trial and error the optimal value of  $\gamma$  is found to be 1.

### **17.2.4 Testing Phase of Proposed MFKNN**

1. Use these group prototypes for the testing purpose instead of using actual patterns in the training set like original KNN.
2. Find the value of  $d$  for which 100% classification rate is achieved. If not then reduce the value of  $d$ , which will result in creation of more number of group prototypes. Use group prototypes created such that they give 100% classification and then go for testing the patterns present in the testing set, i.e., for recognition.

## **17.3 Comments on Group Prototypes**

Since calculation of group prototypes is crucial task in MFKNN, we have decided to explain this step in greater detail for better understanding. If we choose less value of  $d$ , where  $0 < d < 1$ , then the number of group prototypes created will be more and vice versa. This step creates multiple prototypes of the same class. Multiple prototypes of a class give better performance than traditional method of one prototype per class, which is proven in [16]. But we need to carefully avoid the very small value of  $d$ , which will create number of group prototypes equals to number of patterns in the training set, i.e., it will create one group prototype from one pattern in the training set. In such cases and for  $d = 0$ , the MFKNN will be equivalent to original KNN. Thus, these situations must be avoided for better performance of MFKNN.

More the number of group prototypes, more will be the training and recall time of MFKNN. Thus, find out and use a value of  $d$ , such that it will create minimum number of group prototypes with 100% classification. This needs more experimentation. Finding the proper value of  $d$  satisfying above criterion is the training of MFKNN. The appropriate value of  $d$  for different data sets will be different.



## 17.4 Experimental Results

For comparing performance of Original KNN and MFKNN we have collected Skin Segmentation dataset from UCI repository. It consists of 245,057 samples, of which 58,059 are skin samples and remaining are non-skin samples.

The whole dataset is divided into two sets, i.e., training and test sets. The training set contains 70% of the data while the test set contains 30% of the data. When we compare performance of original KNN and MFKNN using this data set it is clear from Table 17.1 that the recognition rate is 87.7 and the time required to recognize testing set is for  $k = 1$ , is 55.11 min. If the same data is given to MKNN it creates 1211 prototypes, which are less than the number of training patterns, i.e., 1440 and thus required 38.11 min to recognize testing set, which is less than the recognition time of original KNN. The recognition rate of MKNN is 89.51, which is better than original KNN. Thus, it is proven that MKNN is faster in recognition than original KNN. Again, the performance of MKNN is better than original KNN when compared with percentage recognition rate.

Table 17.2, shows the prototype creation time for various values of parameter  $d$ , where  $0 < d < 1$ , which a distance that defined group size. This prototype creation time is more but it should not be treated as severe handicap of MKNN, since prototype calculation needs to be done only once. But the recall is very frequent. MKNN has less recall time per pattern as compared to original KNN. From Tables 17.1 and 17.3 it is clear that MKNN performs better than original KNN in terms of percentage classification, percentage recognition, classification time, and recall time.

From Fig. 17.1, we can see that as we increase the value of  $K$  (plotted on horizontal axis) the percentage recognition rate increases almost remain the same up to a certain  $K$  value, in our case that value is  $K = 8$ . After that if we increase

**Table 17.1** Performance of original KNN

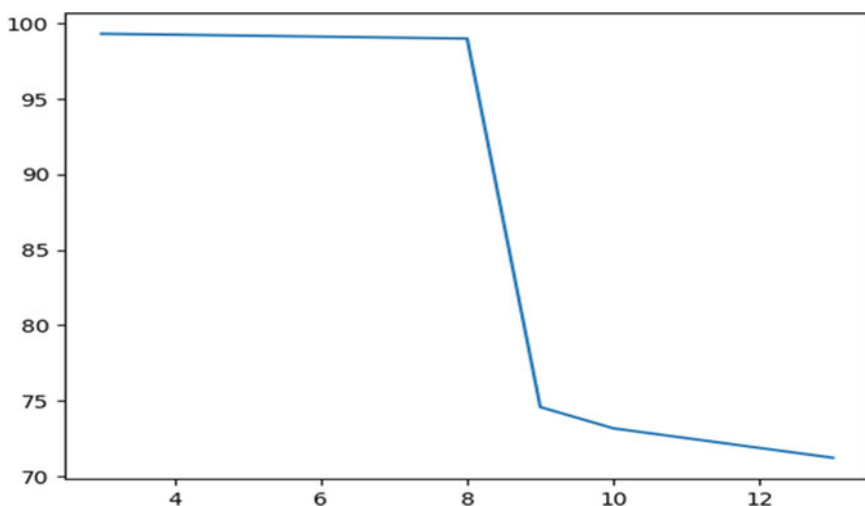
Value of $K$	Percentage classification	Percentage recognition
3	99.71	99.32
8	99.42	99.00
9	99.70	75.23
10	99.58	74.31
13	99.75	72.29

**Table 17.2** Prototype creation in MFKNN for  $K = 3$

Distance $d$	No. of prototypes created	Prototype creation time (min)
0.005	36,295	53.00
0.01	12,069	23.16
0.03	8256	20.02
0.05	1264	19.00

**Table 17.3** Performance of MFKNN for  $k = 3$

Distance	% Classification	Classification time (min)	Percentage recognition	Recognition time (min)
0.005	99.91	53.00	99.94	184.0
0.01	99.83	23.16	99.75	18.0
0.03	99.74	20.02	99.66	17.0
0.05	99.62	19.00	99.32	12.0



**Fig. 17.1** Plot of  $K$  vs. percentage recognition rate

the value we can see a sudden drop in the recognition rate and as we increase the  $K$  value it decreases further.

From Fig. 17.2 it is clearly visible that the number of prototypes created decreases when we increase the value of “ $d$ ” and vice versa. If we put value of “ $d$ ” as 0, then the number of group prototypes created are same as total no of patterns, which is referred as over fitting in pattern recognition.

In Fig. 17.3, skin samples are denoted by green stars and non-skin samples are denoted by blue triangles. The mis-prediction by model are shown in red circles. We can clearly observe that the area where model mis-predicted is overlapping area between skin and non-skin samples.

### 17.5 Conclusions

It is concluded that MFKNN performs better than original KNN in terms of classification time, percentage recognition, and recognition time. Improvement in the recognition performance and recognition time proves that MFKNN has eliminated all the drawbacks of original KNN. Again it has increased scope of MFKNN for the large data sets; previously KNN cannot be applied in such situations.

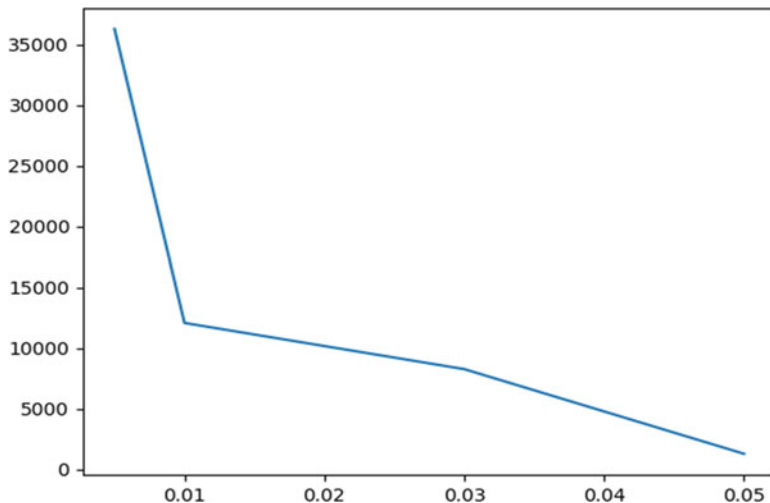


Fig. 17.2 Plot of distance  $d$  versus number of prototypes created

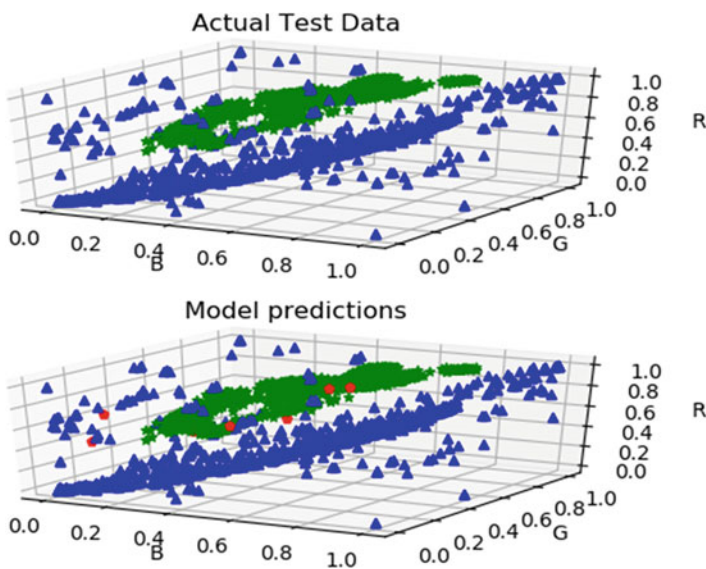


Fig. 17.3 Actual test data verses model prediction

## References

1. E. Rich, K. Knight, *Artificial Intelligence*, 2nd edn. (Tata McGraw-Hill, New Delhi, 1991)
2. R. Isermann, *Fault Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance* (Springer, Berlin, 2006)
3. E. Fix, J.L. Hodges, Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report 4 (USAF School of Aviation Medicine, Randolph Field, San Antonio, TX, 1951)
4. T.M. Cover, P.E. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967)
5. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification* (Wiley, New York, 2001)
6. T.M. Cover, P.E. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **IT-13**(1), 21–27 (1967)
7. T. Mitchell, *Machine Learning* (McGraw-Hill, New York, 1997)
8. M.E. Hellman, The nearest neighbor classification rule with a reject option. *IEEE Trans. Syst. Man Cybern.* **3**, 179–185 (1970)
9. K. Fukunaga, L. Hostetler, k-Nearest-neighbor Bayes-risk estimation. *IEEE Trans. Inform. Theory* **21**(3), 285–293 (1975)
10. S.A. Dudani, The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 325–327 (1976)
11. T. Bailey, A. Jain, A note on distance-weighted k-nearest neighbor rules. *IEEE Trans. Syst. Man Cybern.* **8**, 311–313 (1978)
12. S. Bermejo, J. Cabestany, Adaptive soft k-nearest-neighbour classifiers. *Pattern Recogn.* **33**, 1999–2005 (2000)
13. A. Jozwik, A learning scheme for a fuzzy K-NN rule. *Pattern Recogn. Lett.* **1**, 287–289 (1983)
14. J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy K-NN neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* **SMC-15**(4), 580–585 (1985)
15. P.S. Dhabe, S.G. Lade, S. Pingale, R. Prakash, M.L. Dhore, Modified K- nearest neighbor classifier using group prototypes and its application to fault diagnosis. *CIIT Int. J Data Min. Knowl. Eng. (Impact Factor 0.621)* **2**(5) (2010), May, available online <http://ciitresearch.org/dmkmmay2010.html>
16. ITQON, K. Shunichi, I. Satoru, Improving performance of k-nearest neighbor classifier by test features. Springer Transactions of the Institute of Electronics, Information and Communication Engineers, 2001

# Chapter 18

## Enhancing Cooperative Spectrum Sensing in Flying Cell Towers for Disaster Management Using Convolutional Neural Networks



M. Suriya and M. G. Sumithra

### 18.1 Introduction

Disasters are impromptu events that not only cause significant damage or loss of life but may also hammer out the existing communication networks. The damage caused to the networks along with increased demand in traffic hampers towards the recovery effort. Studies show that spectral utilization is relatively low when examined not just by frequency domain, but also across the spatial and temporal domains. Thus, an intelligent device that is aware of its surroundings and able to dynamically adapt to the existing radio frequency (RF) environment by considering is capable of utilizing spectrum more efficiently and dynamically by sharing spectral resources. Post disaster requires restoration of telecommunications in order to enable first responders to coordinate their responses, and make all affected public to access information and contact friends and relatives [1].

The unmanned aerial vehicles (UAVs), also known as drones, are adapted widely for wireless networking applications. UAVs can be used as flying cell towers, i.e. they are mounted with base stations to enabling communication to wireless networks by providing coverage, reliability and energy efficiency. UAVs can be deployed in order to complement existing cellular systems during post disaster scenario by providing hotspot and network coverage. These devices must be capable enough to operate by adapting to the environment and provide communication. A cognitive radio (CR) is an intelligent device capable of observing the environment and reconfigure based on the surrounding by learning from its experience. This work utilizes the CR technology for disaster responsive and relief networks and makes them as intelligent responsive networks. This work primarily proposes a

---

M. Suriya (✉) · M. G. Sumithra

KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India  
e-mail: [suriya13.ms@gmail.com](mailto:suriya13.ms@gmail.com); [sumithrapalanisamy74@gmail.com](mailto:sumithrapalanisamy74@gmail.com)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_18](https://doi.org/10.1007/978-3-030-19562-5_18)

181

deep learning based technique called SpecCNN to improve spectrum sensing by calculating the signal-to-noise ratio in the CR adapted UAVs for any disastrous scenario [2].

The remaining part of the chapter is organized as follows: Feasibility study based on the proposed works is done and is presented in Sect. 18.2. Section 18.3 presents the system model for intelligent disaster response networks. The deep learning based technique called convolution neural network for cooperative spectrum sensing (CSS) is presented in Sect. 18.4. Section 18.5 covers the proposed SpecCNN model for emergency cognitive radio. The results of the work are provided in Sects. 18.6, and 18.7 concludes the chapter.

## 18.2 Literature Survey

Namuduri in [3] has devised and deployed an aerial communication system that uses AR200 drone mounted with a base station that is weighing about 2 kg. This cell tower works in Band 14 (public safety band) and helps to restore cellular service in the disastrous affected location and also allows share photos and video among communicating users. This innovative public safety responder system provides reliable communication when the coverage is increased as the drone flew higher.

Islam and Shaikh [4] have proposed a disaster management system based on dynamic cognitive radio network technology. In this Artificial Neural Network (ANN) technique was devised for disaster detection based on backward propagation algorithm. He also created a service discovery scheme along with ANN-based spectrum sensing for performing coordination during the time of disaster. The switching time of spectrum sensing scheme was also analysed by calculating the latency of proposed service discovery scheme.

In [5] Grodi et al. work the importance of communication links and its role during any emergency scenario was highlighted. When public telephone networks damaged during disaster, Unmanned Aerial vehicles (UAVs) were used to establish communication between those persons met with an emergency and the rescue team. It explains the use of drones to act as mobile base stations and route wireless communication to the nearest working public telephone network access point.

Mozaffari et al. in [6] explore features of unmanned aerial vehicles such as mobility, flexibility, and its adaptive altitude and suggest the utilization of drones for various wireless systems applications. The work explains how UAVs are deployed as flying mobile terminals by providing additional capacity to hotspot areas and also to enhance network coverage in emergent public safety situations. Also introduces various tools such as optimization theory, machine learning, stochastic geometry, transport theory, and game theory for addressing UAV problems.

Lee et al. in [7] investigates the cooperative spectrum sensing (CSS) in a cognitive radio network (CRN) which consists of multiple secondary users (SUs) at a point where they cooperate to detect when a primary user (PU) arrives. The concept of deep sensing is proposed by utilizing convolutional neural network (CNN). The

sensing samples are trained using CNN instead of traditional mathematical model. The results of the work show a significant increase in places where there is low signal-to-noise ratio (SNR) when the number of training samples were sensible and thus proposed an environment aware CSS.

### 18.3 Intelligent Disaster Response Networks

During natural disaster lot of damage is caused to the telecommunication networks leading to loss of communication among people who are in affected zone and require emergency communication. This rises to build a temporary relief and response system to support communication in places where available telephone network is destroyed. The unmanned aerial vehicles also called drones can help this kind of disastrous scenario by making it mount with a base station weighing about 2 kg and route to nearest public telephone network access point. Figure 18.1 illustrates the deployment of drones as aerial base stations (BSs) to deliver a reliable, cost-effective, and on-demand wireless communications over disaster affected areas. These devices are called as flying cell towers as they can coexist and connect with terrestrial cell towers as long they are deployed in the air. The advantage of these flying cell towers is their ability to establish better line-of-sight (LoS) communication links to the ground users compared to conventional system [8].

The use of dynamic learning algorithms enable drones to effectively adjust their movement, flight path, and motion control to service their ground users. UAVs adapt to any environment dynamically in a self-organizing way and autonomously optimize their trajectory. The training of these dynamic devices using advanced neural networks and performing data analytics will make predict ground user behaviour to track user mobility and thereby effectively operate drones. This feature

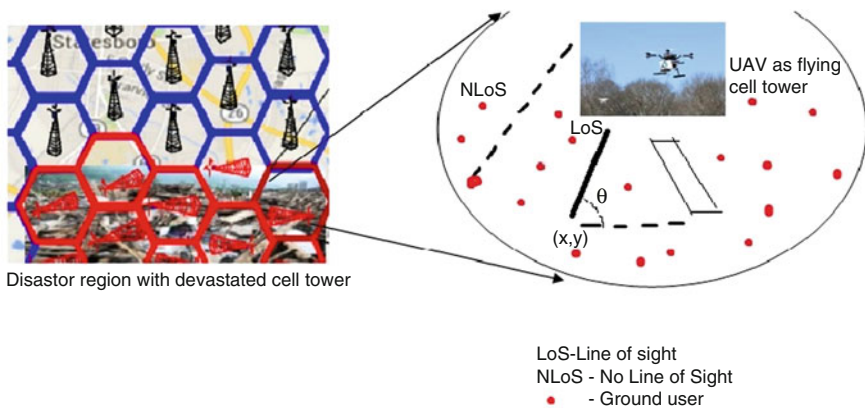


Fig. 18.1 Depiction of intelligent flying cell towers at disaster prone location

helps in designing a special cache enabled UAV system to store users' mobility pattern and dynamically optimize the trajectory to enhance communication [13].

## 18.4 Deep Learning for CSS Using CNN

In [9] a cognitive radio network (CRN), the process of spectrum sensing, helps to identify the presence of primary user when there are multiple secondary users allocated to the available frequency spectrum. The most challenging task in CRN is identifying the tolerable level of all PUs in the network which requires an efficient utilization of spectrum sensing mechanism. Compared to conventional learning techniques deep learning is gaining attention and is being used for many applications such as natural language processing, image processing and various analytical applications. Due to efficient learning process and data support it is widely used for a number of big data applications. A deep neural network (DNN), in Fig. 18.2, consists of more than two hidden layers, hence called multi-layer network that emulate the working of human brain neuron system. Convolutional neural network (CNN) and recurrent neural network (RNN) [10] are used to train huge amount of sample data that requires complex mathematical modelling.

Cooperative communication primarily increases the coverage and minimizes the outage of wireless links for certain channel conditions [11, 12]. DNN has been proposed to wireless communication systems to classify signals where multiple SUs are present in a cooperative spectrum sensing scheme. The CNN technique is applied in spectrum sensing by considering each SUs sensing results as training sample and combining the results to predict PUs efficiently. When CNN is applied to CSS, the sensing results are optimized by combining the movement of PU and location of SU to enhance spectrum sensing.

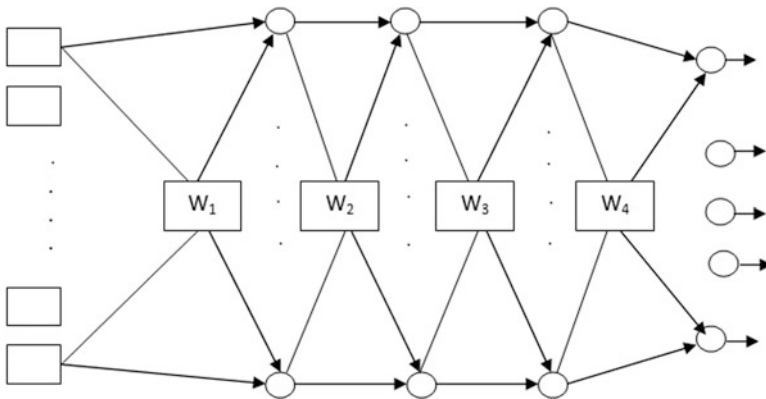


Fig. 18.2 Deep neural network model with three hidden layers



## 18.5 SpecCNN Model for Emergency Cognitive Radio

### 18.5.1 Cyclostationary Signal Feature Extraction

Figure 18.3 depicts how the feature extraction process is performed over the cyclostationary signal (PU) to separate the noise signal and extract different features of the signal to perform spectrum sensing.

A scenario of CRN composed of one PU and multiple secondary users (SUs) can be represented as a binary hypothesis-testing. The following basic hypothesis  $H_0$  and  $H_1$  are considered as in Eqs. (18.1) and (18.2).

$H_0$ : Power of primary user absent at time ‘ $t$ ’

$H_1$ : Power of primary user present at time ‘ $t$ ’

$$H_0 : x(t) = n(t) \tag{18.1}$$

$$H_1 : x(t) = h(t) + n(t), \quad t = 0, 1, \dots, N - 1 \tag{18.2}$$

where

$N$  = number of samples over a period of received signal

$x(t)$  = secondary users signal

$h(t)$  = primary users signal

$n(t)$  = amount of AWGN noise (Additive White Gaussian Noise) with variance  $\sigma_n^2$ .

The PU signal and noise can be distinguished by extracting the different feature of the cyclostationary signal (PU). Suppose that the SU receives the signal is  $h(t)$ , and its cyclic autocorrelation function can be represented in Eq. (18.3).

$$R_x^\alpha = \frac{1}{T_0} \int_0^{T_0} R_x(t, \tau) e^{-j2\pi\alpha t} dt \tag{18.3}$$

where  $\alpha$  = cyclic frequency and  $T_0$  = cycle period.

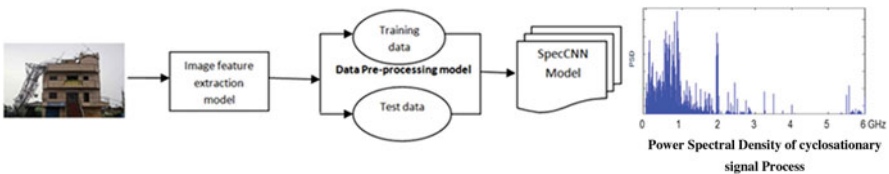


Fig. 18.3 Feature extraction process

The cyclic autocorrelation function is obtained by Fourier transform, and the spectral correlation function (SCF) is obtained as a Fourier transform relation between the conventional power spectral density and the autocorrelation given in Eq. (18.4).

$$S_x^\alpha(f) = \int_{-\infty}^{\infty} R_x^\alpha(\tau) e^{-j2\pi f\tau} d\tau \quad (18.4)$$

where

$S_x^\alpha(f)$  = power spectral density  
 $R_x^\alpha(\tau)$  = autocorrelation function

Based on the binary hypothesis-testing function the (1) autocorrelation function, (2) SCF function and (3) energy function of the extracted signals are obtained as in Eqs. (18.5) and (18.6), respectively.

$$R_x^\alpha = \begin{cases} R_{x,0}^\alpha, & H_0 \\ R_{x,1}^\alpha, & H_1 \end{cases} \quad (18.5)$$

where

$$H_0 = R_{x,0}(t, \tau) e^{-j2\pi\alpha t}$$

$$H_1 = R_{x,1}(t, \tau) e^{-j2\pi\alpha t}$$

$$S_x^\alpha(f) = \begin{cases} S_{x,0}^\alpha(f), & H_0 \\ S_{x,1}^\alpha(f), & H_1 \end{cases} \quad (18.6)$$

where  $H_0, H_1 = \rho_n^2 \sigma(\alpha)$ , since white Gaussian noise.

The energy feature of the extracted signals is obtained as in Eq. (18.7).

$$E_{e,x} = \begin{cases} E_{e,0}, & H_0 \\ E_{e,1}, & H_1 \end{cases} \quad (18.7)$$

where

$$H_0 = \sum_{t=1}^N (n(t))^2$$

$$H_1 = \sum_{t=1}^N (h(t) + h(t))^2$$

Once the features are extracted, the data is pre-processed to make the training data set and the test data set standard.

### 18.5.2 SpecCNN Algorithm

Deep sensing CNN is utilized to analyse the presence of PU signal by combining all the individual sensing results by exploring the signal spectral features based on the amount of noise signal. A spectral image is fed as input to neuron for each sensing result and the CNN performs identification of spectral correlation using adjacent pixels since each image has interrelated spectrum spatial relationship. In Fig. 18.4, a CNN model consists of convolution part (Conv) at the front and fully connected (FC) part at the back. The convolution part extracts features and passes data to the rectified linear unit (ReLu) layer which helps in making system non-linear, and the output of ReLu is fed to the max-pooling layer that generates  $\max(x, 0)$  when input is  $x$ . This layer helps in reducing the size of the data without any loss in the actual input. The FC layer makes final decision with input from the Conv layer and its output fed to softmax function to detect the presence of PU signal or not.

#### 18.5.2.1 SpecCNN Training Algorithm

1. Calculate the output  $\alpha^i$  at each neuron during forward propagation.
2. Calculate error  $\delta^i$  at each neuron during back propagation.
3. Compute weight from neuron  $i$  to  $j$  as gradient weighted value  $\omega^{ij}$ .
4. Apply gradient descent rule to update weight:

$$\delta^i = \phi \left( \omega^{ij} \text{Conv} \left( x^{ij} \right) + b^i \right)$$

where

- $\Phi$  = activation function of output  $i$
- $\omega^{ij}$  = weighted sum at output  $i$
- $b^i$  = bias/error at  $i$

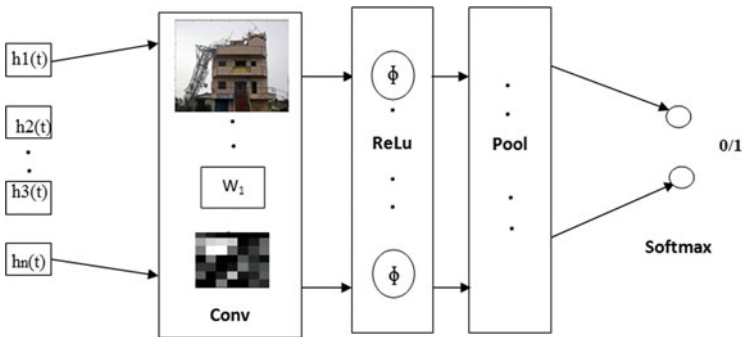
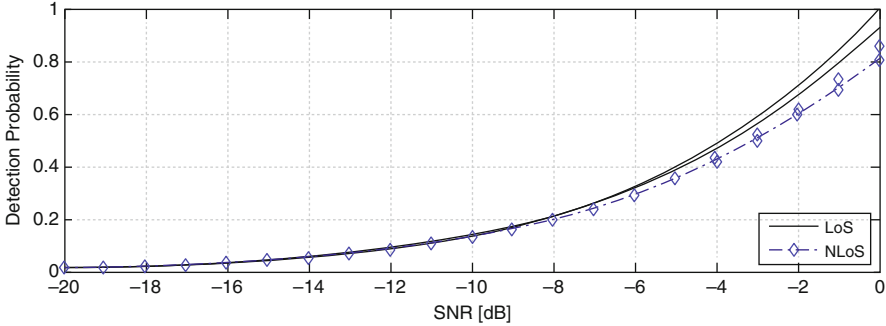


Fig. 18.4 SpecCNN model



**Fig. 18.5** Signal-to-noise ratio vs. detection probability for LoS and NLoS

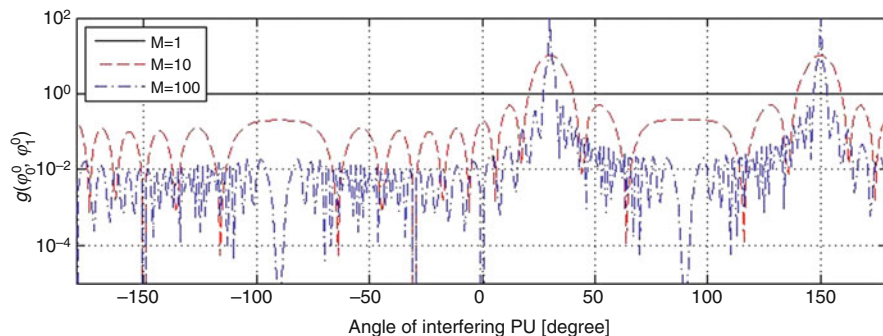
The above proposed SpecCNN model performs computation to verify the presence or absence of PU signal which is considered as index 1 and 0, respectively. Input signal of two-dimensional matrix ( $2 \times 2$ ) is considered that includes cyclostationary and energy signal. The convolution layers considered for the network includes a filter size ( $3 \times 3$ ), stride 1 and padding = 0. The gradient descent stochastic technique is used to train the neurons in each layer.

## 18.6 Evaluation and Discussion

The proposed SpecCNN model for efficient spectrum sensing based on CNN is simulated in MATLAB environment using AWGN channel for 100 samples at a carrier frequency of 10 Hz. The cyclostationary signals are calculated at an SNR value of  $-15$  dB and SCF is extracted using Eq. (18.6) and the energy features are extracted using Eq. (18.7). The resulting data are used as training set to perform CNN process as discussed in the proposed SpecCNN algorithm for training the network and updating weights. The probability of signal detection with a false alarm  $P_f = 0.05$  (see Fig. 18.5) that shows the rate of PU and SU increases with the decrease in SNR value. The amount of interference of cyclostationary signal for the number of users  $m = 1$ ,  $m = 10$  and  $m = 100$  for evaluating primary signal spectrum is analysed (see Fig. 18.6).

## 18.7 Conclusion

The emerging problem of spectrum sensing was considered for disastrous affected areas when a drone deployed the cell tower and acted as flying cell tower. The cooperative spectrum sensing using cyclostationary signal based image feature extraction was implemented using the proposed SpecCNN deep learning algorithm



**Fig. 18.6** Rate of interference of PU signal

model and the test accuracy of 0.9068 (training rate 300) was obtained, which indicated the detection probability increased in the minimal SNR region with the deep learning model of spectrum sensing.

## References

1. A. Trotta, Re-establishing network connectivity in post-disaster scenarios through mobile cognitive radio networks, in *12th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET)*, 2013. ISBN: 978-1-4799-1004-5
2. M.H. Rehmani, A.C. Viana, H. Khalife, S. Fdida, A Cognitive Radio Based Internet Access Framework for Disaster Response Network Deployment. [Research Report] RR-7285, INRIA, 2010
3. K. Namuduri, S. Chaumette, J. Kim, J. Sterbenz (eds.), *UAV Networks and Communications* (Cambridge University Press, Cambridge, 2017). <https://doi.org/10.1017/9781316335765>
4. N. Islam, G.S. Shaikh, Towards a Disaster Response System Based on Cognitive Radio Ad Hoc Networks, 2017, arXiv:1710.02404 [cs.NI]
5. R.D. Grodi, Design, Analysis and Evaluation of Unmanned Aerial Vehicle Ad hoc Network for Emergency Response Communications, Electronic Theses & Dissertations, 2016
6. M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, M. Debbah, A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems, 2018, arXiv:1803.00680
7. W. Lee, M. Kim, D.-H. Cho, R. Schober, D. Sensing, Cooperative Spectrum Sensing Based on Convolutional Neural Networks, 2017, arXiv:1705.08164v1
8. K. Namuduri, Flying cell towers to the rescue. *IEEE Spectr.* **54**(9), 38–43 (2017). <https://doi.org/10.1109/mspec.2017.8012238>
9. V.Q. Do, I. Koo, Learning frameworks for cooperative spectrum sensing and energy-efficient data protection in cognitive radio networks. *Appl. Sci.* **8**, 722 (2018). <https://doi.org/10.3390/app8050722>
10. A. Fotouhi, M. Ding, M. Hassan, Dynamic Base Station Repositioning to Improve Spectral Efficiency of Drone Small Cells, 2017, arXiv:1704.01244v1 [cs.IT]
11. F. Paisana, A. Selim, M. Kist, P. Alvarez, J. Tallon, C. Bluemm, A. Puschmann, L. DaSilva, Context-aware cognitive radio using deep learning, in *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2017, 978-1-5090-2830-6/17

12. P. Rungsawang, A. Khawne, The implementation of spectrum sensing and spectrum allocation on cognitive radio, in *19th International Conference on Advanced Communication Technology (ICACT)*, 2017, <https://doi.org/10.23919/ICACT.2017.7890206>
13. M. Mozaffar, W. Saad, M. Bennis, Y.-H. Nam, M. Debbah, A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems, 2018, arXiv:1803.00680v1 [cs.IT]

# Chapter 19

## Emoticons and Their Effects on Sentiment Analysis of Twitter Data



P. S. Dandannavar, S. R. Mangalwede, and S. B. Deshpande

### 19.1 Introduction

Recent years have seen an uncontrollable increased use of social media such as weblogs, micro-blogs, discussion forums, message boards, and social network services (SNSs). People have taken to social media like never before, to express their opinion on just about anything—person, brand, product, movie—through micro-blog posts, short messages, reviews, etc. Collectively these can be referred to as “consumer-generated content.” The number of such media users is increasing by the day and has resulted in individuals generating overwhelming amount of data with the click of a mouse button. These messages contain rich “sentimental information” with traces of valuable information, i.e., people’s sentiment towards a brand, a product, a person, etc.

It has become crucial for today’s businesses to use tools to analyze such sentiment rich content to understand how consumers perceive their product/brand, etc. At around the same time that SNSs gained unprecedented growth, a need was felt to analyze user-generated content. To this end, efforts were made to develop technology for useful applications. One such tool is SA, which relates to polarity classification of a text. Thus, SA of short informal texts has garnered tremendous interest across different domains (e.g., trend recognition, market prediction, spam detection, decision-making, popularity analysis, and health).

SA is the task of labelling data with a polarity—positive, negative, or neutral—through analysis of the properties contained within the data. Classification can be binary—meaning either positive or negative—or describe a detailed range of

---

P. S. Dandannavar (✉) · S. R. Mangalwede · S. B. Deshpande  
Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi,  
Karnataka, India  
e-mail: [padmad@git.edu](mailto:padmad@git.edu); [mangalwede@git.edu](mailto:mangalwede@git.edu); [sbdeshpande@git.edu](mailto:sbdeshpande@git.edu)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_19](https://doi.org/10.1007/978-3-030-19562-5_19)

191

polarity. To accomplish this, SA uses the emotional words appearing in a text segment to analyze the text, using natural language processing.

There are three standard methodologies that can be used for SA:

1. Lexicon-based methodology—deciding the polarity includes utilizing an arrangement of words (positive and negative), which are contained in a word reference (dictionary).
2. Machine learning-based methodology—utilize machine learning calculations (supervised or unsupervised) to characterize information.
3. Hybrid methodology—utilizes a mix of the over two methodologies.

### ***19.1.1 Text v/s Emoticons***

In regular face-to-face (F2F) conversations, sentiment of the speaker can be gathered from visual cues—smiling, laughing, crying, angry, etc. Such visual cues are however lost in plain text communication. Twitter, for example, has emerged as a very popular micro-blogging service, which permits users to post short textual sentences called “tweets.” In addition to conveying factual information, these tweets also reflect the emotion(s) of the author. Tweets when analyzed can help understand user behavior. Twitter limits every tweet up to a maximum of 140 characters which makes them extremely short. Because of this limitation, people use emoticons, as they help them expressing their emotions better in a short message. A variety of emoticons are used by social media users to express emotions. Over the years, social media users have happily embraced the use of emoticons as an alternate to F2F visual cues. Emoticons act as cues, signalling user’s sentiment in social media content.

An emoticon is typically made up of typographical symbols (such as :, -, ), (, =) and serves as a shorthand for facial expressions. It conveys its meaning through its graphic resemblance to a physical object—for example, a smiling face. Emoticons complement traditional text-based computer-mediated communication.

Since its first use, by Prof. Scott Fahlam of the Carnegie Mellon University in 1982, the use of emoticons has caught on and spread to a much larger community in a very short time. Decades later, emoticons have found their way into everyday social media content.

### ***19.1.2 Emoticons and Sentiment***

Emoticons are strong indicators of sentiment and enable people to express their mood/sentiment/emotion better. Though they can be used (i.e., sent and received) independently, they often accompany a text segment. In such cases, the sentiment of the overall text is dependent on the sentiment carried by the emoticon. However, the



sentiment of an emoticon is independent of its embedding text. When accompanying a text segment, emoticons can be used to:

1. *Intensify the sentiment of the text.* When a “happy” smiley is added to a positive statement, the orientation becomes more positive. For example, “Loved the weather” and “Loved the weather :)”.
2. *Change/negate the sentiment of the text.* For example, “Awesome weather” and “Awesome weather :(”.
3. *Convey sentiment of the accompanying text segment,* especially when the text is ambiguous and does not carry any sentiment. Emoticons carry the only sentiment in such cases. For example, “It rained today :)” or “It rained today :(”.

## 19.2 Previous Work

Several studies have been carried out on SA, in recent years. Most of the researchers have however dismissed emoticons as noisy information and deleted them in the preprocessing stage. By doing so, most SA approaches maybe failing to consider information that is important. For example, in a statement that is objective otherwise, an emoticon present may signal the intended sentiment—“This phone does not work :(”. Work of very few researchers [1–20] have taken into account emoticons in the overall SA process.

Various methods have been used by researchers to handle emoticons in the SA process. Authors in [1] used the technique of clustering emoticons and words (using word2vec and K-means algorithm), to prove that sentiment classification was affected by removing emoticons. The work of [2] highlighted the importance of emoticons in SA by using different examples. Hogenboon et al. proposed [3] a framework for automated SA, which used a manually created emoticon sentiment lexicon to account for the information conveyed by emoticons. The concept of clustering words and emoticons (using k-means algorithm) was again used by Joylin et al. [4] to determine the meaning conveyed by the emoticons. Their work focused on comparing results of SA of text before and after the emoticons are removed. Authors in [5] explored the effects of emoticons in Twitter SA by using and comparing three emoticon preprocessing methods, viz., emoDel, emo2label, and emo2explanation. While work in [6] focused on the study of SA on tweets with emoticons, authors in [7] investigated how social power is related to language use and communication behavior on Twitter. The results in [8] showed that considering the sentiment conveyed by emoticons improved polarity classification. The work of Miller et al. [9] dealt with varying interpretations of emojis. The first emoji sentiment lexicon was provided by Novak [10] using which the authors drew one interesting conclusion—that there was a significant difference in the sentiment reflected by the tweets with and without emojis. MoodLens, the first system for SA of Chinese tweets based on emoticons, was proposed in [11]. References [12, 13] focused on the analysis of methods and the impact of emoticon removal and

SA of short informal text, respectively. The authors in [14] suggested a method to create vectors of emoticons. This vector was created automatically by using the relationship between emotional words and emoticons, while authors in [15] proposed a supervised sentiment classification framework using twitter tags and smiley's as sentiment labels. Seven different emoticon categories and a set of +ve/-ve words were used in [16] to analyze the behavior of people using social media. A detailed examination of the analysis methods and tools used in the process of SA has been presented by Tyagi [17].

The question that arises is—"Should algorithms dealing with sentiment classification account for emoticons?" In this work, we explore the influence of emoticons on SA by testing the hypothesis that "Removing emoticons from text affects Sentiment Classification."

### 19.3 System Design

The overall system can be visualized as consisting of three major components as depicted in Fig. 19.1.

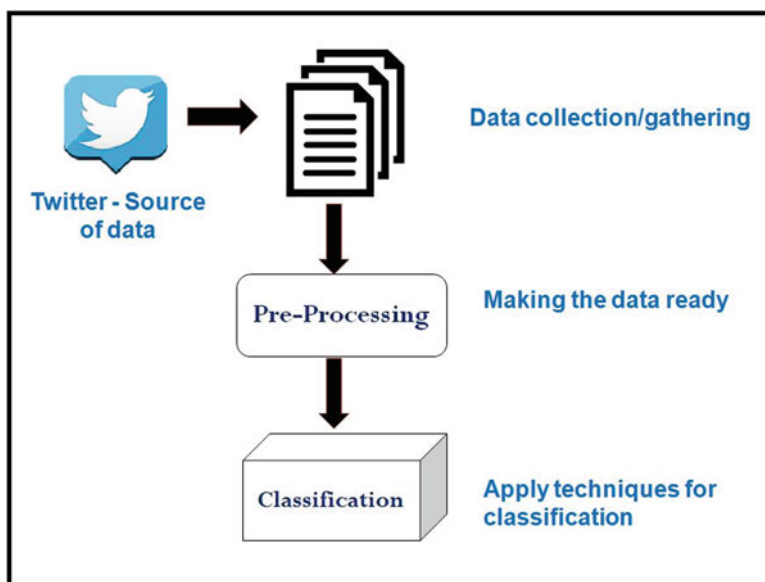


Fig. 19.1 System block diagram

### ***19.3.1 Twitter—Source of Data for Analysis***

Twitter micro-blogging site. Sending and receiving text messages, i.e., “tweets” is restricted to only registered users. Originally restricted to just 140 characters, the length of tweets was doubled for all languages in November 2017, except for Japanese, Korean, and Chinese [21]. It is reported that there are a total of 330 million active twitter users monthly and the total number of tweets tweeted everyday is massive—500 million tweets, as measured in January 2018. This large volume of data (tweets) is openly accessible. Tweets can be gathered using the twitter API which is relatively simple compared to web scraping.

For the current work, data sets for training and testing were collected from Twitter using “Tweepy”—a python library.

### ***19.3.2 Data Preparation***

The tweets that are fetched are raw in nature and consists of many attributes that are redundant for the analysis under consideration. The only relevant and required attribute from the tweet is the “Text.” Rest are eliminated. Data preprocessing is a very important step as it decides the efficiency of the other steps down in line. Preprocessing involves the standard steps of Case Conversion, Stop-words Removal, Punctuation Removal, Stemming, Lemmatization, and POS Tagging.

### ***19.3.3 Machine Learning Tool(s)***

After the tweets are fetched, preprocessed, and the features relevant for sentiment analysis are extracted, SA is performed using a Machine learning classifier. We use the Naïve Baye’s algorithm.

## **19.4 Implementation Details**

As a part of the current study, two experiments were carried out. The details of the two experimental setups are as indicated below.

Two variations of each of the data sets were used: a preprocessed set with emoticons excluded, and a preprocessed set where emoticons have been replaced with sentiment-laden words corresponding to the sentiment value of the emoticon.

The implementation involves the following steps as shown in Fig. 19.2.

The process starts with fetching the tweets using the twitter API. The raw tweets fetched consist of several attributes which are not required for analysis and hence

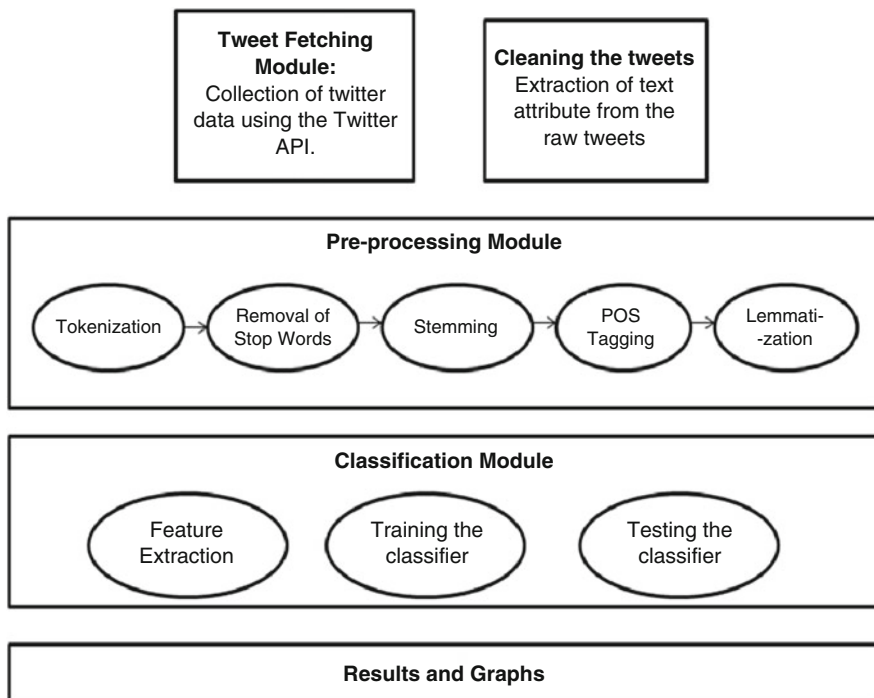


Fig. 19.2 System architecture

must be removed. The only attribute required for analysis is the “Text” attribute which is extracted during the preprocessing phase. For analysis without emoticons, the emoticons appearing in the tweets are removed from consideration and for the analysis with emoticons, the emoticon’s are handled as indicated in Tables 19.1 and 19.2, respectively. After the tweets are preprocessed, the features relevant for SA are selected. Once the desired feature set is available, a machine learning classifier (Naïve Baye’s classifier, in our case) is used for SA, prior to which the classifier must be trained and tested. The classifier is trained and its performance measured training and test data, respectively. Once the model has been “trained” on the dataset, it can be evaluated on new unseen data. New tweets can be fetched on any topic and given to the classifier. The classifier predicts the tweet as carrying a positive/negative/neutral sentiment. The sentiments predicted without emoticons and with emoticons are compared by plotting a graph.

**Table 19.1** Experimental setups

	Setup for experiment 1	Setup for experiment 2
Data source	Twitter	Twitter
Size of dataset	10,000+ tweets	Varying sizes (50 kB, 200 kB, 3 MB, 4 MB)
Technique for handling emoticon's	Converted into actual meaning in the form of text	Assigning a E-Type value (refer Table 19.2) to a group of emoticons
Machine learning algorithm used	Naïve Baye's algorithm	Naïve Baye's algorithm

**Table 19.2** Emoticons and E-type values

Emoticons	Example						E-type
EMOT_SMILEY	:-) :	:	(:	(-:			1
EMOT_LAUGH	:-D :D	:D	X-D XD	xD			2
EMOT_NEUTRAL_NOT_SURE	== -_-	-_-	:-  :-\	:O	:-!		0, 3
EMOT_FROWN	:-( :(	:(	(:	(-:			-1
EMOT_CRY	:'( '(	:(	:'(	:((			-2

## 19.5 Result Analysis

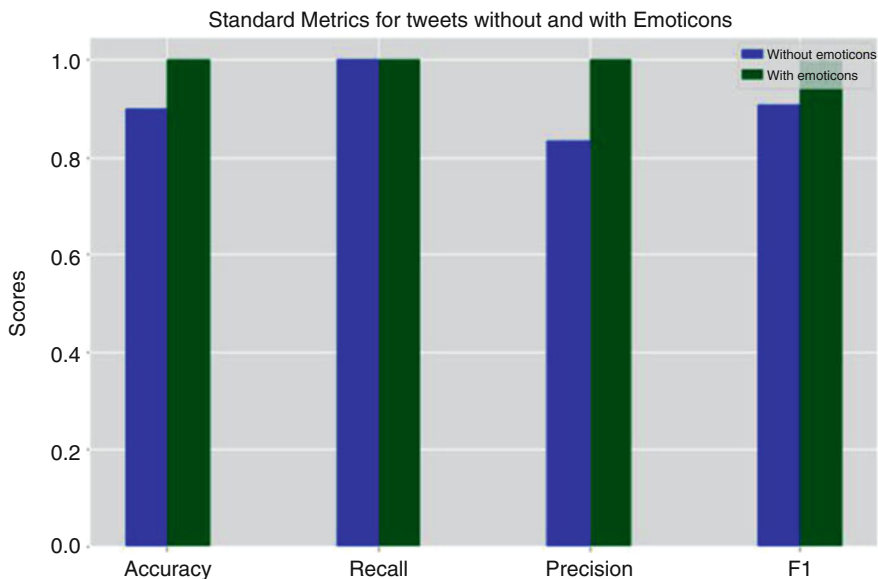
The accuracy of classifiers can be compared by plotting a confusion matrix, which allows us to visualize the performance of a classifier. The confusion matrix [22] makes it easy to identify if the system under consideration is confusing two classes (i.e., mislabelling one as another). It is a  $2 \times 2$  matrix—that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). The following measures can be computed from the confusion matrix and used to determine the performance of a classifier:

1. Accuracy—overall, how often is the classifier correct? [23]
2. Precision—when a yes is predicted, how often is it correct? [23]
3. Recall— $TP/(TP + FN)$ —measure the completeness of the model w.r.t. each class
4. F-score—is estimated as:

$$F\text{-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

The result of applying the Naïve Bayes classifier for the two experimental setups is depicted in the form of various graphs shown below.

As is evident in the graph of Fig. 19.3, the Naive Bayes classifier trained with the tweets that included emoticons has improved accuracy, precision, and F-scores compared to results of classifiers which were trained on tweets without emoticons, which displayed lower accuracy, precision, and recall for the same set of tweets. The percentage of tweets classified as negative was higher when emoticons were considered (60.3%) compared to when emoticons were dismissed (57%), indicating

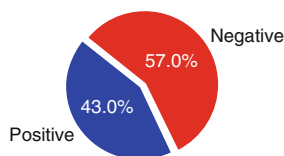


**Fig. 19.3** Bar graph (experimental setup 1)

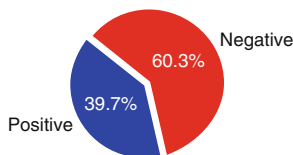
**Fig. 19.4** Pie chart representing the ratio of positive and negative tweets (experimental setup 1)

Percentage of positive and negative tweets

Pie chart without emoticons



Pie Chart with emoticons



that negative tweets were wrongly classified as being positive and removal of emoticons affected this classification. This is evident in the pie chart of Fig. 19.4.

The results obtained for the experimental setup 2 are in sync with the results for setup 1. The results obtained in analysis “with emoticons” have more accuracy (i.e., 92.5%) when compared to analysis “without emoticon” (i.e., 79.37), as depicted in graphical representation of Fig. 19.5.

As is very evident from the graphs, performance measures, viz., accuracy, precision, and F-score, were better for the classifiers that dealt with the both text

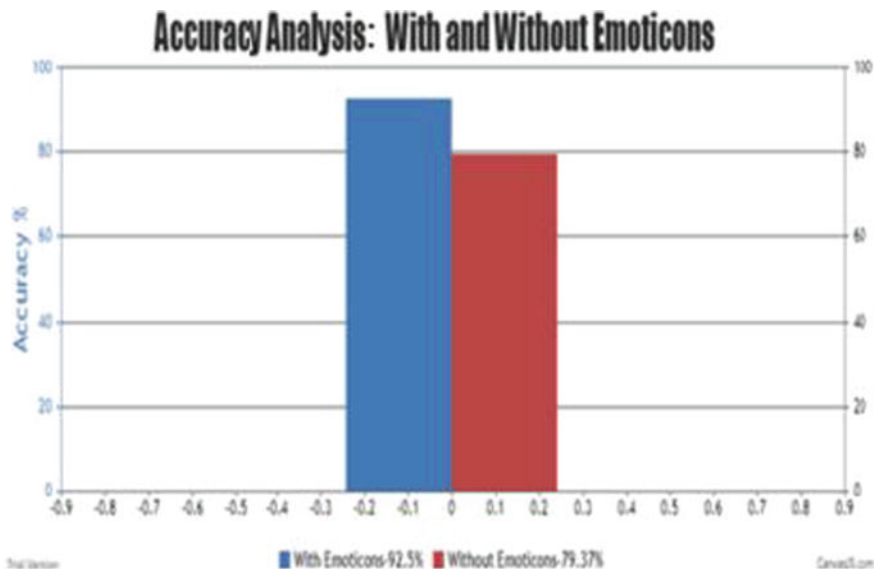


Fig. 19.5 Bar graph (experimental setup 2)

and emoticons in a tweet, compared to classifiers that dealt with only text in tweets and did not account for emoticons. The hypothesis that “Removing emoticons from text affects Sentiment Classification,” thus stands proved.

## 19.6 Conclusion

In the present Internet world, peoples sentiment plays an important role in several fields—like sentiments present in people’s reviews about a product, about a company, health care facility, stock market, about a new movie, etc. The expansive volumes of data contained in micro-blogging sites make them a preferred choice as a source of data for SA. Tweets (which basically may be opinions) can be used to understand and analyze mediocre mindset of a general mass or public.

Most present systems of SA do not take emoticons into consideration. Given the length restriction of short texts, authors can easily convey/describe their feelings by using emoticons and readers of those texts can instinctively comprehend the intention of the author. Emoticons can thus be used to describe the implicit intention that cannot be depicted by dialect. Thus, extraction and examination of emoticon is useful for SA.

This paper takes into account the sentiment expressed by the emoticon along with the text which increase the accuracy of the analysis. After considering emoticons we observe that there is an increase in standard metrics. Hence we conclude that if

emoticons are removed from text and their independent sentiment is not considered, it does affect sentiment classification and thus emoticons play a significant role in conveying the sentiment of a text in totality and hence should be considered.

## References

1. H. Wang, J.A. Castanon, *Sentiment Expression via Emoticons on Social Media* (Silicon Valley Lab IBM, San Jose, CA, 2015)
2. P. Yadav, D. Pandya, SentiReview: sentiment analysis based on text and emoticons, in *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2017
3. A. Hogenboom, D. Bal, F. Frasinca, Exploiting emoticons in sentiment analysis, in *SAC'13*, Coimbra, Portugal, March 18–22, 2013. ACM 978-1-4503-1656-9/13/03
4. B. Joylin, T. Aswathi, N. Victor, Sentiment analysis based on word-emoticon clusters. *Int. J. Pharm. Technol. (IJPTFI)* **8**(4), 25288–25296 (2016). ISSN: 0975-766X
5. K. Węgrzyn-Wolska, L. Bougueroua, H. Yu, J. Zhong, *Explore the effects of emoticons on twitter sentiment analysis*. CSEN, SIPR, NCWC—2016, CS & IT-CSCP, 2016, pp. 65–77. <https://doi.org/10.5121/csit.2016.61006>
6. B. Waghode Poonam, M. Kinikar, *Interpreting the public sentiment with emotions on twitter*. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* **2**(XII), 240–243 (2014). ISSN: 2321–9653
7. S. Tchokni, D.Ó. Séaghdha, D. Quercia, *Emoticons and Phrases: Status Symbols in Social Media* (Association for the Advancement of Artificial Intelligence, Menlo Park, CA, 2014). [www.aaai.org](http://www.aaai.org)
8. A. Hogenboom, D. Bal, F. Frasinca, M. Bal, *Exploiting emoticons in polarity classification of text*. *J. Web. Eng.* **14**(1–2), 22–40 (2015)
9. H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, B. Hecht, “Blissfully Happy” or “Ready To Fight”: *Varying Interpretations of Emoji* (Association for the Advancement of Artificial Intelligence, Menlo Park, CA, 2015). [www.aaai.org](http://www.aaai.org)
10. P.K. Novak, J. Smailović, B. Sluban, I. Mozetič, Sentiment of Emojis. *PLoS One* **10**, e0144296 (2015). <https://doi.org/10.1371/journal.pone.0144296>
11. J. Zhao, L. Dong, J. Wu, K. Xu, MoodLens: an emoticon-based sentiment analysis system for Chinese tweets, in *KDD'12*, Beijing, Aug 12–16, 2012. ACM 978-1-4503-1462-6/12/08
12. A. Pålsson, D. Szerszen, *Sentiment Classification in Social Media—An Analysis of Methods and the Impact of Emoticon Removal* (School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm, 2016)
13. S. Kiritchenko, X. Zhu, S.M. Mohammad, Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **50**, 723–762 (2014)
14. S. Aoki, O. Uchida, Method for automatically generating the emotional vectors of emoticons using weblog articles, in *Recent Researches in Applied Computer and Applied Computational Science*. ISBN: 978-960-474-281-3
15. D.D.O. Tsur, A. Rappoport, Enhanced sentiment learning using twitter hashtags and smileys, in *Coling 2010: Poster Volume*, Beijing, Aug 2010, pp. 241–249
16. C. Mahajan, P. Mulay, *E3: effective emoticon extractor for behavior analysis from social media*, *Second International Symposium on Big Data and Cloud Computing (ISBCC'15)*. *Proc. Comput. Sci.* **50**, 610–616 (2015)
17. E. Tyagi, A.K. Sharma, An intelligent framework for sentiment analysis of text and emotions— a review, in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS 2017)* (IEEE, Piscataway, NJ, 2017), 978-1-5386-1887-5/17/
18. D. Tang, B. Qin, T. Liu, Q. Shi, *Emotion Analysis Platform on Chinese Microblog* (Research Center for Social Computing and Information Retrieval Computer Science Department, Harbin Institute of Technology, Harbin)



19. F. Morstatter, K. Shu, S. Wang, H. Liu, Cross-platform Emoji interpretation: analysis, a solution, and applications, in *Proceedings of ACM Conference*, Washington, DC, Jul 2017
20. A.P. Jain, P. Dandannavar, Application of machine learning techniques to sentiment analysis, in *Second International Conference on Applied and Theoretical Computing and Communication Technology* (IEEE, Piscataway, NJ, 2016), ISSN: 978-1-5090-2399-8/16
21. <https://en.wikipedia.org/wiki/Twitter>
22. <https://www.geeksforgeeks.org/confusion-matrix-machine-learning>
23. <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

# Chapter 20

## Prediction of Customer Churn Using Machine Learning



Saifil Momin, Tanuj Bohra, and Purva Raut

### 20.1 Introduction

Due to the burgeoning consumer market, there has been a long-lasting issue of customer relations and customer retention. As the market becomes more competitive, customer retention becomes more and more expensive. Particularly in the telecommunication industry, cost of customer churn is approximately around \$10 billion per year [1]. Moreover, study suggests that companies end up spending much higher charges in a race to gain new customers than to retain the existing ones [2]. Thus, there is no doubt that an efficient technique to predict the customer churn can greatly reduce the company expenditure in customer relations.

Customer churn datasets often have a great number of variables and features; some of these datasets coming from the data warehouse have more than 100,000 variables [1]. Traditional classification algorithms like logistic regression, naive Bayes, random forest, and decision trees require careful evaluation of these features, as not all features are significant in churn prediction [3, 4]. Thus, feature engineering requires a great amount of effort in terms of computation as well as time.

Deep Learning Algorithms have multiple levels of abstraction, making it suitable for churn prediction as they extract these features from the high-dimensional dataset with minimum end-user efforts [1, 5]. Moreover, deep learning has proved to be an appropriate choice in the case of large datasets.

In this chapter, we compare artificial neural network with traditional classification algorithms to predict the customer churn on IBM's Telco Customer Churn dataset. We achieve this by preprocessing and feeding the data into the constructed models and monitoring the concurrent performance of these models. Our aim is

---

S. Momin · T. Bohra (✉) · P. Raut  
D. J. Sanghvi College of Engineering, Mumbai, India

to accurately predict the churn and simultaneously show the difference in the performance of various algorithms.

## 20.2 Data Research

### 20.2.1 Dataset Selection

IBM Telco Customer Churn dataset, sourced from IBM, is utilized for this study. The dataset consists of various attributes, pertaining to the customer of a telecommunication company that helped us deduce a comprehensible relation between the customer’s behavior and the churn. The dataset consists of 21 attributes and 7044 rows. The deciding attribute: “churn,” describes whether the customer churned or not, indicated by a “Yes” or “No,” respectively. The attributes can further be classified into two fundamental categories: numeric attributes and object type attributes. This breakdown gave us an opportunity to perform tokenization of object type attribute that is discussed in Sect. 20.2.2 (Table 20.1).

### 20.2.2 Data Preprocessing

IBM Telco Customer Churn dataset consists of object type as well as numeric type attributes where in the object type comprises categorical and string type. We applied two separate techniques for preprocessing of these two attribute types: text tokenization for object types and standardization for numeric types, which resulted in the best accuracy of the trained model. The data is further divided into 9:1 ratio, where 90% dataset is utilized as the training dataset and the rest 10% is utilized as the testing dataset.

**Table 20.1** List of attributes

Customer unique ID	If multiple phone lines	If streaming movies
Gender—M/F	If Internet Service	Contract term type
If Senior Citizen	If Online Security	If paperless billing
If Partner exists	If Online backup	Payment method
If Dependents exists	If Device protection	Monthly charges
Tenure	If Technical support	Total charges
If Phone Service	If Streaming TV	Churn—Y/N

### 20.2.2.1 Tokenizer

Tokenization is a task of chopping up a sentence (character sequence) into subsequent pieces that are called tokens. Unique words from the sequence are identified and amalgamated in a list. Tokenization is highly efficacious in the preprocessing of text data.

We selected 16 out of 18 object types, for tokenization (two attributes left out are: “Total Charges” and “churn,” “Total charges” was converted to float type and used alongside the numeric values whereas “churn” is used as the output variable). We formed a character sequence by concatenating these attributes by the format: “attribute\_name” + “|” + “value” (where attribute\_name is the attribute name from the dataset, value is its given value and “|” is the separator required for tokenization).

Sample input (name, value) 1: (gender, Male), (Partner, No), (InternetService, DSL).

Sample output (sequence) 1: gender:Male|Partner:No|InternetService:DSL.

Determinately, we applied tokenization on the output character sequence (shown the sample output). The tokenization function returns vectorization of the text, by assigning them either a sequence of integers or a vector. We padded these sequences to avoid uneven lengths.

Sample input (Sample output 1) 2: gender:Male|Partner:No|InternetService:DSL

Sample output after tokenization 1: [7, 6, 24]

### 20.2.2.2 Standardization

The raw data has a varying magnitude for various attributes; due to this, the contribution of these attributes is not even and results in disproportionate outcomes. To evade this we apply standardization to the numeric type attributes to scale them to a common range. This is achieved by using the standardization formula as in Eq. (20.1)

$$z = \frac{x - \mu}{\sigma} \quad (20.1)$$

where  $x$  is the raw input,  $\mu$  is the mean, and  $\sigma$  is the standard deviation for that attribute.

This redistributes the values such that the mean becomes 0 and the standard deviation is set to 1.

## 20.3 Algorithms

### 20.3.1 Classical Algorithms

#### 20.3.1.1 Logistic Regression

Logistic regression is largely applied to solve problems with two class values. It uses the logit function that basically takes the natural logarithm of the odds. The logit in logistic regression is a special case of a link function in a generalized linear model.

$$\text{logit}(Y) = \text{natural log (odds)} = \ln \left( \frac{P}{1 - P} \right) = c + mx \quad (20.2)$$

In Eq. (20.2),  $c$  is the  $Y$  intercept,  $m$  denotes the regression coefficient, and  $P$  is the probability of the desired result.  $X$  can be both continuous and categorical but  $Y$  is restricted to categorical values [6].

#### 20.3.1.2 Naive Bayes

Naive Bayes is used majorly for real-time single and multi class prediction. It is used for real-time classification as it gives high accuracy and also speed when used on large databases. This algorithm is based on Bayes theorem which says the probability that a row  $X$  is associated with class  $C$ , is stated by Eq. (20.3)

$$p \left( \frac{C_i}{X} \right) = \frac{P \left( \frac{X}{C_i} \right) p(C_i)}{p(X)} \quad (20.3)$$

Since  $p(X)$  is constant only Eq. (20.4) needs to be maximized.

$$p \left( \frac{C_i}{X} \right) = P \left( \frac{X}{C_i} \right) p(C_i) \quad (20.4)$$

Naïve Bayes assumes that for a feature to exist in a class it is not necessary that another feature should also coexist. This is also called as class-conditional independence.

#### 20.3.1.3 Random Forest

Random forest is known to produce good results for prediction due to its default hyper parameters. Random forest is the combination of multiple decision trees merged together to give better prediction. Random forest randomly finds root node and splits the feature nodes while in decision trees this process is not random. The algorithm determines the final output by majority voting of all decision trees.

$$H(X) = \max \sum_{i=1}^k I(h_i(x) = Y) \quad (20.5)$$

In Eq. (20.5),  $H(X)$  gives classification result, and  $I$  is used to denote the utility function.  $h_i(x)$  is the classification result of an individual decision tree, whereas  $Y$  denotes the classification objection [7]. In this study, the amount of trees we have used in the forest is 300 and the splitting criterion used is the value entropy.

#### 20.3.1.4 K-Nearest Neighbors

In K-nearest neighbor classification, neighbors play a significant role determining the labels. KNN classifies the attribute by a majority vote of its neighbors, with the attribute being assigned to the class in which the majority of its  $k$  nearest neighbors are classified. We utilized the Euclidean distance formula as in Eq. (20.6) to calculate the distance between neighbors.

$$\sqrt{\sum_{i=1}^k (a_i - b_i)^2} \quad (20.6)$$

where  $a$  and  $b$  denote the two objects between which distance is to be calculated. In this study, the value of the number of neighbors used for  $k$  neighbors queries is 8 and the power parameter for the Euclidean metric is 2.

#### 20.3.1.5 Decision Tree

In decision tree algorithm the leaf nodes are taken as categories or classes (categorical or continuous values). The selection of decision tree algorithm is a crucial task. ID3 proved to be efficacious in our case. Although ID3 is highly efficient in making simple decision trees, the accuracy of this algorithm to compose good decision trees decreases with the increase in the complication.

Entropy ( $H$ ) is defined as lack of predictability in the dataset  $A$  as in Eq. (20.7)

$$H(A) = \sum_{b \in B} -p(b) \log_2 p(b) \quad (20.7)$$

where  $A$  is the dataset for which we are calculating the entropy and  $B$  denotes the collection of classes in  $A$  where  $B$  is either {yes} or {no},  $p(b)$  is number of elements in class  $b$  divided by number of elements in  $A$ . In ID3, entropy is calculated for all the remaining elements and the element with smallest is used for the split.

### 20.3.2 Deep Learning Algorithm

Deep learning consists of multiple nonlinear layers utilized for feature extraction and transformation, and for classification problems. It has several levels of representations, corresponding to a hierarchy of features that model complex relationships among data [8].

#### 20.3.2.1 Artificial Neural Network

Neural network is arguably the closest that humans have come in replicating the human mind in the computational world. These neurons are represented in the network using a large number of nodes and the interconnection between them is denoted by weighted edges. Every node is capable of receiving input, processing them, and feeding them as an output signal. ANN is a set of an adder to sum the input data according to some allocated weights depending on the strength, and an activation function which limits output of neuron [9] and weighted synapses.

The architecture is made up of an input layer, hidden layer(s), and output layer in the respective order. In Fig. 20.2,  $x_1, x_2, \dots, x_n$  are the features and  $a_1, a_2, \dots, a_n$  are the nodes of hidden layer. The activation function of the nodes in the hidden layer is used to compute the output to the next layer. We utilized the sigmoid (logistic) activation function. The predicted output class is denoted using  $y$ . Additional biased units  $x_0, a_0^2, \dots, a_0^{l-1}$  are added to the input layer and the hidden layers, respectively (Fig. 20.1).

$$a_i^{(j)} = g \left( W_{i0}^{j-1} x_0 + W_{i1}^{j-1} x_1 + \dots + W_{in-1}^{j-1} x_n \right) \tag{20.8}$$

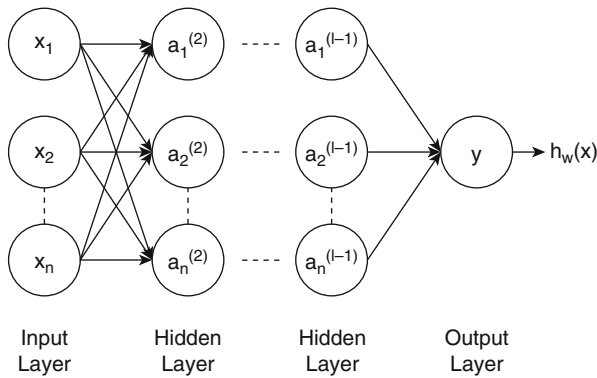


Fig. 20.1 Feed forward neural network architecture

In Eq. (20.8),  $j$  indicates the current layer and  $a_i^{(j)}$  indicates the activation of unit  $i$  in the current layer.  $W^j$  is the matrix of weights and  $n$  denotes the number of nodes in layer  $j - 1$ . The sigmoid activation function is given as:

$$g(z) = \frac{1}{1 + e^{-z}} \quad (20.9)$$

Sigmoid function is parameterized by  $z$ , where  $e$  is the universal constant. The output  $y$  is given by  $a_1^{(l)}$ , where  $l$  denotes the total count of the layers in the neural network.

### 20.3.2.2 Model Representation

We used feed forward multilayered artificial neural network to represent our model. Considering the amount of data available, the total number of hidden layers we have used is 12. The tokenized input from preprocessing is fed to the embedding layer of the model that is initialized with random weights and will learn the embedding for the input of the training data. The layer creates a table that contains the possible vector outputs. The index obtained by tokenization is used to map the corresponding values to these vectors. Further, we applied Spatial Dropout to the embedded input. The role of dropout is to improve the generalization performance by obviating activations from becoming tightly correlated, which causes overfitting [10]. We applied 1D Spatial Dropout that drops entire 1D feature maps, unlike dropout that eliminates individual elements. We found that applying 0.5 1D Spatial Dropout shaped the model better and provided better results. We have implemented a fully connected network that is prone to overfitting that hampers the generalization ability of the overall network. We mitigated this issue by implementing global average pooling, which avails us to select the significant parameters by cutting down on the total number of parameters, and consequently eliminates overfitting [11].

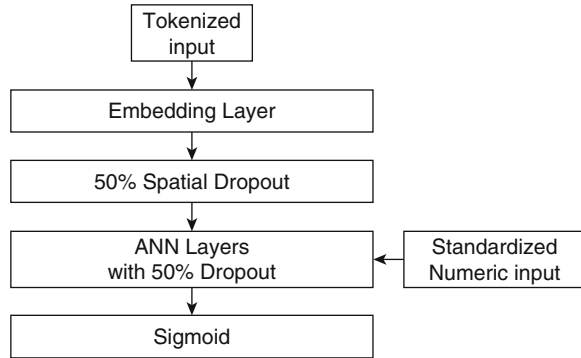
We then concatenated the numeric values to the model and appended the ANN layer that is activated utilizing ReLU activation, which uses (leaky) rectified linear unit. Dropout of 50% is applied to further prevent any overfitting of the model. Finally, we appended a layer that uses sigmoid activation that returns a binary output used to determine the churn (Fig. 20.2).

## 20.4 Results

As seen in Table 20.2, ANN model has the highest accuracy (as given by Eq. (20.10)) of 82.83%. ANN had performed better than the rest of the algorithms (traditional algorithms) in all the test runs. K-nearest neighbor gave better performance than the rest of the traditional algorithms, giving an accuracy of 79.86%.



**Fig. 20.2** Layers in model representation



**Table 20.2** The model accuracies on validation data

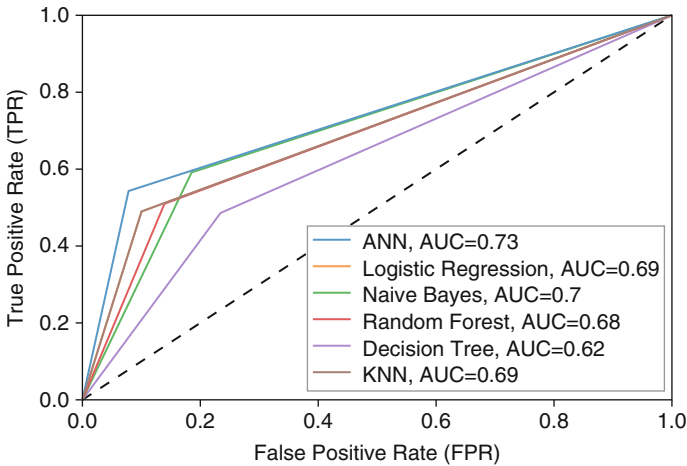
Logistic regression	Naïve Bayes	Random forest	Decision tree	K-nearest neighbor	ANN
78.87	76.45	77.87	73.05	79.86	<b>82.83</b>

Logistic regression gave an accuracy of 78.87% on average. The poorest results on our dataset were observed using the decision tree algorithm that gave an accuracy of 73%. Our ANN model was optimized using Adam optimizer. Adam differs from the classical stochastic gradient descent, as it calculates a separate adaptive learning rate for every parameter iteratively based on the training dataset [12], opposed to the classical stochastic gradient descent which assigns a fixed learning rate which is not updated with training data. Moreover, we utilized binary crossentropy to minimize the loss function during model training. Subsequently, our ANN model reached the highest performance observed on the validation data, achieving an accuracy of 84.54%.

$$\text{Accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \tag{20.10}$$

Here,  $\hat{y}_i$  is the predicted value whereas  $y_i$  is the corresponding true value for that element.  $n$  denotes the total number of samples in training dataset. The ratio of correctly predicted values and the total number of training values ( $n$ ), multiplied by 100 gives the accuracy of the model.

We further used ROC to evaluate the performance of our binary classifier. It depicts the tradeoff between the false positive rate and the true positive rate. ROC curve is utilized to calculate the area under ROC curve (AUC lies between 0 and 1), where 1 indicates 100% correct predictions and vice versa. In our study, as seen in Fig. 20.3, ANN achieved the highest AUC of 0.73.



**Fig. 20.3** ROC curve with individual AUC of all prediction models

## 20.5 Conclusion and Future Scope

In this chapter a study of five classification algorithms, namely logistic regression, naive Bayes, random forest, decision tree, and K-nearest neighbors, and a deep learning algorithm using artificial neural network to predict the customer churn on IBM's Telco Customer Churn dataset has been presented. The results of our experiments indicate that the deep learning algorithm implemented using artificial neural network performs better than the classical algorithms. The reason why ANN performs better than other classification algorithms is its ability to self-learn and detect complex nonlinear relationships between dependent and independent variable. We also noticed that in our case when the technique of tokenization was used for preprocessing the results were significantly better.

The limitation of this experiment was that the amount of data provided by IBM is comparatively less when compared to real-world customer data being generated.

For future work, since the deep learning model can handle increase in the amount of data as well as increase in the number of features, it is scalable [13]. This in turn will help correctly identify the causes of customer churn in large datasets, allowing effective strategies to be applied for customer retention. Even though we have demonstrated the success of the deep learning model for customer churn, it can potentially be applied to tasks like spam filtering [14], credit card fraud detection [15], and other fields where the amount of data keeps on increasing substantially.

## References

1. F. Castanedo, Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network (2014)
2. J. Hadden, A. Tiwari, R. Roy, D. Ruta, Computer assisted customer churn management: state-of-the-art and future trends. *Comput. Oper. Res.* **34**, 2902–2917 (2007). <https://doi.org/10.1016/j.cor.2005.11.007>
3. M. Hassouna et al., Customer churn in mobile markets a comparison of techniques. *CoRR* **8**, 224–237 (2015). arXiv:abs/1607.07792
4. C. Kirui, L. Hong, W. Cheruiyot, H. Kirui, Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. *Int. J. Comput. Sci. Issues.* **10**, 165–172 (2013)
5. M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar, et al., Deep learning applications and challenges in big data analytics. *J. Big Data* **2**(1), 1 (2015). <https://doi.org/10.1186/s40537-014-0007-7>
6. C.-Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **96**(1), 3–14 (2002). <https://doi.org/10.1080/00220670209598786>
7. Z. Li, S. Li, Random forest algorithm under differential privacy, in *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, Chengdu, 2017, pp. 1901–1905
8. L. Deng, D. Yu, *Deep Learning: Methods and Applications*, Foundations and Trends® in Signal Processing, vol 7(3–4) (Now Publishers, Hanover, MA, 2014), pp. 197–387. <https://doi.org/10.1561/20000000039>
9. Z. Zhang, Artificial neural network, in *Multivariate Time Series Analysis in Climate and Environmental Research*, (Springer, Cham, 2018)
10. G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors (2012)
11. M. Lin, Q. Chen, Network in network (2013), arXiv:1312.4400
12. D. Kingma, J. Ba, Adam: a method for stochastic optimization, in *International Conference on Learning Representations*, 2014
13. B.M. Wilamowski, B. Wu, J. Korniak, Big data and deep learning, in *2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES)*, Budapest, 2016, pp. 11–16
14. T. du Toit, H. Kruger, Filtering spam e-mail with generalized additive neural networks, in *2012 Information Security for South Africa, Johannesburg, Gauteng*, 2012, pp. 1–8
15. A.S. Bekirev, V.V. Klimov, M.V. Kuzin, et al., Payment card fraud detection using neural network committee and clustering. *Opt. Mem. Neural Networks* **24**, 193 (2015). <https://doi.org/10.3103/S1060992X15030030>

# Chapter 21

## Prediction of Crop Yield Using Fuzzy-Neural System



Bindu Garg and Tanya Sah

### 21.1 Introduction

Humans have always been allured to have the ability to foretell the future. Over the years this zeal has uplifted forecasting from the realms of voodoo science to a concrete mathematical subject. This has resulted in the extensive usage in the fields where foretelling reaps financial profits like economics, banking, and trading. Considering the today's pressing situation of the food industry it has been realized that having a crop forecast beforehand can be beneficial for the management of resources. This can help the government in better planning and preparation for the crop yield production shocks [1]. Moreover sustaining the burgeoning population is of concern in the twenty-first century. In one of its report, FAO has distinctly mentioned the situation is going to deteriorate in near future and steps should be taken to address it. Previously methods have been developed to forecast crop yield. They mostly depended on the arithmetic relationship between the yields or statistically calculated the production. Traditionally Autoregressive-moving-average model (ARMA) was plied extensively in the previous models for prediction. But this model experiences limitations while applying on the crop yield data set. Though it is an efficient model but not successful in the diminutive input dataset. It requires vast data for prediction whereas in the case of crop yield the data

---

B. Garg (✉)

Computer Engineering Department, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India  
e-mail: [bindu.garg@bharatividyaeeeth.edu](mailto:bindu.garg@bharatividyaeeeth.edu)

T. Sah

Tanya Sah Senior Software Engineer Globallogic India Pvt. Ltd, Noida, Uttar Pradesh, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_21](https://doi.org/10.1007/978-3-030-19562-5_21)

213

is seasonal and we have a very fewer data available for a long period of time. Thus, with the inception of the soft computing, it has been possible to deal with vague imprecise inputs.

## 21.2 Literature Survey

Thinking is a complex process. The way the mind deals with the inexplicit, imprecise, vague data with ease is peculiar [2–4]. The researchers have been beguiled for years to emulate the brain's behavior in electronic systems, but in actual fact, electronics is solely dependent on the classical logic which is a bi-valued function. Boolean logic is incapable of handling the randomness of the data.

The research in the forecasting domain can majorly be tracked down into the following four divisions decided on the basis of the distribution of universe of discourse. In the first category, researchers used the range of the dataset to identify the universe of discourse. The minimum and the maximum value of the sample data was wrapped around to delineate the universal set [5, 6]. In the second category, researchers used the distribution property of the sample data to derive the universe of discourse. In the third category comes the Aladag [7] who used the “Partition method of the optimal fuzzy set with optimization algorithms.” In the fourth category researchers majorly used the clustering algorithms. The clustering algorithms were used to determine the partition of the intervals according to the clustering results. Then, Song and Chissom [5, 6] put forth the fuzzy time series theory grounded on the fuzzy sets. They used fuzzy logical relationships for determining the associations between the observations. The prominent benefit of using a rule system is that it allows deriving conclusions even in the case of partial matching of rules. Song and Chissom gave both time variant and time invariant models for time series prediction. They used the model to forecast the enrollments in the University of Alabama's dataset.

Over the years several modified versions of the algorithms were developed by Huanrg [8, 9], Hwang et al. [10], and Lee et al. [11]. Chen [12] presented a simpler time series forecasting algorithm based on arithmetic calculations. This reduced the number of calculations for the max-min operation. Though this method was lucid and coherent, this failed to incorporate the dependency of the entire data set. Lee et al. going forward used fuzzy candlestick pattern [13] and a Japanese candlestick pattern to forecast financial decisions. To further simplify the convoluted matrix operation involved in fuzzy forecasting models, Huanrg [8, 9] suggested the multivariate heuristic model; this method had an added advantage of assimilating the results from multiple variables. Jilani et al. [14] improvised on this algorithm and implemented it on the car road accident data of Belgium to forecast the accidents rates.

In point of fact, accuracy is of prime importance in forecasting algorithms. In practice, existing models miss out to capture multitudinous aspects of the crop yield forecasting in a single model. Crop yield forecasting data is highly complex and is

**Table 21.1** Fuzzy sets

Fuzzy set	Linguistic variable
A <sub>1</sub>	Very low production
A <sub>2</sub>	Low production
A <sub>3</sub>	Average production
A <sub>4</sub>	Moderate production
A <sub>5</sub>	High production
A <sub>6</sub>	High production
A <sub>7</sub>	Extremely high production

impacted by disparate climatic factors and this makes it a favorable problem to be solved using fuzzy. Considering this we have utilized the results from our previous works [15] to devise a new model addressing the foils of the previous models in crop yield forecasting and effectuated it on the rice yield dataset.

### 21.3 Proposed Algorithm

This section elucidates a new method to forecast the crop yield using the interval based partitioning and using the observation data for the universe of discourse.

Step 1: Foremost it is necessary to define a universe of discourse for the dataset. Thus, for this scenario we have chosen the spread of the observation data to be its Universe of discourse. For the rice yield dataset the minimum production observed was 3219 and the maximum production observed was 4554. In this case  $U = [3200, 5000]$

Step 2: After the demarcation of the Universe of discourse, it is necessary to identify the fuzzy sets from the quantitative values of the crop yield dataset.

The crop yield historical data is in the quantitative form. Using the triangular membership functions map the production of each year into the fuzzified datasets (Tables 21.1 and 21.2).

Step 3: Next step is to determine the logical relationships between the fuzzified dataset. We have used the fuzzy If-Then Rule inference system to determine the correlation between the observations of a particular with other years with the similar relationship. Group all the fuzzy logical relationships obtained and count the repeated patterns just once. The Fuzzy Logical relationships are used to map the values from Cartesian space into the fuzzy interval. These are simplistic relations obtained by implying a basic conditional statement of IF-THEN RULE. This is done by using the membership functions. The strength of a relationship is determined by the order of association in a particular set of a value. The relationship thus obtained has been listed in Table 21.3.

Step 4: Now, using the logical relationships drawn in Table 21.3 calculate the relational matrix using the formula in 21.1.

**Table 21.2** Fuzzy sets

Year	Actual production (kg/ha)	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	A <sub>7</sub>	Fuzzified output
1981	3552	0.24	1	0.75	0	0	0	0	A <sub>2</sub>
1982	4177	0	0	0	0.71	1	0.28	0	A <sub>5</sub>
1983	3372	1	0.885	0	0	0	0	0	A <sub>1</sub>
1984	3455	0.17	1	0.82	0	0	0	0	A <sub>2</sub>
1985	3702	0	0.361	1	0.63	0	0	0	A <sub>3</sub>
1986	3670	0	0.337	1	0.66	0	0	0	A <sub>3</sub>
1987	3865	0	0	0.48	1	0.51	0	0	A <sub>4</sub>
1988	3592	0.27	1	0.72	0	0	0	0	A <sub>2</sub>
1989	3222	1	0.997	0	0	0	0	0	A <sub>1</sub>
1990	3750	0	0.398	1	0	0.62	0	0	A <sub>3</sub>
1991	3851	0	0	0.47	1	0.52	0	0	A <sub>4</sub>
1992	3231	1	0.99	0	0	0	0	0	A <sub>1</sub>
1993	4170	0	0	0	0.71	1	0.28	0	A <sub>5</sub>
1994	4554	0	0	0	0	1	0	0	A <sub>7</sub>
1995	3872	0	0	0.48	1	0.51	0	0	A <sub>4</sub>
1996	4439	0	0	0	0	0.91	1	0	A <sub>7</sub>
1997	4266	0	0	0	0.78	1	0.21	0	A <sub>6</sub>
1998	3219	1	0	0	0	0	0	0	A <sub>1</sub>
1999	4305	0	0	0	0.81	1	0.18	0	A <sub>6</sub>
2000	3928	0	0	0.53	1	0.46	0	0	A <sub>4</sub>

**Table 21.3** Fuzzy logical relationships

A <sub>2</sub> →A <sub>5</sub>	A <sub>5</sub> →A <sub>1</sub>	A <sub>1</sub> →A <sub>2</sub>	A <sub>2</sub> →A <sub>3</sub>	A <sub>3</sub> →A <sub>4</sub>
A <sub>4</sub> →A <sub>2</sub>	A <sub>2</sub> →A <sub>1</sub>	A <sub>1</sub> →A <sub>3</sub>	A <sub>3</sub> →A <sub>4</sub>	A <sub>4</sub> →A <sub>1</sub>
A <sub>1</sub> →A <sub>5</sub>	A <sub>5</sub> →A <sub>7</sub>	A <sub>7</sub> →A <sub>4</sub>	A <sub>4</sub> →A <sub>7</sub>	A <sub>7</sub> →A <sub>6</sub>
A <sub>6</sub> →A <sub>1</sub>	A <sub>1</sub> →A <sub>6</sub>	A <sub>6</sub> →A <sub>4</sub>		

$$R1 = A_1^T \times A_2 \tag{21.1}$$

The fuzzy forecast for any year is obtained by using the model mentioned in Eq. (21.2). To predict the yield of the current year, calculate the effective fuzzified value of the previous year. For this find the A<sub>i-1</sub> which is the fuzzified forecasted of the previous year. The fuzzy forecasted output is obtained by performing the fuzzy min-max composition between last three-year production and R' (the overall relational matrix).

The model used for the forecasting is

$$A_i = A_{i-3} \cdot A_{i-2} \cdot A_{i-1} \circ R' \tag{21.2}$$

**Table 21.4** Fuzzified forecast of rice production

Actual production	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	A <sub>6</sub>	A <sub>7</sub>
3552	–	–	–	–	–	–	–
4177	–	–	–	–	–	–	–
3372	–	–	–	–	–	–	–
3455	1	1	0.75	0.71	1	0.28	0
3702	1	1	0.82	0.71	1	0.28	0
3670	1	1	1	0.63	0	0	0
3865	0.17	1	1	0.66	0	0	0
3592	0	0.36	1	1	0.51	0	0
3222	0.27	1	1	1	0.51	0	0
3750	1	1	0.72	1	0.51	0	0
3851	1	1	1	0	0.60	0	0
3231	1	0.99	1	1	0.60	0	0
4170	1	0.99	1	1	0.60	0	0
4554	1	0.99	0.47	1	1	0.28	0
3872	1	0.99	0	0.71	1	0.28	0
4439	0	0	0.48	1	1	0.28	0
4266	0	0	0.48	1	1	1	0
3219	0	0	0.48	1	1	1	0
4305	1	0	0	0.78	1	1	0

where  $A_i$  is the forecasted production of the current year,  $i$  represents the year, so  $A_{i-1}$  is the production of the previous year, and ‘o’ is the max–min operator. The values thus obtained in this step are listed in Tables 21.4 and 21.5.

Step 5: Defuzzification is the last and the most vital step in the process of forecasting. It is the process by which a fuzzy value is transformed into a crisp value. Defuzzification finally gives us a meaningful forecast value. It optimizes the systems performance index by converting it into Boolean value. We used neural network for the defuzzification of the results and simulated the complete network in the matlab.

For training the data set we used transig as the first transfer function and purelin as the second transfer function for the second layer. The forecasted values thus obtained has been listed in Table 21.6. This lists out a comparison of the forecasts from the Chen [16]

## 21.4 Performance Comparison

This segment tests the efficacy of the propounded model by analyzing and comparing the results with different existing forecasting models. To simplify the results we have graphed all the observations in the form of line charts and bar graphs. In the proposed model, fuzzy logical relationships have been used to determine



**Table 21.5** Defuzzified forecasted production

Actual production ( $A_i$ ) (kg/ha)	Fore. prod ( $F_i$ ) (kg/ha)	Error in fore.
4177	–	
3372	–	
3455	3627.78	0.05000868307
3702	3550.423	0.04094462453
3670	3697.287	0.007435149864
3865	3733.616	0.03399327296
3592	3636.952	0.01251447661
3222	3338.524	0.03616511484
3750	3913.322	0.04355253333
3851	3840.991	0.002599065178
3231	3472.66	0.07479418137
4170	3912.861	0.06166402878
4554	4257.482	0.06511155029
3872	3943.359	0.0184294938
4439	4448.611	0.002165127281
4266	4328.121	0.01456188467
3219	3328.121	0.03389903697
4305	4269.104	0.008338211382
3928	3988.0269	0.01528182281
AFER		0.03067401516

**Table 21.6** Average forecasting error rate

Forecasting method	AFER	MSER
Proposed method	3.06%	20676.06
Chen	13.48%	132162.9

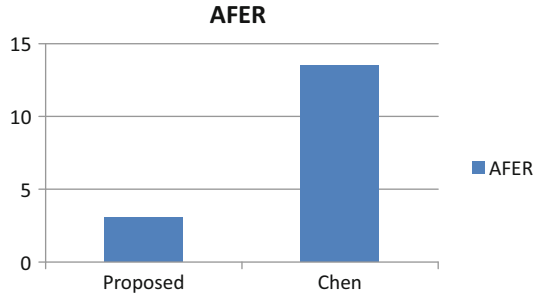
the correlation of the forecast with the previous yield. Comparing the implemented model with the other model the former clearly performs better.

To further corroborate the results, the average forecasting error rate and mean square error has been calculated for the models. Table lists out the AFER and MSE for the proposed models.

The average forecasting error rate in the predicted values can be determined by calculating the mean absolute percentage error. This computation provides a comprehensive view of the deviation from the expected value in terms of percent.

A low AFER value implicates accurate results. Thus, from Table 21.6 it can be established that the proposed model performs better than the existing models and forecast has been closest to the actual yields.

**Fig. 21.1** Average forecasting error rate comparison chart



## 21.5 Conclusion

In this chapter, a fuzzy time series forecasting model has been introduced to predict crop yield based upon the interval based partition, fuzzy logical relationships, and neural networks. The model was applied on historical rice yield dataset obtained from the University of Pantnagar and compared the results against the existing algorithms. Comparison graphs clearly show a reduced AFER and MSE in the prediction through the new model. Though the model has been effectuated on the crop yield it can be successfully applied in the other forecasting domain because of the data independence of the model. This prediction methodology will be an advantage for the government agencies in proper planning and resource administration of the crops. In future, this approach can be expanded for other forecasting domains.

## References

1. C. Musvoto, K. Nortze, B. de Wet, B.K. Mahumani, A. Nahman, Imperatives for an agricultural green economy in South Africa. *S. Afr. J. Sci.* **111** (2015)
2. L.A. Zadeh, The concept of linguistic variable and its application to approximate reasoning-I. *Inform. Sci.* **8**, 199–249 (1975)
3. L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 1 (1973)
4. L.A. Zadeh, Fuzzy sets. *Inform. Contr.* **8**, 338–353 (1965)
5. Q. Song, B.S. Chissom, Fuzzy time series and its models. *Fuzzy Set. Syst.* **54**, 269–277 (1993)
6. Q. Song, B.S. Chissom, Forecasting enrollments with fuzzy time series: Part II. *Fuzzy Set. Syst.* **62**, 1–8 (1994)
7. U. Yolcu, E. Egrioglu, R.V.R. Uslu, M.A. Basaran, C.H. Aladag, A new approach for determining the length of intervals for fuzzy time series. *Appl. Soft Comput.* **9**, 647–651 (2009)
8. K. Huarng, Effective lengths of intervals to improve forecasting in fuzzy time series. *Fuzzy Set. Syst.* **12**, 387–394 (2001)
9. K. Huarng, Heuristic models of fuzzy time series for forecasting. *Fuzzy Set. Syst.* **123**, 369–386 (2002)

10. J.R. Hwang, S.M. Chen, C.H. Lee, Handling forecasting problems using fuzzy time series. *Fuzzy Set. Syst.* **100**, 217,228 (1998)
11. L.W. Lee, L.W. Wang, S.M. Chen, Handling forecasting problems based on two-factors high-order time series. *IEEE Trans. Fuzzy Syst.* **14**, 468–477 (2006)
12. S.M. Chen, Forecasting enrollments based on high order fuzzy time series. *Int. J. Cybernetics Syst.* **33**, 1–16 (2002)
13. C.H.L. Lee, A. Lin, W.S. Chen, Pattern discovery of fuzzy time series for financial prediction. *IEEE Trans. Knowledge Data Eng.* **18**, 613–625 (2006)
14. T.A. Jilani, S.M.A. Burney, C. Ardil, Multivariate high order fuzzy time series forecasting for car road accidents. *Int. J. Comput. Intell.* **4**, 15–20 (2007)
15. B. Garg, S. Aggarwal, J. Sokhal, Crop yield forecasting using fuzzy logic and regression model. *Comput. Electrical Eng.* **67**, 383–403 (2017). <https://doi.org/10.1016/j.compeleceng.2017.11.015>
16. S.M. Chen, Forecasting enrollments based on fuzzy time series. *Fuzzy Set. Syst.* **81**, 311–319 (1996)

# Chapter 22

## Speed Estimation and Detection of Moving Vehicles Based on Probabilistic Principal Component Analysis and New Digital Image Processing Approach



T. V. Mini and V. Vijayakumar

### 22.1 Introduction

The primary reason of accidents in India is vehicle speed violation. The vehicle speed control is a challenging task during travel. Violating the traffic rules such as red light violation is the other primary reason for the accidents which increases the possibility of crashing. The detection of overspeed vehicles and traffic signal violation [1] are essentially automated in India. In smart city traffic management, automatic overspeed vehicle detection and traffic rule violation detection system need to be implemented. Computer vision techniques and pattern recognition plays a significant role in speed detection.

The detection of moving vehicles speed and traffic violation are the main components of an intelligent surveillance and smart city traffic management. Vehicle traffic can be detected with the high volume of vehicles. So, it is essential to identify and estimate speed for knowing road profile based on the vehicle speed. There are many studies for detecting, classifying and counting vehicle objects. Smart traffic monitoring system should have the ability to detect, calculate and estimate vehicle speed [2].

The city traffic monitoring systems collect data regarding vehicle speed while processing in real time. The recent literature presents a review of moving vehicle detection algorithms and it concludes that an accurate vehicle speed detection method need to be developed. The precise speed measurement is dependent on accuracy of detection of vehicles in the video frame. The common vehicle detection

---

T. V. Mini  
Sacred Heart College, Chalakudy, Kerala, India

V. Vijayakumar (✉)  
Sri Ramakrishna College of Arts and Science, Coimbatore, Tamil Nadu, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_22](https://doi.org/10.1007/978-3-030-19562-5_22)

221

equipment are video processing systems, microwave radars, inductive loop detectors and active infrared detectors [3].

Measuring the vehicle speed can be categorized into instantaneous and average speed depending on the methodology employed for the speed estimation. The speed measures at a single point using instantaneous methods. The laser technologies and sensors are bounded in pairs under the road surface. The sensors are positioned at a static position. At least two cameras are installed in a minimum of hundreds of metres [4] for average speed detection.

In this chapter, a novel method for detecting vehicle speed with Probabilistic Principal Component Analysis and digital image processing approach is proposed. The proposed method detected vehicle speed from the surveillance video using the three-stage method. The Spatio-temporal Varying Filter is used to preprocess the extracted frames. Contour finding algorithm is used for the detection of the vehicle. The structure of the proposed chapter is as follows: The detailed review of literature is discussed in Sect. 22.2. Section 22.3 outlined the proposed detection technique of moving vehicles and speed estimation. Experimental results are discussed in Sect. 22.4. Section 22.5 concluded the chapter with future work.

## 22.2 Literature Review

Al-Turjman [5] presented a vehicular speed learning framework for the best traffic load with a latency and throughput conditions. The system communicated with the driver about the speed violation through an in-vehicle wireless receiver. This system gives the high-level intelligence to the driver. It is also used in smartphones as an intelligent sensor. This method is learned autonomously to manage efficient transportation.

Wang et al. [6] proposed a Generalized Cross-Correlation (GCC) approach to detect moving vehicles driven in lanes using two seismic sensors. This method estimates the time delay of arrival. An urban street environment data set was used to examine the vehicle detection accuracy. Guido et al. [7] presented a technique to track moving objects such as vehicles which integrates video processing. The Global Positioning System (GPS) was used as a benchmark data. The results of two case studies show good performance than the existing techniques. However, the detection and tracking of the moving objects will vary from each other. This work is applied only for tracking of moving objects without considering preprocessing step. So it doesn't correctly detect the vehicles accurately.

Kumar and Kushwaha [8] presented a method for detecting and estimating the moving vehicle speed. The vehicles are tracked with position and maintained in the record. The average of detection accuracy was 87.7%. This approach was not suitable for measuring the speed of the vehicles and produced less accurate results.

Wei and Yang [9] discussed a dynamic threshold detection algorithm for adaptable road vehicle detection and speed estimation. They analysed the detection accuracy with different traffic conditions based on tri-axial anisotropic magnetoresistive sensors and wireless sensor network. It gives good detection accuracy and reliability compared with the fixed threshold algorithm.

Wu et al. [10] described a method for speed estimation of individual vehicles. It contains object detection, object tracking and speed estimation. This method improved the accuracy of tracking via a dynamic update of templates and local predictive search. The computational complexity is high when compared to other speed estimation methods.

Rajab et al. [11] proposed a vehicles classification method to recognize the vehicles in a traffic lane by applying a single-element piezoelectric sensor. The lane passing vehicles are detected using diagonal locations of the sensor. It gives 97% classification accuracy in highway sites. The sensor installation procedures, data collection methods and vehicle classification algorithm are briefly discussed. The complexity of the system is increased when refereed with other speed estimation methods. This method produced a high performance for vehicle classification and minimized lane work [10].

Li et al. [12] presented a method for measuring the vehicle speed using video images. With reference to the advanced calibration lines, instantaneous tandem photographs were taken based on the location of the vehicle and charge-coupled device (CCD). The speed measurement algorithm is applied on number plate images. This method will not provide good solution for real-time applications.

Lan et al. [13] developed a Vehicle Speed Measurement (VSM) method. The improved three-frame difference algorithm is applied to detect the contour of moving vehicles. The grey constraint optical flow algorithm is used to compute the speed of the vehicle. The velocity of the vehicles is calculated by the optical flow value of the vehicle contour. This method also measured the corresponding ratio of the image pixels and width of the road. The method gives good optical flow field by minimizing the influence of changing lighting and shadow. It decreases computation by computing the moving target contour's optical flow value. The precision of vehicle speed is moderate for real-time applications.

Jeyabharathi and Dejeey [14] proposed a Vehicle Tracking and Speed Measurement (VTSM) system to discover vehicles speed. This method is used to judge the accidents at a less cost. This method detects the foreground and tracks specified object and computes the object speed. Recognizing the stationary background from moving objects in a surveillance video is a challenge task. Diagonal Hexadecimal Pattern (DHP) is used for dynamic background subtraction and object tracking. The performance of the system is analysed with Metric F-score and Multiple Object Tracking Accuracy (MOTA).

In the existing system, the conventional de-noising filters reduce the noise variance in smooth regions. But it doesn't provide perfect edges in object boundaries. Computing the significant outliers develops more difficulty.

### 22.3 Proposed Methodology

In this chapter, Principal Component Analysis and Digital Image Processing technique are proposed to vehicle movement. The system extracts the video frames from a live video stream. Then, proposed systems detect and track the vehicle with three steps. First, The Probabilistic Principal component analysis (PPCA) is used to detect multiple outliers in frames. Alternate coordinate set is calculated to identify the outliers in video frames. Spatio-temporal Varying Filter (STVF) is applied to preprocess the extracted frames. The contour detection algorithm is used for the detection of the vehicle. The Frame Count Algorithm is used to estimate the vehicle speed. This approach presents high vehicle detection accuracy with high precision and recall rate. The proposed method is shown in Fig. 22.1.

In cloudy time, in dark side area vehicle movement, a region is created in the vehicle area. This dark area produced in the vehicle is the Region of Interest (ROI) as shown in Fig. 22.2. The rectangle denotes the ROI in the image. The frames are extracted from live video stream and are stored as the new database. Spatio-temporal Varying Filter (STVF) is applied on extracted frames to carry out the preprocessing the frame sequence.

#### 22.3.1 Probabilistic Principal Component Analysis (PPCA)

PCA is a Gaussian density model used to reduce dimensions. Maximum-likelihood estimate is calculated for elements related with principal components [15, 16].

Fig. 22.1 Proposed system—flow diagram

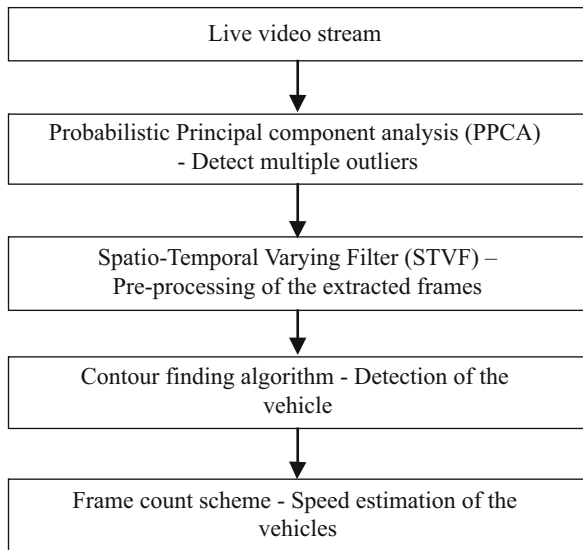




Fig. 22.2 Region of interest

The factor analysis of the Latent variable prototype with linear relationship is given below:

$$\begin{aligned}
 &y \sim Wx + m + \varepsilon \\
 &\text{Variables defined for Latent : } x \sim N(0, I) \\
 &\text{Noise or Error denoted as : } \varepsilon \sim N(0, \psi) \\
 &\text{The mean of Location term : } \mu
 \end{aligned}
 \tag{22.1}$$

An ‘n’-dimensional data are fixed by Principal Component Analysis (PCA). Each axis of the ellipsoid indicates a principal component. If the axis of the ellipsoid is minor value, then the variance along that axis is also lesser. The corresponding principal component of ellipsoid has adequately less amount of information by ignoring the axis. The axes of the ellipsoid are calculated by the mean of each variable. It is detected from the dataset to centre of the data on the origin. It computes the covariance matrix and defines the covariance matrix eigenvalues. The normalized orthogonal eigenvectors develop unit vectors. The orthogonal unit eigenvectors are interpreted as an ellipsoid axis. The choice of basis will transform the covariance matrix into a diagonalized form with the diagonal elements. It represents the variance of each axis. The proportion of the eigenvector variance is calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues [16].

Probabilistic Principal Component Analysis (PPCA):

The Noise variances are controlled to be identical ( $\psi_i = \sigma^2$ ) [17]

The error is measured as:

$$\begin{aligned}
 &\varepsilon \sim N(0, \sigma^2 I) \\
 &y | x \sim N(WX + \mu, \sigma^2 I) \\
 &y \sim N(\mu, Cy),
 \end{aligned}
 \tag{22.2}$$

The covariance matrix for the observed data ‘y’(Cy) is:

$$WWT + \sigma^2 I$$

The PPCA creates progressive different principal components (axes). PPCA is a covariant under alternation of the original data axes.



### 22.3.2 Spatio-temporal Varying Filter (STVF)

The reference frame spatial filtering consists of a sliding window through the image. The window's RGB pixels mean value is calculated and qualified to the window's central pixel value [18]. The temporal filtering method is applied to simplify the balance 'P' Group of Pictures (GOP) frames based on the reference frame. The frame blocks are modified based on their corresponding blocks in the reference image. The block matching algorithm is used to search the Group of Picture similarity to the previously filtered frame. The score matching R is calculated using formula (22.3):

$$R(x, y) = \frac{\sum_{x', y'} [ROI(x', y') - F_{N-1}(x + x', y + y')]^2}{\sqrt{\sum_{x', y'} ROI(x', y')^2 \cdot \sum_{x', y'} F_{N-1}(x + x', y + y')^2}} \quad (22.3)$$

The score matching 'R' is calculated by the square normalized matching method. Zero is a best match and a large value is worst. If the matching succeeds, then all the region of interest's pixels will have the similar values of its similar region's pixels. Else, region of interest is spatially filtered.

### 22.3.3 Contour Finding Algorithm

The Simple Boundary Follower (SBF) method is a straightforward contour-tracing approach. The position of 'S' is recorded to tracer. The tracer 'T' moves in either left or right direction 'd' based on the block 'P'. If the pixel tracer is situated on a contour, the tracer transfers left direction 'd<sub>Left</sub>'; otherwise, it proceeds right 'd<sub>Right</sub>' [19]. The method is discussed in the following algorithm.

#### Algorithm: Simple Boundary Follower (SBF)

1. Procedure SBF
2. Tracer (P,d) ← S(P,d) //where P black
3. Perform the subsequent steps
4. If Block 'P' = black then
5. Tracer (P,d) ← Tracer(P<sub>Left</sub>, d<sub>Left</sub>) //Black Left P<sub>Left</sub>
6. Else Tracer(P,d) ← Tracer(P<sub>Right</sub>, d<sub>Right</sub>)
7. while Tracer(P,d) ≠ S(P,d)

## 22.4 Experimental Results

The data set consists of 18 HD videos, each around one hour duration, captured at six different places. The total vehicle instances in the videos are 20865. LiDAR

is used to interpret the precise speed measurements from optical gates. It is tested with several reference GPS tracks [20]. The implementation is done using MATrixLABoratory (MATLAB). The proposed algorithm performance is evaluated with three performance metrics such as accuracy, precision and recall.

### 22.4.1 Accuracy

It is the percentage of the correctly predicted outcome.

$$\text{Accuracy} = \frac{(\text{Correctly Predicted} + \text{Correctly Not Predicted})}{(\text{Correctly Predicted} + \text{Wrongly Not Predicted} + \text{Wrongly Predicted} + \text{Correctly Not Predicted})} \quad (22.4)$$

where Correctly Not Predicted is a True Negative instance where the case was positive and predicted is negative; Correctly Predicted is a True Positive instance where the case was positive and predicted is also positive; Wrongly Not Predicted is a False Negative instance; Wrongly Predicted is a False Positive instance.

$$\text{Precision} = \frac{\text{Correctly Predicted}}{(\text{Correctly Predicted} + \text{Wrongly Predicted})} \quad (22.5)$$

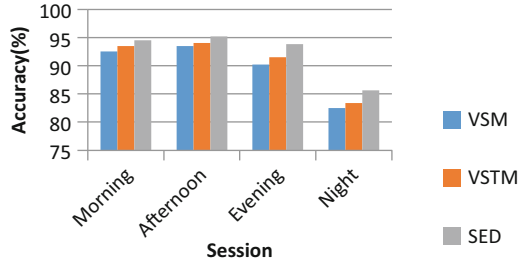
$$\text{Recall} = \frac{\text{Correctly Predicted}}{(\text{Correctly Predicted} + \text{Wrongly not Predicted})} \quad (22.6)$$

For analysing, the proposed method ‘Speed Estimation and Detection’(SED) performance is compared with the two existing methods that are VSM (Lan et al., [13]) and VSTM (Jeyabharathi and Dejeey [14]) for four sessions with ten videos for each session.

Figure 22.3 shows the comparison of three different methods performance and are measured in terms of accuracy. In these four sessions the percentage is reduced in morning up to night. The proposed SED yields higher accuracy results of 94.5%, 95.2%, 93.8% and 85.6% for morning, afternoon, evening and night, respectively. It concludes that the proposed work performs better when compared to other methods as shown in Table 22.1.

The performance analysis for precision SED method gives a good result than the VSM and VSTM shown in Fig. 22.4. The proposed method produces higher precision results of 96%, 97%, 91% and 87% for morning, afternoon, evening and night, respectively. It concludes that the proposed SED method performs better when compared to other methods as shown in Table 22.2.

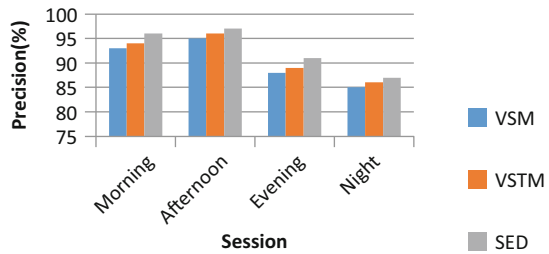
**Fig. 22.3** Performance analysis for accuracy



**Table 22.1** Performance analysis for accuracy

Number of videos	Session	No. of frames	Total no. of vehicles	Accuracy (%)		
				VSM	VSTM	SED
10	Morning	200	8	92.5	93.5	94.5
10	Afternoon	210	10	93.5	94.0	95.2
10	Evening	185	9	90.2	91.5	93.8
10	Night	127	6	82.5	83.4	85.6

**Fig. 22.4** Performance analysis for precision

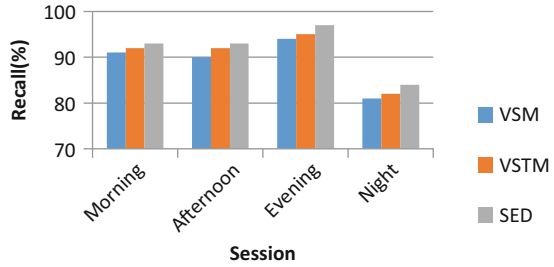


**Table 22.2** Performance analysis for precision

Number of videos	Session	No. of frames	Total no. of vehicles	Precision (%)		
				VSM	VSTM	SED
10	Morning	200	8	93	94	96
10	Afternoon	210	10	95	96	97
10	Evening	185	9	88	89	91
10	Night	127	6	85	86	87

The performance analysis for recall SED method gives a good result than the VSM and VSTM shown in Fig. 22.5. The proposed SED produces higher recall results of 93%, 93%, 97% and 84% for morning, afternoon, evening and night, respectively. The proposed SED method performs better when compared to other methods as shown in Table 22.3.

**Fig. 22.5** Performance analysis for recall



**Table 22.3** Performance analysis for recall

Number of videos	Session	No. of frames	Total no. of vehicles	Recall (%)		
				VSM	VSTM	SED
10	Morning	200	8	91	92	93
10	Afternoon	210	10	90	92	93
10	Evening	185	9	94	95	97
10	Night	127	6	81	82	84

## 22.5 Conclusion and Future Work

This chapter proposed a novel approach for detecting and tracking estimation of the speed. The proposed method is verified and tested on four different databases in videos. It gives high vehicle detection accuracy with high precision and recall rate. It optimally minimizes the wrongly prediction on both sides of the road. The proposed method can be adopted in the current traffic system. In the future work, vehicle detection can be performed using the machine learning classifiers to improve the detection accuracy. Also parallel methods have been introduced to reduce the computational complexity of the classifiers.

## References

1. B.C. Putra, Moving vehicle classification with fuzzy logic based on image processing, Master’s thesis, Sepuluh Nopember Institute of Technology (2016)
2. V. Markevicius, D. Navikas, A. Idzkowski, A. Valinevicius, M. Zilyis, D. Andriukaitis, Vehicle speed and length estimation using data from two anisotropic magneto-resistive (AMR) sensors. *Sensors*17(8), 1–13 (2017)
3. D.F. Llorca, C. Salinas, M. Jimenez, I. Parra, A.G. Morcillo, R. Izquierdo, J. Lorenzo, M.A. Sotelo, Two-camera based accurate vehicle speed measurement using average speed at a fixed point, in *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (2016), pp. 2533–2538
4. F. Al-Turjman, Vehicular speed learning in the future smart-cities’ paradigm, in *IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops)* (2017), pp. 61–65.
5. H. Wang, W. Quan, X. Liu, S. Zhang, A two seismic sensor based approach for moving vehicle detection. *Procedia-Social Behav. Sci.* **96**, 2647–2653 (2013)

6. G. Guido, V. Gallelli, D. Rogano, A. Vitale, Evaluating the accuracy of vehicle tracking data obtained from unmanned aerial vehicles. *Int. J. Transp. Sci. Technol.* **5**(3), 136–151 (2016)
7. T. Kumar, D.S. Kushwaha, An efficient approach for detection and speed estimation of moving vehicles. *Procedia Comput. Sci.* **89**, 726–731 (2016)
8. Q. Wei, B. Yang, Adaptable vehicle detection and speed estimation for changeable urban traffic with anisotropic magnetoresistive sensors. *IEEE Sens. J.* **17**(7), 2021–2028 (2017)
9. W. Wu, V. Kozitsky, M.E. Hoover, R. Loce, D.T. Jackson, Vehicle speed estimation using a monocular camera, in *Video Surveillance and Transportation Imaging Applications*. International Society for Optics and Photonics (2015), Vol. 9407, pp. 704–940
10. S. Rajab, M.O. Al Kalaa, H. Refai, Classification and speed estimation of vehicles via tire detection using single-element piezoelectric sensor. *J. Adv. Transp.* **50**(7), 1366–1385 (2016)
11. Y. Li, L. Yin, Y. Jia, M. Wang, Vehicle speed measurement based on video images, in *3rd International Conference on Innovative Computing Information and Control* (2008), pp. 439–439
12. J. Lan, J. Li, G. Hu, B. Ran, L. Wang, Vehicle speed measurement based on gray constraint optical flow algorithm. *Optik-Int. J. Light Electron Optics* **125**(1), 289–295 (2014)
13. D. Jeyabharathi, D. Deje, Vehicle Tracking and Speed Measurement system (VTSM) based on novel feature descriptor: Diagonal Hexadecimal Pattern (DHP). *J. Vis. Commun. Image Represent.* **40**, 816–830 (2016)
14. H. Abdi, L.J. Williams, Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**(4), 433–459 (2010)
15. R. Bro, A.K. Smilde, Principal component analysis. *Anal. Methods* **6**(9), 2812–2831 (2014)
16. M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis. *J. Roy. Stat. Soc. B* **61**(3), 611–622 (1999)
17. J. Seo, S. Chae, J. Shim, D. Kim, C. Cheong, T.D. Han, Fast contour-tracing algorithm based on a pixel-following method for image sensors. *Sensors* **16**(3), 353 (2016)
18. A.B. Hamida, M. Koubaa, H. Nicolas, C.B. Amar, Spatio-temporal video filtering for video surveillance applications, in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (2013), pp. 1–6
19. A.B. Hamida, M. Koubaa, H. Nicolas, C.B. Amar, Spatio-temporal video filtering for video surveillance applications, in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, (2013), pp. 1–6
20. <https://medusa.fit.vutbr.cz/traffic/research-topics/traffic-camera-calibration/brnocompspeed/>

## Chapter 23

# A Posture Recognition System for Assisted Self-Learning of Yoga by Cognitive Impaired Older People for the Prevention of Falls



K. Ponmozhi and P. Deepalakshmi

### 23.1 Introduction

Falls are dangerous for people who are suffering from osteoporosis. Bones of the people with this disease will be very weak and it may break bone if they fall. The fall issue is most common among elders. An estimate in [1] states that, low- and middle-income countries face 75% fall injuries and especially this is a serious issue in Asian countries among older people [2].

The risk of falling is higher in older people those who are suffering from cognitive impairment compared to the other normal old age people [3]. Studies [4, 5] reveal that deterioration in cognition is one of the factors which can influence the risk of fall.

The risk of falling increases when people with fear of falling will modify their gait [6]. The risk of falls might be reduced by 35% through practicing exercises regularly [7], and interventions [8]. In [9], authors suggested that cognitive restructuring in addition to exercise may give better results than using exercise alone. Improving executive functions through exercise, dual-task training, or cognitive offer better results [10].

Yoga uses a series of physical postures called asanas, breathing control, and meditation. Since Yoga concentrates on both body and mind, it is considered as

---

K. Ponmozhi (✉) · P. Deepalakshmi

Department of Computer Applications, School of Computing, Kalasalingam Academy of Research and Education, Srivilliputhur, Tamil Nadu, India

Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Srivilliputhur, Tamil Nadu, India

e-mail: [ponmozhi@klu.ac.in](mailto:ponmozhi@klu.ac.in); [deepa.kumar@klu.ac.in](mailto:deepa.kumar@klu.ac.in)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_23](https://doi.org/10.1007/978-3-030-19562-5_23)

231

more therapeutic than as an exercise [11]. While doing Yoga, one should stretch major muscle groups, and so it attributes to physical strength and flexibility [12].

A community survey reveals that yoga practice improves leg strength, balance and mobility [13–15] and also the functional measure which have been identified as predictors of fall risk [16]. Yoga can be practiced at home alone or in an organized setting as a group. Yoga is also most suitable for self-learning [17].

Many interventions based on yoga are observed during yoga practice as a community event. Some older people showed discomfort of doing yoga in public view. It is preferable to provide yoga practice in a closed location without any disturbances. Also most yoga trainers are not interested to go to rural areas [18].

This chapter is a part of assisted self-learning yoga practice focusing on identifying the sitting and standing postures which are the basis for almost all asanas. Since these two postures are the basis for all the asanas, we initially tried to identify these postures in this chapter. An Orbbec Astra sensor device is used to capture practitioner's various postures during their yoga session in the form of skeletal information. Angle values of the postures will be calculated and compared with the database. If there is any difference between observed and actual angles values, it will be intimated in the visual form enabling the yoga practitioner to self-learn the asana postures properly.

## 23.2 Related Work

Human skeletal joint coordinates can be accessed easily and reliably from depth sensors. Using the skeletal information, human action can be classified either by applying signal-processing techniques [19] or based on joint coordinates [20].

Sometimes spatial and temporal dictionaries of human parts are used to represent actions [21]. In [22], authors have used 3D geometric relationships of body parts for classification. The moving pose descriptor is defined in [23] to capture postures and skeletal joints. Assistive technology gives major benefits to disabled persons. Assistive technology act defines assistive technology devices as any item, piece of equipment, or product system, whether acquired commercially, modified, or customized, that is used to increase, maintain, or improve functional capabilities of individuals with disabilities [24].

## 23.3 Proposed Work

### 23.3.1 Background

In order to model human bodies, skeleton parts such as head, neck (which are responsible for head movements), arms, hands (to decide on hand postures), and

**Fig. 23.1** Skeletal bone joints in NuiTRACK



legs and feet (for standing posture identification) have been modeled and stored as a vector. Each feature is defined as a set of movable joints. NuiTRACK is the software frame which can be used to calculate joint angles as in Fig. 23.1. The Augmented Reality support provided in this software can be used to instruct and show the correct posture for assisted learning of yoga. At a time, up to six user skeletons can be tracked by [NuiTRACK Skeleton Tracker](#) module. A user whose skeleton is currently being tracked is called an active user. The maximum number of tracked skeletons can be changed via `SkeletonTracker.SetNumActiveUsers`.

NuiTRACK determines the key points of a human body and forms a skeleton shape as a set of 3-component vectors with 0 at the point of sensor location. Sensor recognizes up to 19 points of the skeleton. By passing the `nuiTRACK.Joint` object to function `ToVector3()`, we can get the joint coordinates.

Since, almost all yoga postures are done either in standing posture or in sitting posture, we decided to identify the same. For example, some of the yoga asanas which depend on standing posture are *Utthanasana* (standing forward bend), *Padottanasana* (standing wide legged forward bend), and *Parsvottanasana* (side



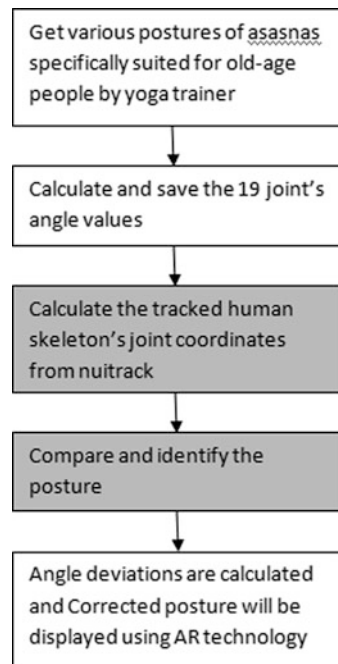
stretch pose). The asanas which depend on sitting posture are Paschimottanasana (seated forward bend) and Janusirsasana (head to knee pose).

### 23.3.2 Flow of Work

In Fig. 23.2, a high-level diagram of the proposed system has been shown, in which the gray area signifies our work. We consider the asanas which are specifically suitable for older people. These asanas will improve the flexibility of joint muscles and improve the balance and motor ability of joints. Angle values of 19 joints in these postures are calculated and stored in a vector.

During yoga practice, human skeleton details will be tracked and joint coordinates are accessed. These values are used to calculate joint angles. These angle values are matched with stored values and determine whether the practitioner is sitting or standing.

**Fig. 23.2** Flow of work  
(source: Nuitrack software website)



### 23.3.3 Angle Calculation

Skeletal angles can be segment angles (like trunk, thigh, leg, and foot angles) or joint angles (like hip, knee, and ankle angles). Joint angle is the smaller angle between two adjacent segments, often termed as anatomical angles as shown in Fig. 23.3a. Figure 23.3b shows the example points  $P_1(x_1 = 4, y_1 = 10)$  and  $P_2(x_2 = 6, y_2 = 4)$  and sample calculation of segment angles.

From segment angles, we can calculate joint angles such as hip angle as in Eq. (23.1).

$$\theta_{hip} = \theta_{thigh} - \theta_{trunk} \tag{23.1}$$

Similarly, other segment and joint angles can also be calculated. Biomechanical angle limits of sitting posture are shown in Table 23.1, and standing posture is shown in Table 23.2.

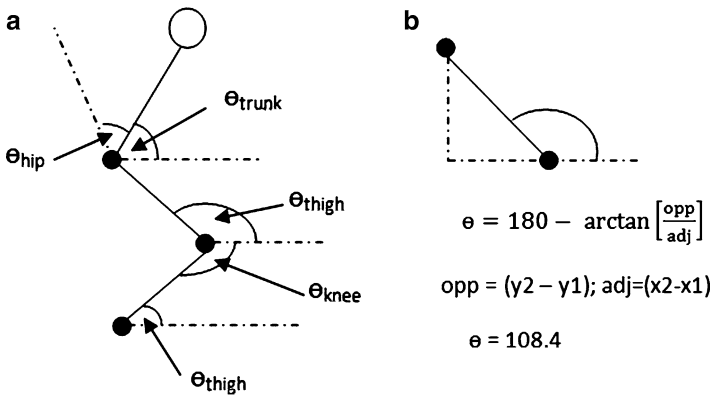


Fig. 23.3 (a) Angle calculation. (b) Example

Table 23.1 Sitting posture angle limits of hip

Angle	Sitting posture
10–80	Sitting: bending forward
85–100	Straight sitting
133–138	Leaning back
+145	Sleeping or tendency to standing pose

Table 23.2 Standing posture angle limits of hip

Angle	Standing posture
$150 < x < 175$	Standing: bending forward
+180 or –180	Straight standing
+190	Leaning back

## 23.4 Results

Joint angle values of three specified standing and sitting postures as specified in Tables 23.1 and 23.2 are stored in a database. Five volunteers were asked to sit in the three different sitting postures and also for three different standing postures for five times. For each of the recorded test case, the joint and segment angles are calculated and compared with the entries in corresponding table. Positive results are averaged for each of the six postures. We calculated recall and precision and out of 50 test, we obtained recall value as 96% and precision as 98%.

## 23.5 Conclusion

Being active during the entire life is essential for self-contained life especially for elderly people. Yoga provides safe as well as multiple good effects while restructuring cognitive skills of older people. In this chapter, the sitting and standing postures which are basic postures of all asanas related to prevention of falls were considered for identification. As a future work, we plan to consider biomechanical limitations of all joints for specific asana postures. Apart from providing skeletal information, Nutrack has a module to handle augmented reality issues. We are planning to use this module for handling corrective action of self-learning practice.

## References

1. H. Hestekin, T. O'Driscoll, J. Stewart Williams, P. Kowal, K. Peltzer, S. Chatterji, *Measuring prevalence and risk factors for fall-related injury in older adults in low- and middle-income countries: results from the WHO Study on Global AGEing and Adult Health (SAGE)* (World Health Organization, Geneva, 2013)
2. J. Jagnoor, L. Keay, R. Ivers, A slip and a trip? Falls in older people in Asia. *Injury* **44**(6), 701–702 (2013)
3. F.E. Shaw, Prevention of falls in older people with dementia. *J. Neural Transm.* **114**, 1259–1264 (2007)
4. K.J. Anstey, C. von Sanden, M.A. Luszcz, An 8-year prospective study of the relationship between cognitive performance and falling in very old adults. *J. Am. Geriatr. Soc.* **54**, 1169–1176 (2006)
5. D.B. Welmerink, W.T. Longstreth Jr., M.F. Lyles, A.L. Fitzpatrick, Cognition and the risk of hospitalization for serious falls in the elderly: results from the Cardiovascular Health Study. *J. Gerontol. A* **65**, 1242–1249 (2010)
6. T.T. Huang, M.L. Chung, F.R. Chen, et al., Evaluation of a combined cognitive-behavioural and exercise intervention to manage fear of falling among elderly residents in nursing homes. *Aging Ment. Health* **20**, 2–12 (2016)
7. C. Sherrington, A. Tiedemann, N. Fairhall, J.C. Close, S.R. Lord, Exercise to prevent falls in older adults: an updated meta-analysis and best practice recommendations. *NSW Pub. Health Bull.* **22**(3–4), 78–83 (2011)

8. A. Kumar, K. Delbaere, G.A. Zijlstra, et al., Exercise for reducing fear of falling in older people living in the community: cochrane systematic review and meta-analysis. *Age Ageing* **45**, 345–352 (2016)
9. S. Tennstedt, J. Howland, M. Lachman, et al., A randomized, controlled trial of a group intervention to reduce fear of falling and associated activity restriction in older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **53**, P384–P392 (1998)
10. O. Segev-Jacobovski, T. Herman, G. Yogev-Seligmann, A. Mirelman, N. Giladi, J.M. Hausdorff, The interplay between gait, falls and cognition: can cognitive therapy reduce fall risk? *Expert Rev. Neurother.* **11**, 1057–1075 (2011)
11. A.S. Chan, Y.C. Ho, M.C. Cheung, M.S. Albert, H.F. Chiu, L.C. Lam, Association between mind-body and cardiovascular exercises and memory in older adults. *J. Am. Geriatr. Soc.* **53**, 1754–1760 (2005)
12. M. Van Puymbroeck, L.L. Payne, P.C. Hsieh, A phase I feasibility study of yoga on the physical health and coping of informal caregivers. *Evid. Based Complement. Alternat. Med.* **4**, 519–529 (2007)
13. K.P. Roland, J.M. Jakobi, G.R. Jones, Does yoga engender fitness in older adults? A critical review. *J. Aging Phys. Act.* **19**(1), 62–79 (2011)
14. M. Krishnamurthy, S. Telles, Effects of yoga and an Ayurveda preparation on gait, balance and mobility in older persons. *Med. Sci. Monit.* **13**(12), LE19–LE20 (2007)
15. A. Tiedemann, S. O'Rourke, R. Sesto, C. Sherrington, A 12-week Iyengar yoga program improved balance and mobility in older community-dwelling people: a pilot randomized controlled trial. *J. Gerontol. A Biol. Sci. Med. Sci.* **68**(9), 1068–1075 (2013)
16. A. Tiedemann, H. Shimada, C. Sherrington, S. Murray, S. Lord, The comparative ability of eight functional mobility tests for predicting falls in community-dwelling older people. *Age Ageing* **37**(4), 430–435 (2008)
17. S. Boros, B. Csala, E. Szilágyi, Yoga practice for the elderly: good choice to avoid falls. *J. Exerc. Sports Orthoped.* **5**(1), 1–4 (2018)
18. M. Sullivan, M. Leach, J. Snow, S. Moonaz, Yoga's effect on falls in rural, older adults. *Complement. Ther. Med.* **35**, 57–83 (2017)
19. T. Batabyal, A. Vaccari, S.T. Acton, Ugrasp: A unified framework for activity recognition and person identification using graph signal processing, in: *2015 IEEE International Conference on Image Processing (ICIP)*, (2015), pp. 3270–3274
20. T. Batabyal, T. Chattopadhyay, D.P. Mukherjee, Action recognition using joint coordinates of 3d skeleton data, in *Image Processing (ICIP), 2015 IEEE International Conference on, IEEE* (2015), pp. 4107–4111
21. C. Wang, Y. Wang, A.L. Yuille, Mining 3d key-pose-motifs for action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2639–2647
22. R. Vemulapalli, F. Arrate, R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group, in *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 588–595
23. M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in *Time of Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ed. by M. Grzegorzec, C. Theobalt, R. Koch, A. Kolb, (Springer, Berlin, 2013), pp. 149–187
24. L. Columna, L.J. Lieberman, R. Lytle, K. Arndt, Special education terminology every physical education teacher should know. *J. Phys. Edu. Recreation Dance* **85**(5), 38–45 (2014)

# Chapter 24

## Improved UFHLSNN (IUFHLSNN) for Generalized Representation of Knowledge and Its CPU Parallel Implementation Using OpenMP



Priyadarshan S. Dhabe and Sanman D. Sabane

### 24.1 Introduction

Pattern classification involves classification of an input pattern into correct class. Dataset to be classified will contain multiple patterns with numeric feature vectors and associated classes. When such patterns are served as an input to a training algorithm, it will learn a classification model. Upon application of a new pattern this classification model classifies the pattern. Pattern classification is used in many domains such as medical diagnosis, weather prediction, and fraud detection character recognition. There are various types of classifiers suggested in [10, 11] but hybrid fuzzy neural classifiers are found better suited for real-world complex pattern recognition tasks due to their ability to learn like humans.

#### 24.1.1 Introduction to Fuzzy Neural Network Hybrid System

The meaning of term fuzzy is imprecise. For handling imprecise information, Sir Lofti Zedah introduced fuzzy sets [2], which can be described as follows. Let  $X$  be a nonempty set. A fuzzy set  $A$  in  $X$  is characterized by its membership function  $A : X \rightarrow [0; 1]$  and  $A(x)$  is interpreted as the degree of membership of element  $x$  in fuzzy set  $A$  for each  $x \in X$ . The value zero is used to represent complete non-membership, one is used to represent complete membership, and values in between zero and one are used to represent partial memberships.

---

P. S. Dhabe (✉) · S. D. Sabane

Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, India

e-mail: [priyadarshan.dhabe@vit.edu](mailto:priyadarshan.dhabe@vit.edu)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_24](https://doi.org/10.1007/978-3-030-19562-5_24)

239

Artificial Neural Networks [3] are parallel structures which are made up of multiple processing elements connected to each other with the links. Each link has a specific property that can be represented with weights assigned to that link. As the execution progresses, these weights get adjusted using some learning rule. These qualities of neural network provide various features such as learning like humans (experiential learning), adaptation, learning on –fly and fault tolerance ability.

## 24.2 Parallel Computing with CPU

### 24.2.1 Introduction to OpenMP

OpenMP (Open Multi-Processing) [12] is an Application Programming Interface (API) that supports multi-platform shared memory multiprocessing programming in C, C++, and Fortran [13]. It consists of a set of compiler directives, library routines, and environment variables that influence runtime behavior [14, 15]. OpenMP is an implementation of multithreading, a method of parallelizing whereby a master thread forks a specified number of slave threads and a task is divided among them. Figure 24.1 illustrates fork join model in OpenMP. The runtime environment allocates threads to different processors and threads then run concurrently. The section of code that is meant to run in parallel is marked accordingly, with a preprocessor directive that will cause the threads to form before the section is executed [15].

By default, each thread executes the parallelized section of code independently. To divide a task among the threads, work-sharing constructs can be used so that each thread executes its allocated part of the code.

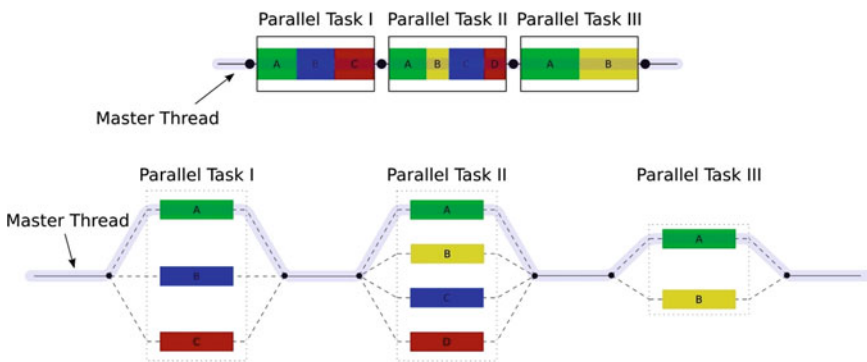


Fig. 24.1 Multithreading with OpenMP (source: <https://en.wikipedia.org/wiki/OpenMP>)

### 24.2.2 IBM POWER8 [16]

Power Architecture is a registered trademark for similar Reduced Instruction Set Computing (RISC) instruction sets for microprocessors developed and manufactured by such companies as IBM, Freescale/NXP, AppliedMicro, LSI, Teledyne e2v, and Synopsys. The governing body is [Power.org](http://Power.org), comprising over 40 companies and organizations [17]. The name “POWER” was originally presented as an acronym for “Performance Optimization With Enhanced RISC” [17].

Released in August 2013 at the Hot Chips conference, POWER8 is a superscalar symmetric multiprocessors based on the POWER Architecture. The designs are available for licensing under the OpenPOWER Foundation, which is the first time for such availability of IBM’s highest-end processors [17]. POWER8 is designed to be a massively multithreaded chip where each core has eight threads implemented using simultaneous multithreading (what IBM calls SMT8) [18].

### 24.3 Proposed IUFHLSNN

We proposed improved UFHLSNN called IUFHLSNN for generalizing the learned knowledge obtained in the form of generated HLS for better recognition. The HLS are calculated in the same manner as of the UFHLSNN [5]. Each HLS is defined with two  $n$  dimensional end points  $v = (v_1, v_2, \dots, v_n)$  and  $w = (w_1, w_2, \dots, w_n)$  are stored in matrix  $V$  and  $W$ . After obtaining the HLS we generalize the obtained knowledge by calculating the middle point of every HLS and store it in matrix  $M$ . Data stored in matrix  $M$  represents the knowledge in a coarse manner [4, 6]. These mid-points are used as HLS in the recall phase. For an  $i$ th HLS its mid-point is calculated and stored as described in (24.1)

$$M [i, j] = \left[ \frac{v_j + w_j}{2} \right] \quad \text{for } j = 1, 2, \dots, n \quad (24.1)$$

For parallel implementation with OpenMP we launch every thread which is available with the respective CPU. We used two test benches for our implementation powered by two different CPU’s, namely Intel’s Xeon E5-2620 and IBM’s POWER8. With IBM POWER architecture we had 2 POWER8 CPU on a single machine with 10 cores each. The POWER8 processor is 8 way threaded (SMT8) which gives us total 160 threads to work with. With Intel x86 architecture we had 2 Intel Xeon E5-2620 with 6 cores each. The Intel’s Xeon E5-2620 is 2 way threaded which gives us 24 threads to work with. The `pragma omp parallel` directive launches all the available threads with the CPU and divides the workload within those threads.

If you have 1000 iteration of an operation running in a OpenMP parallel for loop with 24 threads available on the machine, then each thread will get  $1000/24$ , i.e., approx. 42 iterations to process. In our implementation each thread with assigned range of iterations calculates membership to their HLS and reduction clause helps us to find out the maximum membership.

### 24.3.1 IUFHLSNN Training Algorithm

The IUFHLSNN training algorithm is supervised and is same as UFHLSNN and is described in Algorithm 1. It takes input as training patterns and outputs the HLSs learned out of them. Threshold  $0 < \theta \leq 1$  that governs maximum size of HLS is a user-defined parameter that need to be decided in learning. Let training set contain  $\lambda$  patterns, by adjusting threshold  $\theta$ , we can control the number of HLS's created, and thus the classification rate.

In training phase an input pattern  $R_h$  is applied and its membership is calculated with all existing HLS's. Let  $\alpha$  be the number of existing HLS's. If the index of pattern  $R_h$  is less than  $\psi$ , membership is calculated serially, otherwise it is calculated in parallel using OpenMP on CPU. If no HLS's are created before application of  $R_h$ , where  $1 \leq h \leq \lambda$ , then a new HLS is created from it. If the pattern index meets the condition  $h > \psi$ , then membership of  $h$ th training pattern will be calculated in parallel on CPU in all existing HLS's. By using *reduction* [14] clause gets the index of HLS giving maximum membership to this input pattern. In this way HLS are created in the training phase.

### 24.3.2 UFHLSNN Recognition Algorithm

Recognition algorithm is used to identify correct class of supplied pattern. Based upon trained neural network recognition rate may vary. We parallelize the recognition algorithm with OpenMP. We have obtained mid-points of all the created HLS, as in (24.1), and stored it in matrix  $M$ . Then, the test patterns are applied. For each testing pattern we launch all the available threads with the CPU to calculate membership of  $i$ th pattern in all existing HLS. We are using *reduction* [14] clause while launching multiple threads which will find out the maximum membership of  $i$ th pattern within the set of formed HLS and stored in  $M$ . Let we have  $\rho$  number of patterns in the testing phase and  $\alpha$  HLS are there in matrix  $M$ .



**Algorithm 1: OpenMP UFHLSNN training algorithm**

**Require:** All training pattern and matrices  $V$  and  $W$  for storing two end point of created HLS's

**Ensure:** HLS's is created and stored in matrices  $V$  and  $W$

```

1: Begin
2: for  $i=1$  to  $\lambda$  do
3:   if  $i < \psi$  then
4:     for  $j=1$  to  $\alpha$  do
5:       Calculate membership of  $i^{th}$  pattern to  $j^{th}$  HLS
6:     end for
7: Step 1: Compute index of HLS  $\beta$  awarding maximum membership to  $i^{th}$ 
   Pattern.
8: Step 2: If  $i^{th}$  pattern falls on  $\beta^{th}$  HLS then no need to update  $V$  and  $W$ .
   Otherwise try to extend  $\beta^{th}$  HLS to include pattern  $i^{th}$  as per [1]. If we cannot
   extend this HLS to include  $i^{th}$  input pattern, then create a new HLS and
   increase  $\alpha$  by 1.
9:   end if
10: Step 3: Launch available CPU threads to calculate membership of  $i^{th}$  pattern
   in all  $\alpha$  HLS.
11: Step 4: Use reduction clause with OpenMP for getting index  $\gamma$  of all HLS
   giving maximum membership to the  $i^{th}$  pattern.
12: Step 5: Go to Step 2.
13:   Update newly created HLS.
14: end for
15: End

```

**Algorithm 2: IUFHLSNN recognition algorithm**

**Required:** Mean of matrix  $V$  and  $W$  stored in matrix  $M$

**Ensure :** Identity correct class of a pattern

```

1: Begin
2: for  $i=1$  to  $\rho$  do
3:   Step 1: Launch available CPU threads to calculate membership of  $i^{th}$ 
   pattern in all  $\alpha$  HLS stored in  $M$ .
4:   Step 2: Use reduction clause with OpenMP for getting index  $\gamma$  of all HLS
   giving Maximum membership to the  $i^{th}$  pattern.
5:   Step 3: Assign  $\gamma$  HLSs class to current pattern.
6: end for
7: end for
8: Calculate percentage recognition rate.
9: End

```

## 24.4 Experimental Results

In this section we compare serial and parallel execution time required for classification and recognition. We also compare parallel execution time on different CPU's. The specification of both the test bench is tabulated in Table 24.1. Each test had  $2 \times$  CPU of the following specification which resulted in 160 usable threads on IBM POWER8 machine and 24 usable threads on Intel Xeon E5-2620 machine.

We used three datasets for our experimentation. The details of the dataset are tabulated in Table 24.2. All the instances in the dataset are not used for both the classification and recognition stage. We calculated speed up and percentage gain in time using (24.2) and (24.3.).

$$\text{Speed up} = \left( \frac{\text{Serial execution time}}{\text{Parallel execution time}} \right) \quad (24.2)$$

$$\% \text{Gain in time} = \left( \frac{\text{Serial execution time} - \text{Parallel execution time}}{\text{Serial execution time}} \right) \times 100\% \quad (24.3)$$

The results of serial and parallel experimentation for every dataset are summarized in Tables 24.3, 24.4, and 24.5, respectively. Both the results are executed on the test bench powered by Intel Xeon E5-2620. Table 24.3 shows the serial and parallel runtime comparison for skin segmentation dataset for both the classification and recognition phase. We achieved 21.68 times speed up in classification phase and 19 times speed up in recognition phase. We observed 95.38 % gain in time in classification phase and 94.73% gain in time in recognition phase. Total 81,476 HLS were created. We achieved 99.50% classification rate and 98.83% recognition rate.

**Table 24.1** Test bench specification

Parameter	IBM's POWER8	Intel's Xeon E5-2620
No. of cores	10	6
No of threads	80	12
Clock speed	3.6 GHz	2.5 GHz
RAM	256 GB	32 GB
Operating system	Ubuntu 16.04	CentOS 7
Compiler	gcc	gcc

**Table 24.2** Dataset details

Dataset	Total instances	Instances used	No. of attributes	No. of classes
Poker [7]	1,025,010	300,000	10	10
QtyT40I10D100K [8]	3,960,456	300,000	3	10
Skin segmentation [9]	245,057	245,057	3	2

**Table 24.3** Skin segmentation dataset serial vs. parallel runtime comparison (Intel XeonE5-2620)

Task	Serial time (s)	Parallel time (s)	Speed up	%Gain in time
Classification	5574	257	21.68	95.38
Recognition	228	12	19	94.73

**Table 24.4** QtyT40I10D100K dataset serial vs. parallel runtime comparison (Intel XeonE5-2620)

Task	Serial time (s)	Parallel time (s)	Speed up	%Gain in time
Classification	4479	281	15.93	93.72
Recognition	176	1	176	99.43

**Table 24.5** Poker dataset serial vs. parallel runtime comparison (Intel XeonE5-2620)

Task	Serial time (s)	Parallel time (s)	Speed up	%Gain in time
Classification	16,252	1223	13.28	92.47
Recognition	110	3	36.66	97.27

Table 24.4 shows the serial and parallel runtime comparison for QtyT40I10D100K dataset for both the classification and recognition phase. We achieved 15.93 times speed up in classification phase and 176 times speed up in recognition phase. We observed 93.72% gain in time in classification phase and 99.43% gain in time in recognition phase. Total 87,489 HLS were created. We achieved 51.33% classification rate and 40.71% recognition rate.

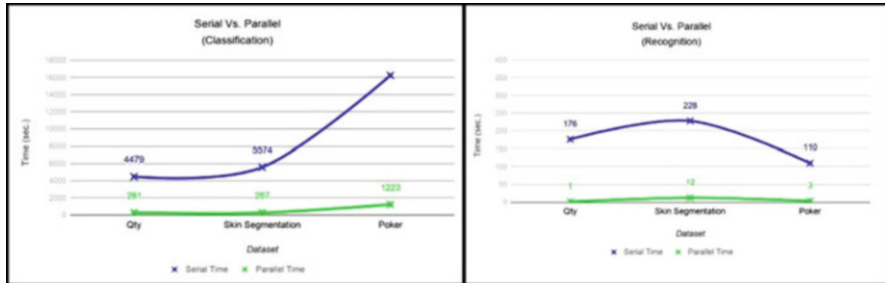
Table 24.5 shows the serial and parallel runtime comparison for Poker dataset for both the classification and recognition phase. We achieved 13.28 times speed up in classification phase and 36.6 times speed up in recognition phase. We observed 92.47% gain in time in classification phase and 97.27% gain in time in recognition phase. Total 149,936 HLS were created. We achieved 76.06% classification rate and 50.30% recognition rate.

For the comparison of parallel execution time of two architectures IBM's POWER8 and Intel's Xeon E5-2620, we used 245,057 instances of skin segmentation dataset [9], 900,000 instances of QtyT40I10D100K [8] dataset, and 300,000 instances of Poker dataset [7]. We found that POWER8 is 2.57 times faster in classification w.r.t. Intel's x86 and 3 times faster in recognition phase for skin segmentation dataset. Similarly, POWER8 is found 2.68 times faster in classification stage w.r.t. Intel's x86 and 2.44 times faster in recognition phase for QtyT40I10D100K dataset. For Poker dataset we used 300,000 instances and observed that POWER8 is 1.74 times faster than Intel's x86 architecture in classification and 1.6 times faster in recognition phase. Thus, we recommend IBM's POWER8 over Intel Xeon E5-2620 for parallel implementation of IUFHLSNN on CPU (Table 24.6).

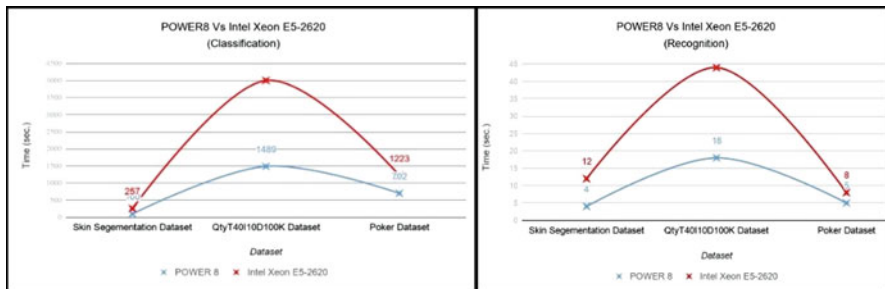
Figure 24.2 shows serial and parallel time plot for classification and recognition with every dataset used. We also compared the parallel runtime comparison on different architectures. Figure 24.3 shows classification and recognition time comparison between POWER8 and Intel Xeon E5-2620

**Table 24.6** Parallel runtime (in s) comparison of IBM’s POWER8 and Intel’s Xeon E5-2620

Dataset	Task	Execution time on IBM’s POWER8	Execution time on Intel’s Xeon E5-2620
Skin segmentation	Classification	100	257
	Recognition	4	12
QtyT40I10D100K	Classification	1489	4001
	Recognition	18	44
Poker	Classification	702	1223
	Recognition	5	8



**Fig. 24.2** Comparison of serial and parallel execution time in classification and recognition



**Fig. 24.3** Timing comparison of POWER8 and Xeon E5-2620 for classification and recognition

From Fig. 24.2, we can say that, on average, the obtained speedup is 2.37 in classification phase and 2 times in recognition phase.

From observation of Fig. 24.3 one can conclude that POWER8 is on average, 2.3 times faster than Xeon E5-2620.

The fact that a single core can spawn into 8 thread on the IBM POWER 8, gave it an edge over Intel Xeon E5-2620. Although IBM POWER8 has 160 cores in total, the serial part of initial HLS creation (serial) executed slower when compared to Intel Xeon E5-2620. But when it comes to parallel execution IBM POWER8 took a substantial lead because of the ability to spawn 8 threads (SMT8) from a single core.

## 24.5 Conclusion

In this chapter we proposed IUFHLSNN along with its learning and recall algorithms. In IUFHLSNN generalized knowledge is used in recall phase in terms of mid-points of HLS. The proposed IUFHLSNN is proven superior to UFHLSNN in terms of recognition rate.

OpenMP parallelization of the IUFHLSNN with generalized knowledge gave us average speedup of 16.96 times in classification phase and 77.22 times speedup in recognition phase with all the datasets used. The percentage gain obtained on average is 92% for classification phase and 94% for recognition phase.

As per our comparison of IBM POWER 8 and Intel Xeon E5-2620, we can conclude that IBM POWER8 is 2.33 times faster than Intel's Xeon E5-2620 in classification and 2.34 times faster in recognition phase for all the used datasets.

## References

1. U. V. Kullarni, T. R. Sontakke, G. D. Randale, Supervised fuzzy hyper line segment neural network for rotation invariant handwritten character recognition, in *IJCNN 01, International Joint Conference on*, vol. 4, pp. 2918–2933, 2001.
2. L.A. Zadeeh, Fuzzy logic computing with words. *IEEE Trans. Fuzzy Syst.* **4**(2), 103 (May 1996)
3. Jacek M. Zurada, Neuron modelling for artificial neural systems, in *Fundamental Concepts and Models of Artificial Neural Systems* (West Publisher Company, St. Paul, 1992), ch. 2, sec. 2.1, pp. 30–36
4. P. M. Patil, P. S. Dhabe, T. R. Sontakke, Recognition of handwritten characters using modified fuzzy hyper line segment neural network, in *The 12 IEEE International Conference*, vol. 2, Aug 2003
5. P.S. Dhabe, Vyas Prashant, Kulkarni Aditya, Pattern classification using Updated Hyperline Segment Neural Network and its GPU parallel implementation for large datasets using CUDA, in *2016 International Conference on Computing, Analytics and Security Trends*, 2016
6. A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Occam's razor. *Inform. Process. Lett.* **24**(6), 377–380 (1987)
7. Poker Dataset. <https://archive.ics.uci.edu/ml/datasets/Poker+Hand>
8. QtyT40I10D100K DataSet. <https://archive.ics.uci.edu/ml/datasets/QtyT40I10D100K>
9. Skin Segmentation Dataset. <https://archive.ics.uci.edu/ml/datasets/skin+segmentation>
10. R. Lippmann, An Introduction to computing with neural nets. *IEEE ASSP Mag.* **4**(2), 4–22 (1987)
11. J. Vieira, F.M. Dias, A. Mota, Neuro-fuzzy systems: a survey. *Measurement* **67**, 126–136 (2015)
12. OpenMP Tutorial at Supercomputing, 2008. <https://www.openmp.org/uncategorized/openmp-tutorial-at-supercomputing-2008/>
13. OpenMP Compilers. [OpenMP.org](http://OpenMP.org). 10 Apr 2013. Retrieved 14 Aug 2013
14. Using OpenMP—Portable Shared Memory Parallel Programming. <https://www.openmp.org/uncategorized/download-book-examples-and-discuss/>
15. OpenMP: A Proposed Industry Standard API for Shared Memory Programming [Online]. Available: [http://www.cse.iitd.ernet.in/openmp\\_p5.pdf](http://www.cse.iitd.ernet.in/openmp_p5.pdf)

16. POWER8 Processor User's Manual for the Single-Chip Module. [https://www.setphaserstostun.org/power8/POWER8\\_UM\\_v1.3\\_16MAR2016\\_pub.pdf](https://www.setphaserstostun.org/power8/POWER8_UM_v1.3_16MAR2016_pub.pdf)
17. POWER Architecture. [https://en.wikipedia.org/wiki/Power\\_Architecture](https://en.wikipedia.org/wiki/Power_Architecture)
18. H.B. Bakoglu, G.F. Grohoski, R.K. Montoye, The IBM RISC System/6000 processor: hardware overview. *IBM J. Res. Dev.* **34**(1), 12–22 (1990). <https://doi.org/10.1147/rd.341.0012>

# Chapter 25

## Performance Evaluation of Multihop Multibranch DF Relaying Cooperative Wireless Network



M. Dayanidhy and V. Jawahar Senthil Kumar

### 25.1 Introduction

For achieving high data rate and reliable communication, a network model designed with multiuser–single antenna handset background is called cooperative network. The model consists of a source node (S), multiple intermediate relay nodes (R), and a destination node (D). We utilize the intermediate relay node to receive symbol from source and forward to destination node. Emerging cooperative networks as a feasible solution to close the gap in the end-to-end data rate and transmission range. By selection combining of symbols received through intermediate relay nodes, spatial diversity is exploited at destination. The performance of relay node channels is studied in [1–3].

Four major relaying protocols are used by relay nodes in cooperative networks. Amplify and Forward (AF): The relay nodes amplify the received symbol and forward to the neighbor node, not considering the error in the symbol. Decode and Forward (DF): the intermediate relay nodes regenerate the source symbol and forward the encoded symbol of his confirm to the other node. Compress and Forward (CF): the compressed version of received symbol is forward to neighbor node, compression technique is not unique to all nodes. Coded Cooperation (CC): forwarding of channel coded symbol to other nodes. We consider the intermediate relay nodes are operated with DF protocol.

The performance of multihop single branch relaying without diversity are studied in [4], where the end-to-end probability outage of lower bound of Nakagami

---

M. Dayanidhy (✉)  
ECE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

V. Jawahar Senthil Kumar  
ECE, CEG, Anna University, Chennai, Tamil Nadu, India  
e-mail: [veerajawahar@annauniv.edu](mailto:veerajawahar@annauniv.edu)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_25](https://doi.org/10.1007/978-3-030-19562-5_25)

249

channel fading is derived. In [5], end-to-end bit error probability and the outage probability is studied for four different cooperative channel models with simple AF cooperative protocol. In [6], the study of lower bound error performance of a one hop, multihop and multibranch cooperative diversity with DF relaying is derived. The analysis of asymptotic error probability of coherent demodulator for a single branch cooperative is derived.

In this scenario, we use cumulative distributed function and probability density function of the SNR to exact the SEP of the two proposed model using DF relaying. We derive the derivation for SEP of single branch multihop DF relay network in Sect. 25.2. In Sect. 25.3, we derived the expression for SEP for multibranch multihop DF relaying network. In Sect. 25.4, computer simulated results and numerical results are compared to verify the accuracy of the analysis and in the Sect. 25.5, conclusion.

## 25.2 Multihop DF Relay Model with Single Branch

### 25.2.1 System Model

Single branch multihop model consists of a source node (S), multiple intermediate relaying node (R) with DF and a destination node (D) as shown in Fig. 25.1. The information is transmitted through  $(N - 1)$  intermediate nodes  $R_k$ , Where  $R_0$  is source and  $R_N$  is the destination [7].

In the above network, assume all relay nodes are identically independent and experience by Flat Rayleigh fading. Two operating phases is suggested in the model. In the first phase, source node (S) broadcast the MPSK symbol to the relay node. The transmitting symbol  $S_k$  having energy of  $2E_s$  and belong to the one of the M complex constellation  $S_k = \{S_1, S_2, \dots S_N\}$ , which is given by

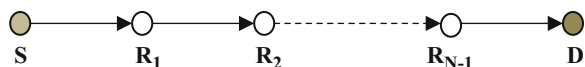
$$S_k = \sqrt{2E_s} \exp\left(j \frac{2\pi (n - 1)}{N}\right), n = 1, 2, 3 \dots N \tag{25.1}$$

where  $j = \sqrt{-1}$ . The received signal with complex baseband in the intermediate relay node in phase 1 is given by

$$r_{SR_k} = h_{SR_k} S_k + n_{SR_k} \tag{25.2}$$

The intermediate relay nodes regenerate the received symbol as  $\hat{S}_k$  and forward the symbol to the destination node [8]. In the destination node, the received baseband complex signal during phase 2 is given by

**Fig. 25.1** Single branch multihop relay model





$$r_{R_k D} = h_{R_k D} \hat{S}_k + n_{R_k D} \tag{25.3}$$

where  $k = 1, 2, 3, \dots, N$  represent the intermediate nodes between the source and the destination. The random complex fading gains of the S-R and R-D links are given by  $h_{SR_k}$  and  $h_{R_k D}$ . Further,  $n_{SR_k}$  and  $n_{R_k D}$  denote the complex circularly symmetric Gaussian noise with zero mean and variance  $\Omega_{SR_k}$  and  $\Omega_{R_k D}$ . The noise is modeled as complex Gaussian variable with variance of  $2N_0$  that is  $\mathcal{CN}(0, 2N_0)$  and zero mean.

The instantaneous SNR of the links  $SR_k$  and  $R_k D$  is given by,

$$\gamma_{SR_k} = \frac{E_s}{N_0} |h_{SR_k}|^2, \quad \gamma_{R_k D} = \frac{E_s}{N_0} |h_{R_k D}|^2 \tag{25.4}$$

and the average SNR of corresponding above links are given by,

$$\Gamma_{SR_k} = E[\gamma_{SR_k}] = \frac{E_s \Omega_{SR_k}}{N_0}, \quad \Gamma_{R_k D} = E[\gamma_{R_k D}] = \frac{E_s \Omega_{R_k D}}{N_0} \tag{25.5}$$

$E[.]$  denote the expectation operator and the detected symbol in intermediate node is given by

$$\hat{s} = \arg \left\{ \max_{s \in S} \text{Re} (s^* h_{SR}^* r_{SR}) \right\} \tag{25.6}$$

where  $(.)_*$  denotes the complex conjugate.

### 25.2.2 SEP of Single Branch Multihop Network

Based on the random variable  $\gamma$ , the conditional error probability of MPSK modulation scheme is given by

$$P_e(\gamma) = \frac{1}{\pi} \int_0^{\frac{\pi(M-1)}{M}} \exp \left( -\frac{\gamma \sin^2 \left( \frac{\pi}{M} \right)}{\sin^2 \phi} \right) d\phi \tag{25.7}$$

Since  $\gamma$  is an exponential distributed random variable, its PDF and CDF are given by eqs. (25.8) and (25.9)

$$f(x) = \left( \frac{1}{\Gamma_{XY}} \right) \exp \left( -\frac{x}{\Gamma_{XY}} \right) \tag{25.8}$$

$$F(x) = 1 - \exp\left(-\frac{x}{\Gamma_{XY}}\right) \tag{25.9}$$

$\Gamma$  denote the average SNR and a another random variable  $V_i$

$$V_i = \min(\gamma R_{1,2}, \gamma R_{2,3}, \gamma R_{3,4}, \dots, \gamma R_{N-1,D}) \tag{25.10}$$

The overall CDF ( $F_{V_i}$ ) of the SD link with above random variable is given by,

$$F_{V_i}(u) = \left(1 - \exp\left(-\frac{u}{\Gamma_{XY}}\right)\right)^N \tag{25.11}$$

The average SEP of the S-D links is derived by the averaging eq. (25.7) over the exponential statistics of the links under the condition that  $\gamma SD > \min(\gamma R_{1,2}, \gamma R_{2,3}, \gamma R_{3,4}, \dots, \gamma R_{N-1,D})$  and is given by

$$P_{eD} = \int_0^\infty P_e(x) F_{U_i}(x) f_{SD}(x) dx \tag{25.12}$$

Substituting eqs. (25.7), (25.8) and (25.11) in eq. (25.12) gives

$$P_{eD} = \frac{1}{\pi} \int_0^\infty \int_0^{\frac{\pi(M-1)}{M}} \exp\left(-\frac{x \sin^2\left(\frac{\pi}{M}\right)}{\sin^2 \phi}\right) X\left(1 - \exp\left(-\frac{u}{\Gamma_{XY}}\right)\right)^N \cdot \left(\frac{1}{\Gamma_{XY}}\right) \exp\left(-\frac{x}{\Gamma_{XY}}\right) d\phi dx \tag{25.13}$$

where  $\Gamma_{XY} = \Gamma_{R_{n,n+1}}$

By integrating eq. (25.13) with respect to  $x$ , we get

$$P_{eD} = \sum_{n=0}^{N-1} \binom{N-1}{n} \frac{(-1)^n}{\pi \Gamma_{R_{n,n+1}}} \int_0^{\frac{\pi(M-1)}{M}} \left(\frac{1}{\frac{\sin^2\left(\frac{\pi}{M}\right)}{\sin^2 \phi} + \frac{2}{\Gamma_{R_{n,n+1}}}}\right) d\phi \tag{25.14}$$

## 25.3 Multihop Multiple Branch DF Relaying Model

### 25.3.1 System Model

The system model of multihop multilink relaying cooperative network is shown in Fig. 25.2, which consists of  $K$  branches and  $N_k$  hops in the  $K$ th branch. The Source (S) communicates the destination through the  $K$  branches, where the branch nodes are considered to have the properties of the previous model. First the source broadcast the symbol to the first relay node in each branch. Hereby, each branch carries the symbol information independent to other branches. Intermediate relay nodes decode and forward the received symbol to the neighboring node by partial CSI. If the relay node  $R_{k,N_{k-1}}$  receives the correct symbol, this node forward the symbol to the final node otherwise discarded. In destination node, the received symbols from each branch are combined using Maximum ratio combing technique [9].

Consider the symmetry of the MPSK constellation, the conditional probability error of any symbol  $S_k$  transmitted, conditioned on the  $\gamma_{S,R}, \gamma_{R,R_K}, \gamma_{R_K D}$ . Which we denote as  $Pe(\gamma_{S,R}), Pe(\gamma_{R,R_K}),$  or  $Pe(\gamma_{R_K D})$ , respectively.

From the tree diagram shown in Fig. 25.3, which gives the possible outcomes of events when source transmits symbol  $S_k$ , and  $S_M$  is the error symbol decoded by the node [10]. From Fig. 25.3 it is clear that the conditional probability of the decision during forwarding, conditioned on  $\gamma_{R,R_K}$  and  $\gamma_{R_K D}$ , is given by

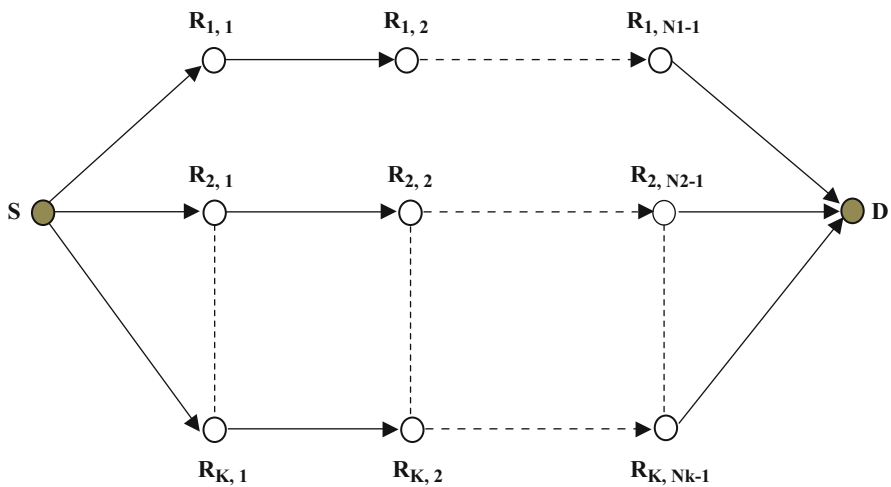


Fig. 25.2 Multihop multiple branch network model

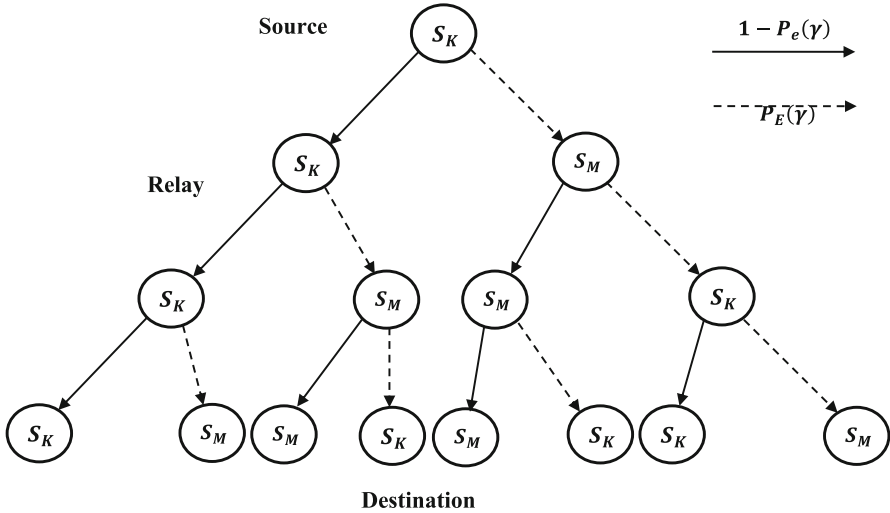


Fig. 25.3 Tree diagram depicting the outcomes of events when user 1 transmits symbol  $S_K$

$$P_{S_K} = \sum_{n=0}^N (P_e(\gamma))^n (1 - P_e(\gamma))^{N-n} \tag{25.15}$$

where  $N$  = number of Hops and  $n = 0, 2, 4, 6, 8 \dots$  even

### 25.3.2 SEP of Multihop Multibranch

Considering the branches that have decoded the correct symbol at the final relay node, then eq. (25.15) is summated overall possible active branches [11]. Considering the  $K$ th branch to be active and decoded the symbol correctly to the destination. The error probability of such  $K$ th branch is given by,

$$P_{e,S_K} = \sum_{k=0}^K \sum_{n=0}^N (P_e(\gamma))^n (1 - P_e(\gamma))^{N-n} \tag{25.16}$$

Therefore, the SEP of the multihop multilink network at destination is analysis by taking the averaging eq. (25.7) to the over exponential statistics of average error probability of the SRD links.

$$P_{e,D} = \frac{1}{\pi} \int_0^\infty \int_0^{\frac{\pi(M-1)}{M}} \exp\left(-\frac{x \sin^2\left(\frac{\pi}{M}\right)}{\sin^2 \phi}\right) \cdot \left(\frac{1}{\Gamma_{XY}}\right) \exp\left(-\frac{x}{\Gamma_{XY}}\right) d\phi dx$$

$$x \sum_{k=0}^K \sum_{n=0}^N \exp\left(\frac{-x}{\Gamma_{XY}}\right)^n \left(1 - \exp\left(\frac{-x}{\Gamma_{XY}}\right)\right)^{N-n} \tag{25.17}$$

Integrate eq. (25.17) with respect to  $x$ , we get

$$P_{e,D} = \sum_{k=0}^K \sum_{n=0}^N \binom{N-n}{n} \frac{(-1)^n}{\Gamma_{R_{n,n+1}}} \int_0^{\frac{\pi(M-1)}{M}} \left(\frac{1}{\frac{\sin^2\left(\frac{\pi}{M}\right)}{\sin^2 \phi} + \frac{N+n+1}{\Gamma_{R_{n,n+1}}}}\right) d\phi \tag{25.18}$$

### 25.4 Evaluation and Simulation Result

In this section, the performance of the multihop multilink DF relay network is analyzed by graph. The network model represented is considered to be identically and independent distributed (i.e.,  $\Gamma_{SR} = \Gamma_{RR} = \Gamma_{RD} = \Gamma_{XY}$ ). For simplicity of exposition, a MPSK modulation used with separate orthogonal channel. In the analysis of both models, the energy is considered to be equal to all the nodes. For the path loss, the model uses the average power of the channel coefficient between the source and destination by effective distance [12–15].

Figure 25.4 compares the theoretical SEP and the simulated result for a single branch multihop relaying network model. The relay distance between the relays is given by  $d_{R_k R_{k+1}} = 1/N$  and the energy of symbol is given by  $E_K = E_S/N_0$ . Performance is better by adding more intermediate relay nodes between source and destination. The diversity order also remains same for the increase in hops and also observed the same level of curve for high SNR.

The performance comparison graph of the model for MPSK modulation for a two-hop network is shown in Fig. 25.5. As we anticipate that the plot of BPSK gives much better performance than the QPSK, 8-PSK and 16-PSK.

The destination node receives the multiple copies of symbol and increase the diversity order of multibranch relaying. Figure 25.6, shows the results of SEP vs. average SNR of multihop multibranch relaying for  $B = (1, 2, 3, 4)$  with MPSK using partial CSI. The result shows that diversity order increase with increase of branches. The performance is improving by increasing the intermediate relay nodes and branches. For example, for a SEP of  $10^{-3}$  the SNR gain is about 2 dB with  $N = 3$ . The performance gives a saturation in the gain for value of  $N = 9$ .

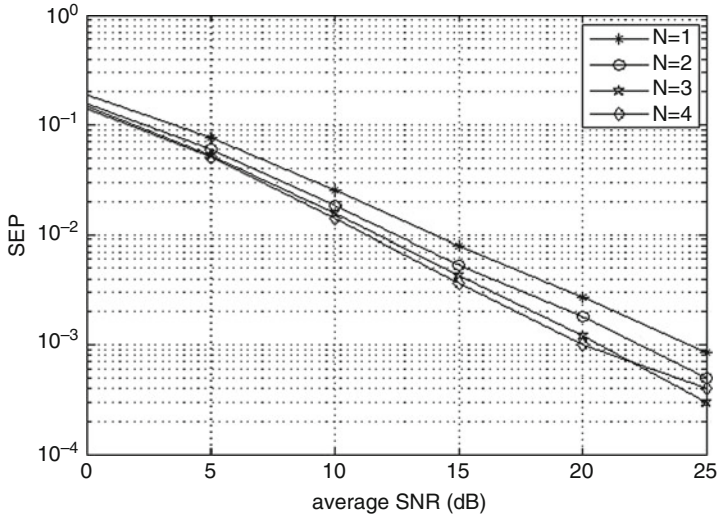


Fig. 25.4 Computation and simulation results of multihop ( $N = 1, 2, 3, 4$ ) single branch relaying

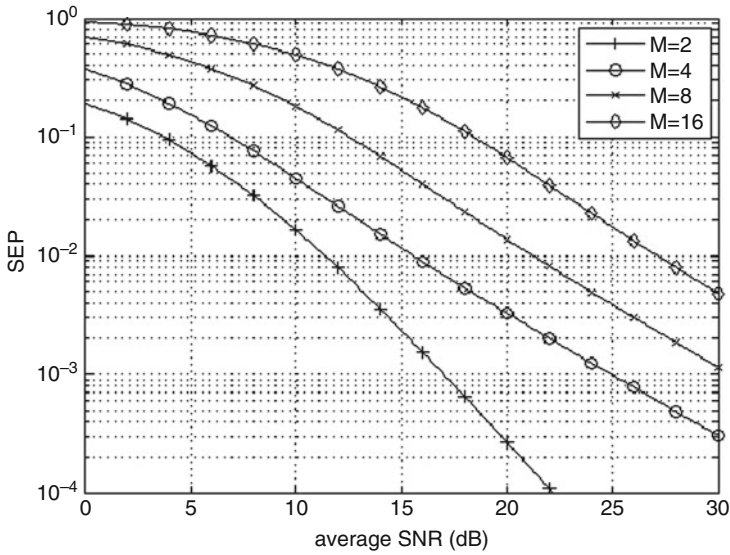


Fig. 25.5 MPSK modulation performance comparison ( $M = 2, 4, 8, 16$ )

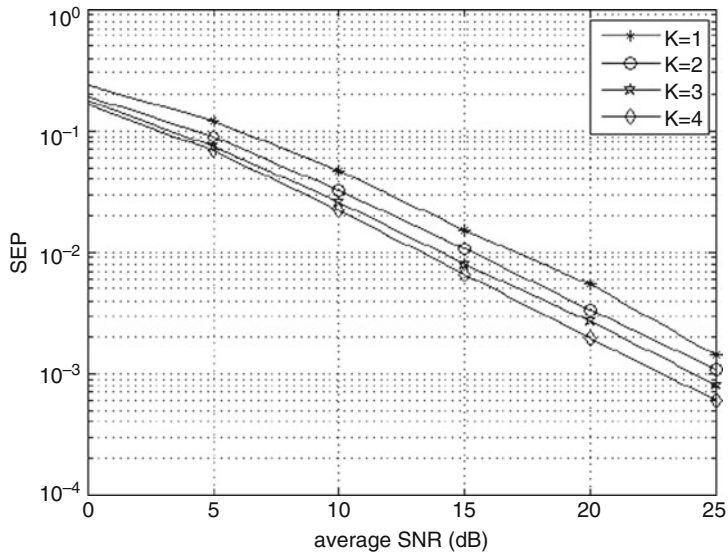


Fig. 25.6 SEP of multihop ( $N = 1, 2, 3, 4$ ) multibranch ( $K = 1, 2, 3, 4$ ) relaying

## 25.5 Conclusion

In this chapter, two models of cooperative relay network are considered, single branch with multi relay nodes and multibranch multihop relaying with partial CSI. The symbol error probability of the networks is derived with MPSK and flat Rayleigh fading channel. The results inferred that number of branches equals the diversity order. The simulation and numerical results are validated with our analytical expressions.

## References

1. A. Nosratinia, E.H. Todd, A. Hedayat, Cooperative communication in wireless networks. *Commun. Mag.* **42**(10), 74–80 (2004)
2. M.D. Selvaraj, R.K. Mallik, Error analysis of the decode and forward protocol with selection combining. *IEEE Trans. Wireless Commun.* **8**(6), 3086–3094 (2006)
3. J.N. Laneman, D.N.C. Tse, G.W. Wornell, Cooperative diversity in wireless networks: efficient protocols and outage behavior. *IEEE Trans. Information Theory* **50**(12), 3062–3080 (2004)
4. M.O. Hasnaand, S. Alouini, Outage probability of multihop transmission over naka-gami fading channels. *IEEE Commun. Lett.* **7**(5), 216–218 (2003)
5. A. Paul, M. Kaveh, Exact symbol error probability of a cooperative network in a Rayleigh-fading environment. *IEEE Trans. Wireless Commun.* **3**, 1416–1421 (2004)
6. H. Jeremiah, N.C. Beaulieu, Performance analysis of decode-and-forward relaying with selection combining. *Commun. Lett.* **11**(6), 489–491 (2007)

7. S. Amara, H. Boujemaa, Multihop multibranch DF relaying for cooperative systems. *IEEE Trans. Wireless Commun.* **9**(3), 144–148 (2011)
8. J. Boyer, D.D. Falconer, H. Yanikomeroglu, Multi-hop diversity in wireless relaying channels. *IEEE Trans. Commun.* **52**(10), 1820–1830 (2004)
9. M.D.Selvaraj, Ranjan K. Mallik, Full CSI selection combining for multi-relay Cooperative diversity systems, in *Communications (NCC), 2012 National Conference on IEEE* (2012)
10. M. Dayanidhy, V.J.S. Kumar, Performance investigation of multi-relay cooperative diversity networks. *Comput. Electrical Eng.* **60**, 151–160 (2017)
11. M. Dayanidhy, V.J.S. Kumar, SEP of the min-max selection combining over Rayleigh fading channel, in *Wireless Communications, 2016 International National Conference on. IEEE* (2016)
12. T.Q. Duong, V.N.Q. Bao, H.-J. Zepernick, On the performance of selection decode-and-forward relay networks over Nakagami-fading channels. *IEEE Commun. Lett.* **13**(3), 172–174 (2009)
13. J. Yindi, H. Jafarkhani, Single and multiple relay selection schemes and their achievable diversity orders. *Wireless Commun.* **8**(3), 1414–1423 (2009)
14. R. Alejandro, X. Cai, G.B. Giannakis, Symbol error probabilities for general cooperative links. *Wireless Commun.* **4**(3), 1264–1273 (2005)
15. S.-I. Chu, Performance of amplify-and-forward cooperative diversity networks with generalized selection combining over Nakagami-m fading channels. *Commun. Lett.* **16**(5), 634–637 (2012)



# Chapter 26

## Predicting Property Prices: A Universal Model



E. Poovammal, Mayank Kumar Nagda, and K. Annapoorani

### 26.1 Introduction

Property rates of any region are always a topic of discussion within the society. Factors such as supply and demand, demographics, and location affect property rates of a region. Also, these factors might not be the same for all the locations. In the real-world situation there are negotiations while purchasing a property. Hence, there can only be a range of actual property rates and not the exact rate itself.

This chapter not only suggests a new approach to predict property prices in a particular region but also proposes a model which can further be used to predict property prices globally.

As of now other researchers [2–4, 7] have concentrated on a single region for property prediction models due to constraints such as variable property influential factors and data [1]. This chapter is a step towards developing a flexible model which can be used in all regions universally to predict property prices. This chapter uses Improved Linear Regression model to predict regional property prices.

Clustering is used to universally represent regions which are having same dependency factors as different clusters of property based on the predicted prices.

---

E. Poovammal (✉) · M. K. Nagda · K. Annapoorani  
Department of Computer Science and Engineering, SRM Institute of Science and Technology,  
Kattankulathur, Tamil Nadu, India  
e-mail: [poovammal.e@ktr.srmuniv.ac.in](mailto:poovammal.e@ktr.srmuniv.ac.in)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_26](https://doi.org/10.1007/978-3-030-19562-5_26)

259

## 26.2 Related Work

The quest for predicting property rates has been studied for a long time in many fields. This includes statistics, patterns recognition and exploratory data analysis. Analysing the explicitly stored data and going beyond the data provides better knowledge about a business.

Most of the prediction models [2–4] developed so far have a common way of collecting property rates in a given region. A suitable mathematical model is then used to generalize the collected data for that region which helps in property prediction. The prediction models entirely depend on the mathematics used and also they hold true for a particular region where rates have fewer variations. These models can't be used in regions with large variations in property rates.

A universally generalized model is one which can be used as such in any region and is flexible enough to adapt itself according to that region. While doing so, model should also have competitive mathematical success.

### 26.2.1 Limitations

The real problem of property prediction lies in the market itself. Property rates keep on changing and there are too many factors affecting it [1]. In some areas, it may be a new supermarket which can influence property rates. In other example facilities yet to come could drive up property rates. It becomes very tough to predict property rates with a model which is one to two years old as they don't count in the new facilities and their effect on the property market. Hence, with these many factors and problems in the way, most of the researchers limit their model to a particular location, e.g., London, Boston, and Montreal, due to the datasets available. But what about other places? This can be achieved using a generalized dynamic flexible model.

## 26.3 Proposed System

Our aim is to develop a universal model which can be used in all the regions and is not just limited to a particular region. To generalize any model universally, some factors need to be considered which affect the property rates of that region. These factors may not be the same for all the regions. A generalized model should be flexible enough to change itself with any particular region according to its unique factors.

Property rates for any region are never constant. They keep on changing with the market transition [13, 14]. The generalized dynamic model should also be intelligent enough to change itself with the market growth. What is the use of a model which

gives outcomes of the past? A real prediction model is one which can predict the future property rates and trends.

Once the generalization problem is solved, different mathematical techniques such as Improved Linear Regression and Clustering can be applied to predict property rates.

All the factors which affect the property prices are relative from location to location and so is their fundamental meaning. Therefore, it is not possible to have a generalized model which has some fixed values and factors to predict the property rates.

The solution to this problem is to divide different location with varying prices into different clusters. In these clusters, the factors which affect the property prices will have the same meaning and influence. The motto is to prepare a universal model which can predict the property rates in these individual clusters.

The Final Model comes into play after combining all the clusters into one to get one single contour as the solution.

## 26.4 General Trends in Property Rates

Trends in property rates change very fast. A property lying dead somewhere in a corner of the city can suddenly become a hot spot if there is a positive market change nearby, such as an opening of new supermarket or university in that area. These factors drastically affect the property prices.

Property rates in any given area largely depend on the factors such as:

1. Location
2. Supply and demand
3. Market/economic growth
4. Condition
5. Neighbourhood

These factors vary with respect to different areas. Property rates of a good location in a tier 2,3 city differs from that of a tier 1 city. Globally there can't be anything such as a good location or a bad location. It always depends on the neighbourhood.

A good neighbourhood in Texas has a very different meaning than that of in Miami and that influences property rates very much.

### 26.4.1 *Factors Affecting Regional Property Rates*

The rates of a property in a specific region also depend on the same factors such as location, supply/demand, and neighbourhood, as discussed in the beginning of Sect. 26.4. Yet, these factors can be addressed with more specific and accurate values.

It is similar to a situation where instead of knowing the overall total amount (gross price) of our purchase in a retail shop, we prefer to know which item of purchase (specific item and its quantity) contributes to what percentage of our total purchase.

Every region has its own quality or influential factors, which are nothing but the factors most sought after by the people living in that region [12]. The target now is to survey these regions and calculate desirability of the location from the local people. The more desired area, the higher its cost will be.

## 26.5 Predicting Property Prices: A New Approach

In this section, a new approach to predict regional property rates is visualized. This approach is exclusively dependent on a particular region as well as can be further changed with time when the trends shift. This approach is discussed with a sample scenario.

### 26.5.1 Surveying Regional Factors

Since, every region has some desirable locations affecting its property rates, it is best to do a survey. The deciding factors can be a nearby school, college, shopping mall, hospital etc. Target regions are surveyed among the people and the desirable locations are collected as L1, Location 1; L2, Location 2; L3, Location 3; ... Ln Location  $n$  and stored as Survey-1.

The survey data gives us the desirable locations (L1–Ln) where people want their property to be close by. But the preferences differ from person to person, according to their life style. So, to make it more accurate one more survey is conducted to ask people about the priority given by them (in terms of points) to that particular location/place collected as P1, Mean priority points for Location 1(L1); P2, Mean priority points for Location 2(L2); P3, Mean priority points for Location 3(L3); ... P $n$ , Mean priority points for Location  $n$ (Ln) and stored as Survey-2.

Every region has different locations which affect the property prices of that region. Table 26.1 represent those factors/locations and the priority points, chosen by the residents living in that region.

### 26.5.2 Prediction Algorithm

A universal prediction algorithm is formed as shown in Fig. 26.1 which uses the survey data collected in Sect. 26.5.1

**Table 26.1** Priority given to the considered location factors

Locations	Average priority points (out of 10)
Office/college/school	6.2
Transportation facility (bus/rail/air)	5.8
Super market/shopping complex	9.1
Park/garden	8.5
Good restaurants	8.5
Hospital	8.9
Places of religious importance	5

**Input:** Property dataset of a particular region

**Output:** Prediction data/values

**Function 1: Calculating Weight points.**

*Function Starts*

*//To calculate weight points-*

**1:** Calculate distance ( $d1...dn$ ) of input properties from the locations ( $L1... Ln$ )

**2:** Location Points =  $(10 - \text{distance}) * \text{Priority points}(P1..Pn)$  of each Location.

**3:** Weight point= summation of allocation points of a property.

*Function Ends*

**Algorithm Initialization:**

**1:** Select a property ( $O1$ ) in a region for which prediction is to be done.

**2:** Get selling price ( $S1, S2, S3..Sn$ ) of some properties in that region which were sold in last few months.

**3:** For each of the selected properties, calculate their weight points ( $W1, W2, W3..Wn$ ) by calling Function-1.

**4:** Dataset now has weight points of properties ( $W1, W2, W3..Wn$ ) with their selling price ( $S1, S2, S3..Sn$ ).

**5:** Calculate weight point ( $Ow1$ ) of the preferred property ( $O1$ ) by calling Function-1 and use that in a customized improved linear regression model with ( $W1, W2, W3..Wn$ ) and ( $S1, S2, S3..Sn$ ) to calculate  $Os1$  (predicted selling price of the selected property).

**Ends**

**Fig. 26.1** Universal prediction algorithm

### 26.5.3 Improved Linear Regression

Property prices for any region follow a particular continuous trend; that is why linear regression is the most applicable method to use. But sometimes due to some anomalies in the survey data, some of the data points are not correct. These data sets

can have drastic effect on the regression model [5, 6]. To overcome these human anomalies a new improved regression technique should be followed.

**Improved Regression Algorithm.** Property prices depend on the human factor. Sometimes, some might pay too high for a property and some might get it cheaply also based on the negotiations. There is always a possibility of having some inconsistent data in the survey. To take care of such constraint an improved method is used as shown in Fig. 26.2 as algorithm.

Figure 26.3 is generated by data collected from sandiego.edu. After applying Simple Linear Regression on unit prices and property lot size, the accuracy calculated is 87.54%.

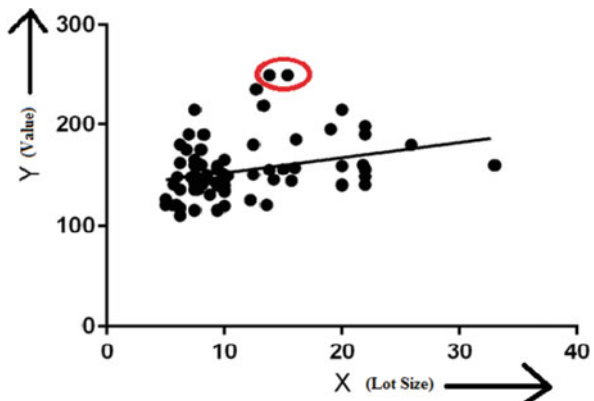
**Input:** Linear Regression Model  
**Output:** Regression Model with improved accuracy  
**Algorithm Initialization**

- 1: Make Linear Regression Model of data points using the default method and plot it.
- 2: Select 95% (significance) of total data sets using Euclidean Square distance symmetrically on both sides of the line.
- 3: Draw 2 pseudo parallel lines on both side of regression line having equal distances so that the area under pseudo lines should cover 95% of the data.
- 4: Select a data point which is farthest from the regression line (perpendicular Euclidean distance) and does not lie in the 95% of the data.
- 5: Remove the data point selected in Step 4 and perform Linear Regression again with the edited data.
- 6: If there is increased accuracy then repeat Step4-6
- 7: If accuracy decreases, retrieve the last deleted data.
- 8: The new regression line is now obtained from the remaining data sets for prediction.

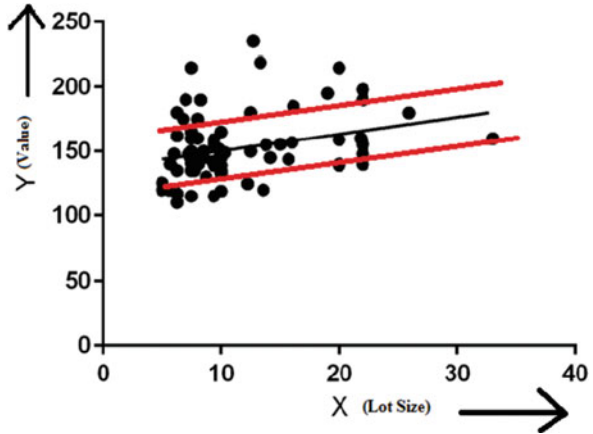
**Ends**

Fig. 26.2 Algorithm for improved regression

Fig. 26.3 Linear regression on property prices



**Fig. 26.4** Improved regression model with extended pseudolines



**Table 26.2** Calculating weight points of a given property

Locations	Priority points ( $p$ )	Distance of $S1$ from locations ( $s$ )	$(10 - s)*p$
Office/college/school	6.2	0.2	60.76
Transportation facility (bus/rail/air)	5.8	0.3	56.26
Super market/shopping complex	9.1	0.1	90.09
Park/garden	8.5	0.3	82.45
Good restaurants	8.5	0.2	83.30
Hospital	8.9	0.1	88.11
Places of religious importance	5	0.6	47.00
$\Sigma$ (Total)			507.97

Applying Improved Regression discussed in the beginning of Sect. 26.5.3, the two marked data points shown in Fig. 26.3 are removed after two iterations of the Improved Regression method.

Figure 26.4 is generated after two iterations of the algorithm discussed in Sect. 26.5.3. After two iterations of improved algorithm the accuracy calculated is: 89.16%.

### 26.5.4 Calculation of Weight Points of Property

Assuming known selling price of  $S1 = \text{Rs. } 1000/\text{sq-ft}$  and  $S2 = \text{Rs. } 2000/\text{sq-ft}$ , the selling price of new property  $S3$  is to be calculated.

- $d1 =$  distance of 1st Location (L1) from  $S1$
- $10 - d1 =$  individual point of each location

Weight points for  $S1$  and  $S2$  are calculated as shown in Table 26.2 using the Universal Prediction Algorithm.

Similarly, weight points of properties (W1, W2, W3) are calculated. Selling Prices (S1, S2) are already known. Applying Linear Regression technique, S3 can be calculated using steps 1–3.

Step-1 Put W1, W2 are on x-axis.

Step-2 S1, S2 on y-axis.

Step-3 Then for equation  $y = a + bx$

$$a = \left( \sum y \right) \left( \sum x^2 \right) - \left( \sum x \right) \left( \sum xy \right) / \left( n \left( \sum x^2 \right) - \left( \sum x \right)^2 \right) \quad (26.1)$$

$$b = n \left( \sum xy \right) - \left( \sum x \right) \left( \sum y \right) / \left( n \left( \sum x^2 \right) - \left( \sum x \right)^2 \right) \quad (26.2)$$

Now, in the equation where  $a, b$  are known values;  $W3$  can be substituted as  $x$ , and hence  $y$  can be calculated. Value of  $y$  is our predicted value  $S3$ .

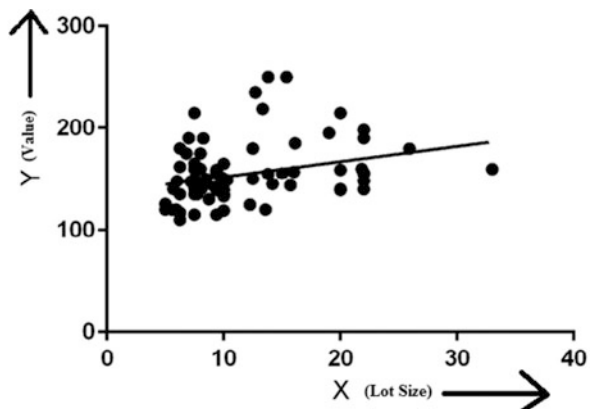
All the Weight Points with the Selling Prices are stored in the database as shown in Table 26.3 So that it can be used for further prediction in that region.

Figure 26.5 shows Improved Linear Regression graph plotted by taking Weight Points on  $x$ -axis and Property Prices on  $y$ -axis. It can be seen that for a particular region there is not much deviation in the property prices, hence linear regression is a suitable option to use in this scenario.

**Table 26.3** Database of weight points and selling price

Property ID	Weight points	Selling price/predicted price
1	W1	S1
2	W2	S2
3	W3	S3

**Fig. 26.5** Linear regression graph





### 26.5.5 Clustering Model

Regional property rates can be calculated using the model discussed in Sect. 26.5 by surveying the factors which affect the property prices in a particular region with their priority points. By using these priority points, we calculate weight points of each property in a particular region. The known selling price and weight points are then used in Improved Linear Regression algorithm for predicting the regional rates of a property. Using those regional rates, we can calculate our Clustering Model.  $k$ -means clustering is used to cluster the regions. The clustering will group the data together with properties having same dependency factors in a particular region.

## 26.6 Analysis of Result

The Algorithm and technique proposed in Sect. 26.5 can be used universally in any region to calculate property rates.

If there are any changes in the desirability location, such as any new shopping mall opens or anything else. We can simply add one more location attribute and can carry on with the same model. Hence this model is flexible enough to adopt with future changes, if any, in a region.

Also, an Improvised Linear Regression Model is discussed to remove the anomalies caused in the process of survey or human error. In machine learning, improving the error by 0.01 with a particular algorithm can be considered a significant breakthrough [8–10]. However, it is arguable whether these improvements will translate into any useful applications in everyday life [11]. With the Clustering Model, property trends can be visualized easily from the clusters.

## References

1. R.J. Shiller, *Understanding recent trends in house prices and home ownership*. National Bureau of Economic Research, Working Paper No. 13553 (2007). doi: <https://doi.org/10.3386/w13553>.
2. N. Pow, E. Janulewicz, L.D. Liu, *Applied Machine Learning Project for Prediction of real estate property prices in Montreal*. SJSU ScholarWorks (2017, Spring)
3. N. Bhagat, A. Mohokar, S. Mane, House price forecasting using data mining. *Int. J. Comput. Appl.* **152**(2), 975–8887 (2016)
4. A. Ng, *Machine learning for a London housing price prediction mobile application*. Thesis, Imperial College of London (2015)
5. D. Belsley, E. Kuh, R. Welsch, *Regression diagnostics: Identifying influential data and source of collinearity* (Wiley, New York, 1980)
6. J.R. Quinlan, *Combining instance-based and model based learning* (Morgan Kaufmann, San Mateo, CA, 1993)
7. A. Caplin, S. Chopra, J.V. Leahy, Y. LeCun, T. Thampy, Machine learning and the spatial structure of house prices and housing returns. Available at SSRN 1316046 (2008)

8. Y. Bengio, X. Glorot, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of AISTATS 2010*. Sardinia, Italy: Chia Laguna Resort (2010, May), (Vol. 9, pp. 249–256)
9. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
10. J. Schmidhuber, Multi-column deep neural networks for image classification, in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. CVPR '12* (Washington, DC: IEEE Computer Society, 2012, pp. 3642–3649. . ISBN: 978-1-4673-1226-4
11. K. Wagstaff, Machine learning that matters. CoRR., [abs/1206.4656](https://arxiv.org/abs/1206.4656) (2012)
12. I.R. Lake, A.A. Lovett, I.J. Bateman, I.H. Langford, Modelling environmental influences on property prices in an urban environment. *Comput. Environ. Urban Syst.* **22**(2), 121–136 (1998)
13. K. Tsatsaronis, H. Zhu, What drives housing price dynamics: Cross-country evidence (March 1, 2004). *BIS Quarterly Review* (2004, March 1). Retrieved from <https://ssrn.com/abstract=1968425>
14. T. San Ong, Factors affecting the price of housing in Malaysia. *J. Emerg. Issu. Econ. Finan. Bank.* **1**(5), 414–429 (2013)

# Chapter 27

## Facial Based Human Age Estimation Using Deep Belief Network



Anjali A. Shejul, Kishor S. Kinage, and B. Eswara Reddy

### 27.1 Introduction

Due to lot of real-time applications in surveillance, electronic vending machines, security control, forensic art, entertainment, cosmetology, and many more, facial based human age estimation has obtained huge popularity in research domain. This has become interesting and attractive topic. Human face reflects large amount of information such as gender, expression, and age. This paper focuses on predicting age information from face images. As age progresses human face exhibits remarkable facial changes. These changes are mainly happens in two stages, one is from birth to adulthood and other is from adulthood to old age. In the first stage changes mainly occur in craniofacial growth that is in shape of face and skull. In the second stage facial changes occur in skin texture, i.e., changes in skin and muscle elasticity [1]. Facial age estimation can be considered as multiclass classification or regression approach. For multiclass classification multiple classes of age ranges, i.e., continuous values, need to be considered, whereas in regression approach each image is assigned to exact discrete age value.

---

A. A. Shejul (✉)

Department of CSE, JNTUA, Ananthapuramu, Andhra Pradesh, India

K. S. Kinage

Department of E&TC, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India

B. E. Reddy

Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_27](https://doi.org/10.1007/978-3-030-19562-5_27)

269

### **27.1.1 Organization of the Paper**

The remaining part of the paper is organized as follows.

Section 27.2 describes the related work. Section 27.3 presents in detail methodology of the work. This section elaborates different steps used in age estimation. Section 27.4 briefs facial aging database used in this paper. Performance measures used by researchers are presented in Sect. 27.5. Experimentation carried out and corresponding results are mentioned in Sect. 27.6. Finally conclusion and future scope is discussed.

## **27.2 Related Work**

In literature most authors considered age estimation as a multi class classification approach [2–5] or a regression approach [6–10]. Kuang-Yu Chang et al. [11] introduced a cost sensitive ordinal hyper planes ranking algorithm for facial based human age estimation. Author applied relative order information among the age labels for rank prediction. Yuan Dong et al. [12] present age estimation based on deep learning algorithm. In this author trained deep convolution neural network (DeepConvNets) using transfer learning strategy. New loss function is introduced for age classification task. Shengzheng Wang et al. [13] proposed that asymmetric information can be used to improve the generalizability of the trained model. Author introduced learning using privileged information (LUPI) framework, for this attributes of support vector machine (SVM+) are carefully defined. Author termed specific setting as relative attribute SVM+ (raSVM+). Sun et al. proposed DeepID structure to extract features [14]. DeepID crops 60 patches from face image and each patch is fed to independent network [15].

## **27.3 Methodology**

Facial based human age estimation predicts age for the given input image and also calculates mean absolute error (MAE). Figure 27.1 shows steps of age estimation process. All steps are divided in training phase and testing phase. During training phase images from input dataset are preprocessed. To reduce dimensionality of images vital features are extracted using active appearance model and scattering transform. These features are fed to build deep belief network classification model by training partitioned images. In the testing phase, model is tested on partitioned testing images to find age. MAE is also calculated to check error rate. Each step is elaborated in below subsection.

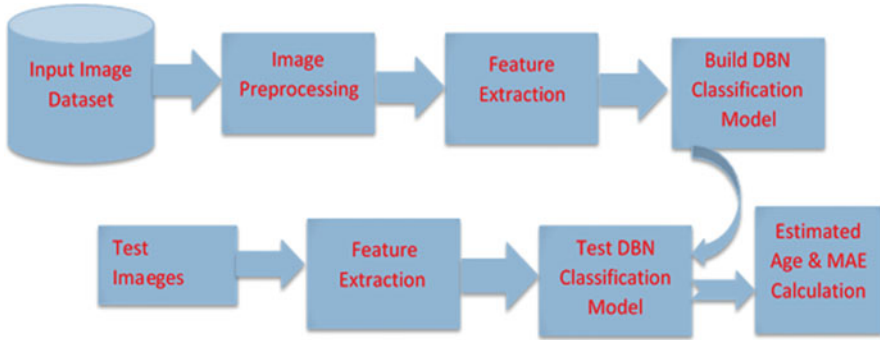


Fig. 27.1 Steps for age estimation using DBN

### 27.3.1 Preprocessing

Many times images captured from camera are not suitable for analysis. Such kind of images is preprocessed for noise removal, histogram equalization, and illumination correction. Along with this face detection need to be carried out to focus on only important region. For this Cascade object detector method from computer vision toolbox is used. Cascade object detector uses Viola Jones face detection algorithm to detect face resulting in bounding box around face.

### 27.3.2 Feature Extraction

Extracting informative, non-redundant, and robust facial features are important for better results of facial age estimation. Features can be of local, global, and hybrid type. From birth to adulthood facial changes are mainly in global features and from adulthood to old age facial changes are in local features. Global feature deals with changes in craniofacial growth that is changes in shape of face and skull. Whereas local features deal with changes in skin texture including muscles elasticity. Hybrid features are combination of local and global features. Literature mentioned following facial feature extraction models:

- Anthropometric model
- Active appearance model (AAM)
- Aging pattern subspace
- Age manifold

Anthropometric model considers only facial geometry that is shape and ratio of face. Because of this anthropometric is applicable for younger ages only. AAM considers both geometry and texture so applicable for all ages. Aging pattern subspace is personalized age estimation model. This uses sequence of personal

images sorted in time order. Age manifold learns common pattern of aging for more than one person at different ages. In this paper AAM and scattering transform feature extraction technique is used.

### 27.3.2.1 Active Appearance Features

Cootes et al. [16] proposed this model. AAM focuses not only on craniofacial features but also on texture features [17]. AAM is based on principal component analysis (PCA). AAM marks landmark points on facial images. AAM uses aging function defined by

$$\text{age} = f(b) \quad (27.1)$$

where age is the age of a person in the picture,  $b$  is a vector, it includes parameters learned from AAM, and  $f$  is an aging function. The function defines the relationship between person's age and facial description parameters. Different variations of aging functions are: quadratic aging function, linear aging function, cubic aging function, and others.

### 27.3.2.2 Scattering Features

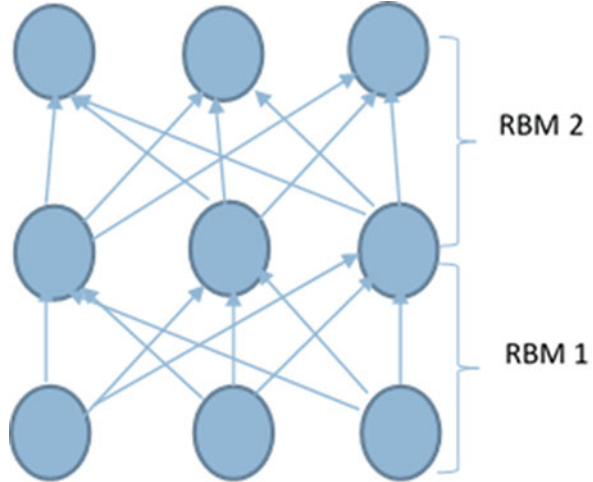
Scattering Transform extracts features based on scattering model discussed in [18]. Scattering transform has multiple layers and is invariant to various transformations such as translation and rotation. Scattering features are extracted using a deep wavelet based architecture named as ScatNet. Every layer produces scattering transform coefficients with different scales and orientations by convoluting averaging filter on the input image. Higher layer produces high frequency information whereas lower gives low frequency information.

### 27.3.3 DBN Classification Model

DBN is a hierarchical model having multiple layers. DBN comprises multiple restricted Boltzmann machine (RBM). Number of RBMs depends upon our requirements. Figure 27.2 shows deep belief network having two RBM Layers. RBM is a two-layer network with input and hidden layer. Each layer of RBM consists of various nodes. Nodes from different layers communicate to each other but nodes from the same layer never communicate. So there is no intralayer communication, only interlayer communication happens. This is the reason RBM is called as restricted.

Features of DBN are

**Fig. 27.2** Deep belief network



- There exists a fast and greedy algorithm that can find good set of parameters.
- Despite being unsupervised algorithm it can be applied to labelled data.
- Reconstructs input using contrastive divergence algorithm to increase performance.
- DBN contributes in solving vanishing gradient descent problem

Vanishing gradient problem is when any model learns by going in the direction of the gradient, many times gradient approaches to nearly zero. So the name is vanishing gradient, due to this vanishing gradient, learning speed reduces yielding local minima. DBN tries to solve this vanishing gradient descent problem [19]. RBMs are used to extract features and to reconstruct input. DBN uses greedy training approach and contrastive divergence method to train RBM. In this work we used two RBMs.

RBM training undergoes below steps:

1. Initialize weights, assume bias and learning rate.
2. Positive phase

Update weights of hidden units ( $H_j$ ) in parallel using conditional probability with eq. (27.2).

$$P(H_j = 1|V) = s\left(B_j + \sum_{i=1}^m W_{ij}V_i\right) \quad (27.2)$$

where  $B_j$  is bias,  $W_{ij}V_i$  is weight associated with hidden unit ( $H_j$ ) and visible unit ( $V_i$ ),  $\sigma$  is activation function.

3. Negative Phase

Update weights of visible units ( $V_i$ ) in parallel using conditional probability with eq. (27.3).

$$P(V_i = 1|H) = s \left( A_i + \sum_{j=1}^n W_{ij} H_j \right) \quad (27.3)$$

where  $A_i$  is bias,  $W_{ij}H_j$  is weight associated with hidden unit ( $H_j$ ) and visible unit ( $V_i$ ),  $s$  is activation function.

4. Update weights of edge ( $W_{ij}$ ) using positive and negative phase with the learning rate

$$W_{ij} = W_{ij} + L^* (\text{Positive}(E_{ij}) - \text{Negative}(E_{ij})) \quad (27.4)$$

where  $L$  is Learning Rate

5. Repeat steps 2–4 until we get some threshold accuracy. Once we have optimal weights, Freeze weights.

## 27.4 Database Description

For the experimentation IMDB face dataset is used. This dataset has a collection of 460,723 images of several celebrities, and 62,328 images from the Wikipedia source. Thus, in total, the IMDB face database has 523,051 face images. The database provides age and gender labels for each image.

## 27.5 Evaluation Metrics

To evaluate performance mean absolute error (MAE) measure is used. MAE identifies the deviation of the output of classifier from actual output to be obtained, and it is expressed as,

$$\text{MAE} = \frac{1}{N} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y}_i)^2} \quad (27.5)$$



where  $Y_i$  and  $\bar{Y}_i$  indicate the actual outcome of the classifier and desired output, respectively.

### 27.6 Experimental Results and Discussion

Experimentations are performed on IMDB database images. Figure 27.3 shows few sample images and corresponding feature extracted images from IMDB database.

Figure 27.4 shows comparative analysis of mean absolute error (MAE) obtained for support vector machine (SVM), neural Network (NN), K-nearest neighborhood (KNN), and deep belief network (DBN) classification algorithm for varying training percentage.

Table 27.1 shows MAE obtained using different classifier for 70% training percentage. This shows that DBN is outperforming.

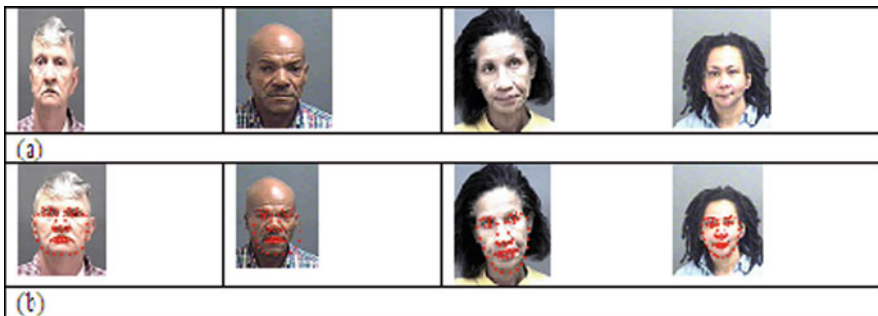
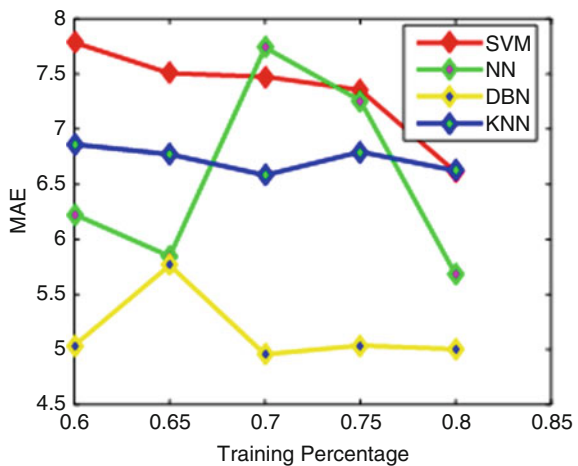


Fig. 27.3 (a) Sample images from IMDB database. (b) Features extracted of sample images

Fig. 27.4 Comparative analysis using the IMDB face database based on MAE



**Table 27.1** MAE obtained for different classifier

Classifier	MAE
SVM	7.4763
NN	7.7415
KNN	6.5818
DBN	4.9553

## 27.7 Conclusion and Future Work

In this paper we proposed human age estimation using deep belief network model (DBN). DBN overcomes vanishing gradient descent problem and so DBN outperforms compared to other classifiers. Features are extracted using active appearance model that extracts not only shape features but texture features also. In future we are planning to work more on feature extraction method to improve accuracy.

## References

1. X. Geng, Y. Fu, K.S. Miles, Automatic facial age estimation. *11th Pacific Rim Int. Conf. Artif. Intell.*, 1–130 (2010)
2. A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **34**(1), 621–628 (2004)
3. Z. Yang, H. Ai, Demographic classification with local binary patterns. *Adv. Biometrics* **4642**, 464–473 (2007)
4. Chung-Chun Wang, Yi-Chueh Su, Chiou-Ting Hsu, Bayesian age estimation on face images, in *2009 IEEE International Conference on Multimedia and Expo*, New York, NY, 2009, pp. 282–285.
5. X. Geng, Z. Zhou, S. Member, K. Smith-miles, Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(200343), 2234–2240 (2007)
6. Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold. *IEEE Transact. Multimedia* **10**(4), 578–584 (2008)
7. B. Ni, Z. Song, S. Yan, Web image mining towards universal age estimator, in *Proc. seventeen ACM Int. Conf. Multimed.—MM '09*, p. 85, 2009
8. G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features. 2009 *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work.* **2009**, 112–119 (2009)
9. G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.* **17**(7), 1178–1188 (2008)
10. Y. Zhang and D.-Y. Yeung, Multi-task warped Gaussian process for personalized age estimation, in *Comput. Vis. Pattern Recognit. (CVPR), 2010 IEEE Conf.*, pp. 2622–2629, 2010.
11. K.-Y. Chang, C.-S. Chen, A learning framework for age rank estimation based on face images with scattering transform. *IEEE Trans. Image Process.* **24**(3), 785–798 (2015)
12. Y. Dong, Y. Liu, S. Lian, Automatic age estimation based on deep learning algorithm. *Neurocomputing* **187**, 4–10 (2016)
13. D. T, J.Y.S. Wang, Relative attribute SVM+ learning for age estimation. *IEEE Trans. Cybern.* **46**, 827–839 (2016)
14. Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10 000 classes, in *CVPR*, pp. 1891–1898, 2014.

15. A.A. Shejul, K.S. Kinage, B.E. Reddy, Comprehensive review on facial based human age estimation. *International conference on Energy, Data Analytics & Soft Computing (ICECDS)*, 3211–3216 (2017)
16. T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models. *Proc. Eur. Conf. Comput. Vis.* **2**, 484–498 (1998)
17. P. Pandey, R. Singh, M. Vatsa, Face recognition using scattering wavelet under Illicit Drug Abuse variations, in *2016 Int. Conf. Biometrics, ICB 2016*, 2016
18. M. Hayes, Adaptive active appearance models, no. December 2005, 2014
19. G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)

# Chapter 28

## Randomized Agent-Based Model for Mobile Customer Retention Behaviour Prediction



N. Sandhya, Philip Samuel, and Mariamma Chacko

### 28.1 Introduction

Mobile technology has become an integral part in the day today business [1]. Customer retention is considered to be the main factor to investigate as it is the centre of interest for developing profitable relationship with customers. It is important for telecommunication providers to retain their customers, as 20–40% mislaying of their customers happen each year [4, 5]. Retention of existing customer is more advantageous to the telecom company since enticing new users to the telecommunication industry is more expensive [6]. Hence, it is important to develop good customer retention techniques.

In the telecommunication industry customer switching from prepaid to postpaid and vice versa is an on-going process [2]. This scenario affects the telecom market in a negative manner [2]. To avoid such situation it is important to find suitable solution to this problem, i.e. by retaining the customers before customer attrition [7, 8]. Different data mining techniques, artificial intelligence methods and other statistical methods are adopted to handle this problem. Usually telecommunication service providers offer different retention policies as trial-and-error methods to entice new customers [11, 15, 16].

---

N. Sandhya (✉)

Information Technology, School of Engineering, Cochin University of Science and Technology, Kochi, India

P. Samuel

Department of Computer Science, Cochin University of Science and Technology, Kochi, India  
e-mail: [philips@cusat.ac.in](mailto:philips@cusat.ac.in)

M. Chacko

Department of ship Technology, Cochin University of Science and Technology, Kochi, India  
e-mail: [mariamamma@cusat.ac.in](mailto:mariamamma@cusat.ac.in)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_28](https://doi.org/10.1007/978-3-030-19562-5_28)

279

Another issue in telecommunication industry is the large volume of data generation which is difficult to analyse and process. This challenge is handled using MapReduce [9] big data technology with machine learning techniques. Excessive volume of unstructured dataset is processed using MapReduce functions in a distributed environment. In order to rectify this customer churn situation to overcome the loss in the telecommunication industry, service providers look forward for different statistical methods and machine learning algorithms [14, 17]. Previous history of the customer details have to be analysed using machine learning methods to make some customer predictions [12, 13].

An object-based model is required for simple and normal computational analysis [19]. Complex computational analysis that is intellectual and autonomous behaviour in nature required an agent-based computational model. Agent-based models are utilized to analyse complex systems [20, 21].

This paper is organized as follows. Section 28.2 discusses related work. In Sect. 28.3 we provide the details of proposed method. Section 28.4 is result analysis. Section 28.5 is the conclusion of the paper.

## 28.2 Related Work

Qureshi et al. [18] describe certain data mining techniques for churn prediction. Customer DNA dataset is used for prediction of 106,000 users traffic data. The customer usage has been analysed for 3 months. The statistical methods implemented are regression analysis method, artificial neural networks, K-means, decision tree including CHAID, Exhaustive CHAID, CART and QUEST. The potential churners have been identified using above methods and Exhaustive CHAID was found to be the most accurate. The prediction accuracy of this method found to be 75.4%.

Proposed paper [23] explains how agent-based model (ABM) is used to express the real-world problems into the computer programming scenario. Real-world entities are demonstrated with the help of algorithms and mathematical formulas. In the computational processing these algorithms and equations are developed using high-level programming languages. This paper also describes how design phase and the testing phase of the computational product are handled by ABM.

The existing system does not handle input data cleaning or preprocessing redundancy techniques. Our proposed systems handle data redundancy problem of the customer call details data set.

## 28.3 Proposed Method

In this paper during data preprocessing phase Randomized Method is applied to clean up the unwanted data from input dataset, i.e. the customer call dataset is

the crucial input data set which is used for customer retention. The data set we obtained is unstructured, and to deal with this large volume of data, we proposed Randomized Method with Map-Reduce technique. Telecommunication customer retention is predicted using agent-based model.

The unstructured user dataset should be cleaned by removing unwanted data to obtain better results in the application execution. All the entities in the dataset are not required for prediction analysis. Some entities are not complete; some of the entities have 'null' values. Incomplete or null value entity fields are removed. So data duplication has to be removed and data integration is required for the better performance of the system. Unstructured redundant input data is given as input. This dataset is cleaned by removing redundant unwanted data using Randomized Method, thus obtaining an irredundant dataset which can be used to train the agent-based customer retention feature prediction model.

### 28.3.1 *Randomized Method with Map-Reduce Technique*

Map-Reduce technology has made large complex data processing more easy and efficient [9]. This technology is applied for batch processing of large volumes of data. Map-Reduce [9] is a parallel data processing approach for execution on computer cluster [10].

As shown in Fig. 28.1 Map-Reduce programming model defines two user defined functions *Map* and *Reduce*. Input data has been divided into chunks. Each chunk generates a *Map* function which performs sorting and filtering of the input data. The input data to *Map* function is in the form of key/value pair [9]. The *Map* function in the Map-Reduce job is fed with data stored on the distributed file system, which are split across nodes. The Map tasks are started on the compute nodes and Map function applied to each key/value pair and outputs intermediate key/value pairs. This intermediate key/value pairs are stored in the local file system and sorted by the keys.

The Reduce function performs the addition operation. The Reduce function takes the intermediate key/value pair as the input, A *Reduce* function is applied to all values grouped for one key and in turn generates key/value pairs. Key/value pairs from each reducer are written on the distributed file.

In the preprocessing stage, the Randomized Method allocates value '0' and '1' for each variable feature of the input dataset. Each feature is compared in the map phase and duplicate variable feature is assigned value '1'. In the reduce phase variable feature with value '0' only counted and feature with value '1' will be eliminated.

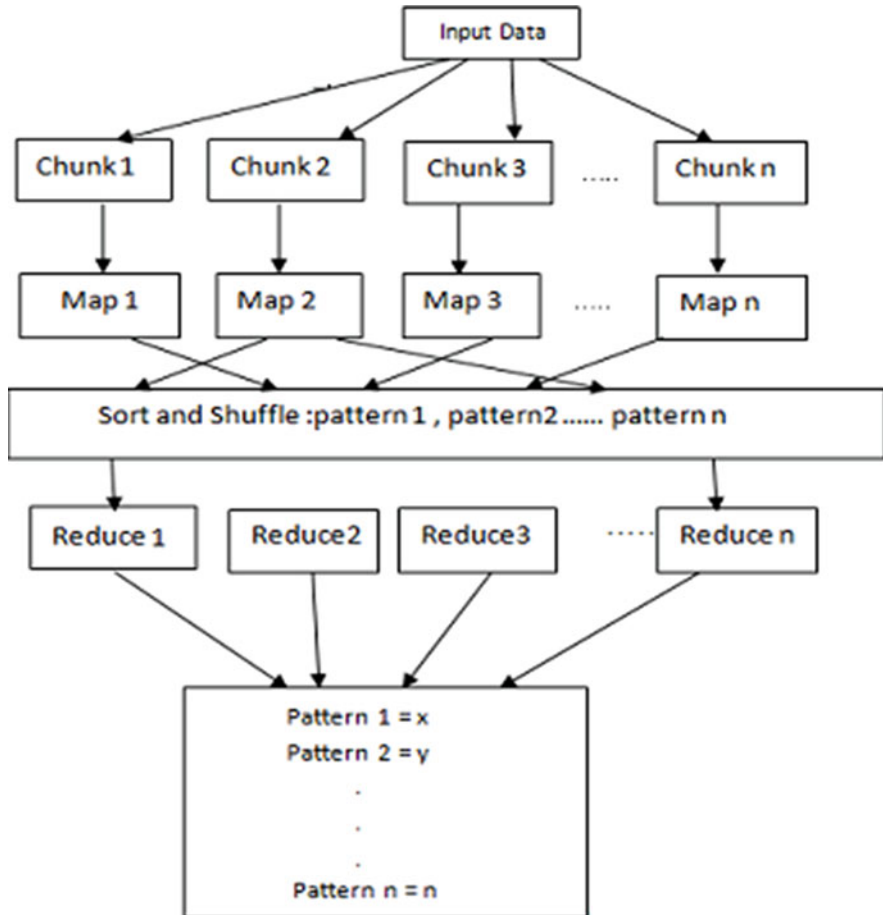


Fig. 28.1 Architecture of MapReduce Model

### 28.3.2 Agent-Based Model

Human behaviour is complex. Multi-agents are useful in understanding different perspectives of human behaviour. Agent-based model (ABM) acts as actions and communications between agents for collective entities such as organizations [22]. The ABM is used to predict some complex phenomenon such as human behaviour which depends on some situations or it affects the system or group as a whole to deal with it [23].

The customer variable features depend on customer behaviour. This customer behaviour is varies at different situations and it can be complex. To predict customer retention is difficult, hence we can implemented an Agent-based model. Using this model customer retention behaviors complexity lessens. Agents are the functional

units of this model. The positive side of this model is that ABM is flexible as increasing the number of agents to the model, the performance of the system can be improved. Addition of the agents to the agent-based model is simple. The variable features are agents in agent-based model. Analysing the customer dataset as shown in Table 28.1, we designed a model to suggest the customer retention variable feature which affects the user churn.

Behaviour of customers affects the mobile user retention; sudden change over to different networks is easier nowadays. Agents are designed as follows churn agents and retention agents. Both agents are again subdivided into some subgroups based on CALL\_DURATION (CD), CALL\_CHARGES (CC), CUSTOMER\_SERVICE\_CALL (CSC), SIM\_EXPIRY (SE) and NETWORK\_PLANS (NP).

Agent interactions occur in dynamic manner. The coding rules of the agents are as follows: Churn Agents are considered as CH\_A, Retention Agents as R\_A. Thus, state space of the agents is defined in Table 28.2.

As shown in Table 28.2, the three state space of the agent interactions have been defined. First state space 'nCoR' in which the state with agent interactions cannot be

**Table 28.1** Variable features of customer call dataset

No.	Variable feature	Description	No.	Variable feature	Description
1	Contract	Plan duration of the customer (PD)	11	Day_Charg	Daytime call charges
2	MO_Churn	Monthly customer churn	12	Eve_Calls	Number of calls in the evening
3	Day_Mins	Duration of call at day (DCD)	13	Eve_Charge	Evening call charges
4	Eve_Mins	Duration of call at evening (ECD)	14	Night_Calls	Number of calls at night
5	Night_Mins	Duration of call at night (NCD)	15	Night_Charge	Call charges at night (CCN)
6	Intl_Mins	Duration of international calls	16	Intl_Calls	Number of international calls
7	CustServ_Calls	Customer service calls (CSC)	17	Intl_Charge	Number of customer service calls (IC)
8	Total_Churn	Customer churn during last 1 year	18	Area_Code	Area code of the customer(NB)
9	Int'l_Plan	Customers taken international plan (IP)	19	Monthly_Chg	Monthly charges of the customer (MCC)
10	Day_Calls	Number of calls at daytime	20	State	Customers state code

**Table 28.2** State space of agents features

(neither CH_A or R_A) as nCoR
(only CH_A) as oC
(only R_A) as oR



defined between the churn agent or retain agent. The state 'oC' is described as the only churn agent. The explicit behavioural feature interactions describe the churn agent. The 'oR' state space is only retention agents with respect to the interaction between the variable features.

These three state spaces are described for this system. The system can change the status at any point of time that makes the system shows complexity depending on customer variable features. The relationships between the features are defined for the proposed model based on the state space features. The rules of the agent-based model prediction for customer retention are defined as follows:

1. If CD\_DCD increase for a user CC\_MCC increases OR CD\_NCD increases then CC\_MCC increases the probability of oC increases.
2. If NP\_PT to NP\_PP or wise versa change over occurs then CSC enquiry should accompany over the customer otherwise oC occurs.
3. If SE then intent to CSC\_PB applicable then chances of nCoR increases.
4. If IC rises then NP\_IP with reduction will increase the chance of oR.
5. If CSC continuously enquired the customer based on CSC\_PB, CSC\_IB and CSC\_NB with customer satisfaction during SE, then high chance of oR.

The proposed prediction agent-based model is implemented using MapReduce function with these defined agent-based rules. The dataset [3] is 1 year customer call details which have been used for training customer retention agent-based model and this system predicts customer retention features best suited for the specific customer based on the variable features.

## 28.4 Result Analysis

Customer retention feature prediction using Randomized Method and MapReduce with agent-based model illustrates the customer retention requirements. Depending on the customer behaviour churn out reasons are characterized as variable features which can be modelled as agents in the system. Table 28.3 shows percentage of customers based on their agent-based behaviour. According to the agent-based rule prediction 55,000, 65,000 and 75,000 customers are analysed based on the ABM. Hence it is seen that customer behaviour depend on the variable features of telecommunication. Customer retention behaviour can be maintained based on the variable feature rate.

**Table 28.3** Customers versus agent-based behaviour

No. of customers	oR behaviour (%)	oC behaviour (%)	nCoR behaviour (%)
55,000	65	25	10
65,000	60	27	13
75,000	56	30	14

## 28.5 Conclusion

Day by day telecommunication market is passing through different transformational development phases. Customer retention in this industry is a challenging situation. Customer satisfaction is an important criterion for mobile customer retention. Our proposed model helps in the customer retention feature using big data and machine learning technique. Data redundancy of large volume of input dataset is removed by Randomized Method in a simple manner using Map function and Reduce function.

Customer retention depends on complex human behaviour. This can be analysed based on the customer call details and predict customer retention using simple global rules in our model. If macroscopic behaviour of the agents in the system is observed then the performance of the system can be improved. Using agent-based method we assimilate intelligence in agents using some rules to make them competent of learning from past experience. The benefit of this work is to develop a simple, flexible and cost-effective model to predict telecommunication customer retention.

## References

1. A.M. Almana, M.S. Aksoy, R. Alzahrani, A survey on data mining techniques in customer churn analysis for telecom industry. *J. Eng. Res. Appl.* **4**(5), 165–171 (2014)
2. W.-H. Au, K.C.C. Chan, X. Yao, A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Trans. Evolutionary Comput.* **7**(6), 532–545 (2003)
3. Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, Telco Churn Prediction with Big Data, in *The 2015 ACM SIGMOD International Conference* (2015, May), <https://doi.org/10.1145/2723372.2742794>
4. S. Ahmed, Z. Kobti, R.D. Kent, Predictive data mining driven architecture to guide car seat model parameter initialization, in *Intelligent Decision Technologies* (Springer, 2011), pp.789–797
5. C. Apte, S.J. Hong, Predicting equity returns from securities data with minimal rule generation, in *Advances in Knowledge Discovery and Data Mining*, ed. by P. S. U. Fayyad, G. Piatetsky-Shapiro, R. Uthurusamy, (American Association for Artificial Intelligence, Menlo Park, CA, 1996), pp. 541–560
6. J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction. *Expert Syst. Appl.* **36**(3), 4626–4636 (2009)
7. K. Coussement, D. Van den Poel, Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. *Expert Syst. Appl.* **34**(1), 313–327 (2008)
8. K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, A. Joshi, Social ties and their relevance to churn in mobile telecom networks, in *EDBT*, pp. 668–677, (2008)
9. F. Li, B. C. Ooi, M. Tamer Ozsu, S. Wu, Distributed data management using MapReduce, *ACM Comput. Surv. (CSUR)*, vol. 46, Issue 3, pp. 42, Jan. (2014).
10. C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, K. Olukotun, MapReduce for machine learning on multicore, in *NIPS '06* (MIT Press, 2006), pp. 281–288
11. W.M.C. Bandara, A.S. Perera, and D. Alahakoon, Churn prediction methodologies in the telecommunications sector: a survey, in *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)* (2013), pp. 172–176

12. E.J. de Fortuny, D. Martens, F. Provost, Predictive modeling with big data: is bigger really better? *Big Data* **1**(4), 215–226 (2013)
13. R.E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification. *J. Machine Learning Res.* **9**, 1871–1874 (2008)
14. K. Dahiya, S. Bhatia, Customer churn analysis in telecom industry, in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)* (2015), pp. 1–6
15. J. Hadden, A. Tiwari, R. Roy, D. Ruta, Computer assisted customer churn management: State-of-the-art and future trends. *Comput. Oper. Res.* **34**(10), 2902–2917 (2007)
16. N. Kim, K.-H. Jung, Y.S. Kim, J. Lee, Uniformly subsampled ensemble (use) for churn management: Theory and implementation. *Expert Syst. Appl.* **39**(15), 11839–11845 (2012)
17. S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, A review of urban computing for mobile phone traces: Current methods, challenges and opportunities, in *KDD Workshop on Urban Computing* (2013), 2–9
18. E. Bonabeau, Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. U. S. A.* **99**(3), 7280–7287 (2001)
19. P.L. Brantingham, U. Glässer, B. Kinney, K. Singh, M. Vajihollahi, Modeling urban crime patterns, Viewing multi-agent systems as abstract state machines, in *Proc. ASM'05*, **3**, 101–1173 (2005)
20. A. Amin, C. Khan, I. Ali, S. Anwar, Customer churn prediction in telecommunication industry: With and without counter-example, in *Nature-Inspired Computation and Machine Learning*, ed. by A. Gelbukh, F. C. Espinoza, S. N. Galicia-Haro, (Springer, 2014), pp. 206–218. [https://doi.org/10.1007/978-3-319-13650-9\\_19](https://doi.org/10.1007/978-3-319-13650-9_19)
21. S.H. Han, S.X. Lu, S.C. Leung, Segmentation of telecom customers based on customer value by decision tree model. *Expert Syst. Appl.* **39**(4), 3964–3973 (2012). <https://doi.org/10.1016/j.eswa.2011.09.034>
22. V. Folcik, Orosz C., An agent-based model demonstrates that the immune system behaves like a complex system and a scale-free network. SwarmFest, University of Notre Dame, South Bend, IN, June, (2006)
23. O. Baqueiro, Y.J. Wang, P. Mcburney, F. Coenen, Integrating data mining and agent based modeling and simulation, in *Advances in Data Mining. Applications and Theoretical Aspects*, ed. by P. Perner, (Springer, Berlin, 2009), pp. 220–231

# Chapter 29

## Keyword-Based Approach for Detecting Civil Unrest Events from Social Media



J. Joslin Iyda and P. Geetha

### 29.1 Introduction

Nowadays, online social media plays a vital role in our daily lives and are the major way through which individuals interact on the Internet. The social networking sites like Facebook, Twitter, LinkedIn and MySpace enables the user to communicate with other users, or to find people with similar interests to one's own. And also online profiles can be created by the users in which they post daily updates about their lives in the form of pictures, videos, and related content. Facebook and Twitter have more than billions of users and it grows every day. Everybody started using social media ranging from normal people to celebrities, politicians, and media houses. They become prominent news source and can disseminate the information much faster than the traditional news media. Many real-world examples have shown the effectiveness and the timely information reported by Twitter during disasters and social movements. The following are the representative examples: the bomb blasts in Mumbai in November 2008, [1] the flooding of the Red River Valley in the United States and Canada in March and April 2009, the U.S. Airways plane crash on the Hudson River in January 2009, the devastating earthquake at Haiti in 2010, the demonstrations following the Iranian Presidential elections in 2009, and the "Arab Spring" in the Middle East and North Africa region.

---

J. J. Iyda (✉)

Anna University, Chennai, Tamil Nadu, India

Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

e-mail: [josliniyda.j@rajalakshmi.edu.in](mailto:josliniyda.j@rajalakshmi.edu.in)

P. Geetha

Department of Information Science and Technology, CEG, Anna University, Chennai, Tamil Nadu, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_29](https://doi.org/10.1007/978-3-030-19562-5_29)

287

As online social networking utilization turned out to be progressively interlaced with the occasions in the online world, people and organizations have discovered approaches to abuse these stages to spread wrong information [2], to assault and calumniate others [3], or to mislead and control. Clients with some tricky expectation may utilize this to spread bits of gossip, issue threats, give the wrong direction to their adherents and impart their tentative arrangements to their community [4, 5]. Criminal gangs and terrorist organizations like ISIS receive web-based social networking for purposeful publicity and enlistment [6]. Fraudulent action and social bots have been utilized to facilitate planned protest campaigns, to control political decisions and stock markets [7]. The absence of compelling substance confirmation frameworks and insufficient technical solutions to timely detect and ruin improper use on a considerable lot of these platforms, including Twitter and Facebook, raises concerns when more youthful clients disclose to cyber-bullying, harassment, or hate speech, initiating dangers like gloom and suicide. Moreover, online communications such as highly powerful social media are often used as a way of shouting out people's intentions before engaging in their acts of violence and also to coordinate criminal activities [8]. Being able to automatically detect negative material is beneficial to the managers of websites that allow users to post content or as part of an early warning system to authorities on possible threats to public safety [9]. The automatic detection of potentially dangerous words can help to ensure the safety of the public with minimum disruption. Thus monitoring social media posts and discussions, then figuring out how participants are reacting to a brand or event can improve the business [10–12]. Extraction of useful information from social media is more challenging than classic information extraction, i.e., extraction from trusted sources like traditional news media and well-formed grammatical texts. The actual challenge is in accessing that data and transforming it into something that is usable and actionable. Social media text [13] is typically very short, noisy, a high uncertainty of the reliability of the information conveyed in the text messages compared to conventional news media, and many social media support multi-lingual languages.

In this paper, we propose a keyword-based approach for detecting civil unrest events from twitter dataset. This system can automatically learn keywords from the dataset and the dataset is filtered based on these identified keywords. Then clustering analysis is performed in order to detect tweets promoting civil unrest and analyze the impact of the protest on the public. Finally, extensive experimental evaluation and performance analysis are performed.

## 29.2 Related Work

In recent years much attention is given to Online Social Network Mining due to the availability of enormous volume of uncensored data posted by people, which focuses on Social Recommendations, Opinion Mining, Sentiment Analysis, Topic Detection and Tracking, Community Detection, Event Detection, and Forecasting. This section presents related works in the following areas: (1) Spatiotemporal

mining of Social Media; (2) Event Detection and Forecasting; (3) Early detection of Suspicious Behaviors in Social Media; and (4) Civil Unrest event forecasting from Social Media.

### ***29.2.1 Spatiotemporal Mining of Social Media***

Considerable research work has been carried out by the researchers for studying the spatiotemporal event that is mainly relevant to the tweets posted within a certain geographical neighborhood. Thus, forecasting of such events requires an examination of spatial features and their correlations in addition to the temporal dimension. Ting Hua [14] reviewed several methods of spatiotemporal event detection and event forecasting. Judith Gelernter proposed a method for identifying locations and associating them with people by mining social media text conversations. Bo Hu [15] developed a probabilistic model for location recommendation by capturing the spatiotemporal aspects of user check-ins. Andrade [16] adopted a temporal approach for analyzing the cross-correlation between rainfall gauge data and rainfall-related Twitter messages by means of temporal units and their lag-time.

### ***29.2.2 Event Detection and Forecasting***

Most prior event detection research has focused on keywords present in the text also they rely on templates, dictionaries or presence of a specific pattern in the text. Wei Wang [17] extracted key sentences promoting civil unrest contain fields like participants, purpose, location and time using multiple instance learning. Yiming Yang [18] adapted the traditional hierarchical and non-hierarchical clustering techniques for online event detection based on semantic and temporal properties of events. Fang Jin [19] detected civil unrest events by representing the spatiotemporal structure of user activity in twitter in the form of graph wavelets. Minglai Shao [20] proposed a method to indicate the forthcoming or ongoing events in dynamic multivariate networks by measuring the significance of evolving sub graphs and subsets of attributes.

### ***29.2.3 Early Detection of Suspicious Behaviors in Social Media***

Considerable research work has been carried out in the area of Social Media Analysis. However, there has been relatively little work with respect to the early detection of Suspicious Behaviors targeting civil unrest, by observing text-based user's conversations. Some of the significant works are presented in this section.

Myriam Munezero [21] developed a framework to search for linguistic features that pertain to Anti Social Behaviors (ASBs) in order to use those features for the automatic identification of suspicious activities in texts. Dongjin Choi [22] proposed a method by using word similarity based on WordNet hierarchy and n-gram data frequency for distinguishing articles about terrorism. Burnap [23] built models that predicted information flow size and survival on Twitter following the terrorist event in Woolwich, London in 2013. Emilio Ferrara [24] has proposed a method to identify criminal networks from communication media such as mobile phones and online social networks that leave digital traces in the form of metadata.

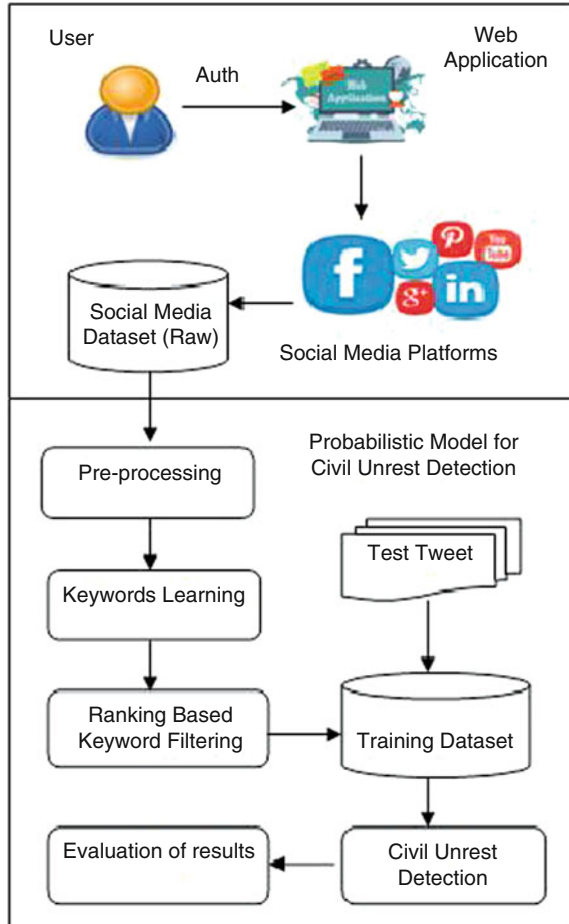
### ***29.2.4 Civil Unrest Event Forecasting from Social Media***

Many events with a large number of people gathering to support a common case are not civil unrest events [25] rather it is typically defined by law enforcement as a gathering of three or more people, in reaction to an event, with the intention of causing a public disturbance in violation of the law. Ryan Compton and Jiejun Xu [26] proposed a strategy by simply applying various filters like keyword filter, future dates filter, and location filter for early detection of civil unrest from social media. Congyu [27] proposed to locate the predictive power of social media in its function as a protest advertisement and organization mechanism from the Global Database of Events, Location, and Tone (GDELT).

## **29.3 System Framework for Civil Unrest Detection**

Social network analysis (SNA) has long been used for identifying social groups and for determining the relationships among the members of social groups. Figure 29.1 depicts the overall architecture of civil unrest detection system. It is divided into the following steps. First, all tweets between two dates are collected and preprocessed, where basic pre-processing steps are taken to clean the tweets and make them suitable for further processing. Second, automatic keyword learning is done based on the highest term frequency and significant keywords representing a particular protest are identified. Third, using this set of keywords the preprocessed tweets are filtered and the features used for detecting civil unrest are extracted from the resulting tweets. Fourth, clustering analysis is done to detect the essence of unrest content in those tweets in order to understand the influence of that protest on society.

**Fig. 29.1** The overall process of civil unrest detection



### 29.3.1 Preprocessing

The extracted tweets contain many unwanted words, symbols, white spaces, acronyms, etc., and such unwanted elements must be eliminated so that they can be easily processed in future and yield results with maximum accuracy. So the raw tweets were cleaned and preprocessed in order to remove the stop words, punctuations, and unwanted symbols. And the tweets written in natural languages are translated into English by Google Translate in order to process the tweets incrementally.



### 29.3.2 *Keyword Learning and Filtering*

Then the average term frequency and inverse document frequency score for each word are calculated and words were listed in decreasing order. Then the top ranked 100 words were selected and they were highly related to the cause for protest, place of protest and the key actors of protest. And the keyword matching was applied to the complete dataset using these protest-related terms. Keyword matching method is used to measure the tweets containing information about the upcoming protest. We measured the volume of tweets containing protest-related keywords and future-oriented words. First, we applied the keyword matching method. Since the tweets were extracted in the period of BusFareHike protest we tried with the basic keywords related to that protest like #BusFareHike, #TNBusStrike were the most popular hashtags of that protest. The tweets containing these keywords were selected and aggregated by day and thus we collected a huge volume of tweets containing the post of twitter for the period of 8 days for each protest.

### 29.3.3 *Clustering Model for Civil Unrest Detection*

The unsupervised learning is highly useful in social media monitoring as it enables us to obtain an overview of the public opinion about an event by applying various clustering techniques. Clustering is the technique of collecting the similar type of components in one cluster. Tweets containing information about the same event express collective behavior. This can be used to make different clusters having keywords representing various civil unrest events like #SaveFisherMen, #BusFareHike, and #Jallikattu. Simple TF-IDF algorithm is used for making clusters.

#### **Algorithm**

Civil Unrest detection based on keyword extraction will be performed in four general steps as below:

**Input:** Document containing tweets.

**Output:** Number of Clusters each representing different protest events.

**Step1:** Remove stop words and repeated tweets from each posts.

**Step2:** Extracting keyword of the user tweets based on TFIDF method:

TF-IDF value is composed of two components TF and IDF values. The logical basis of TF value is that more frequent words in a document are more important than less frequent words. TF value in a document is the number of times a given term appears in that document. The IDF, which measures the importance of a term in the collection. Dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient gives the value.

$$tf(i, j) = \frac{n(i, j)}{\sum_k n(k, j)} \quad (29.1)$$

$n(i, j)$ : The number of occurrences of the considered term in document  $d_j$   
 $\sum_k n(k, j)$ : The number of occurrences of all term in document  $d_j$

$$idf(i) = \log \left( \frac{|D|}{|d_j : t_j \in d_j|} \right) \quad (29.2)$$

$|D|$ : The total number of documents in the corpus  
 $|d_j : t_j \in d_j|$ : Number of documents where the term  $t_i$  appears

$$tfidf(i, j) = tf(i, j) \times idf(i) \quad (29.3)$$

**Step 3:** Calculate cosine distance between each tweet as a measure of similarity such that

$$\cos \theta = \frac{x \cdot y}{|x| \cdot |y|} \quad (29.4)$$

where  $x$  and  $y$  are term frequency-inverse document frequency (TF-IDF) vectors corresponding to documents  $x$  and  $y$ .

**Step4:** Clustering the tweets using the  $K$  Mean clustering algorithm.

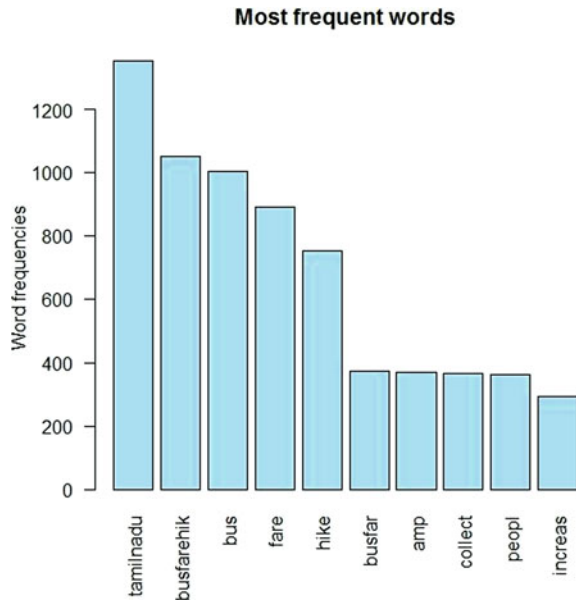
## 29.4 Results and Discussion

The implementation process starts with the data collection. Twitter API allows the users to extract information needed by providing them separate login and access credentials. These credentials are used to handshake with the R tool. The tweets were extracted using the Twitter API and R tool. The twitter posts were called tweets and that were collected in the period of 22/01/2018 to 29/01/2018 for #TNBusFareHike protest. We retrieved about 35,000 tweets; which contains people's opinions against the Tamil Nadu government for suddenly increasing the Bus Fare. Similarly, the dataset for #SaveFisherMen and #HydroCarbon protest was collected during the days of protest and they were aggregated by day. Thus we collected a huge volume of tweets for different protests.

Figure 29.2 shows the word cloud that is formed using the protest-related keywords identified from the tweets. The words that appear in bigger size are the words that appear frequently in the tweets. TF-IDF is the product of TF and IDF. When the Term Frequency is high and the Document Frequency is low (IDF is high)



**Fig. 29.3** Frequent words that appear on #busfarehike tweets



**Table 29.1** Keywords extracted to identify tweets of different protests

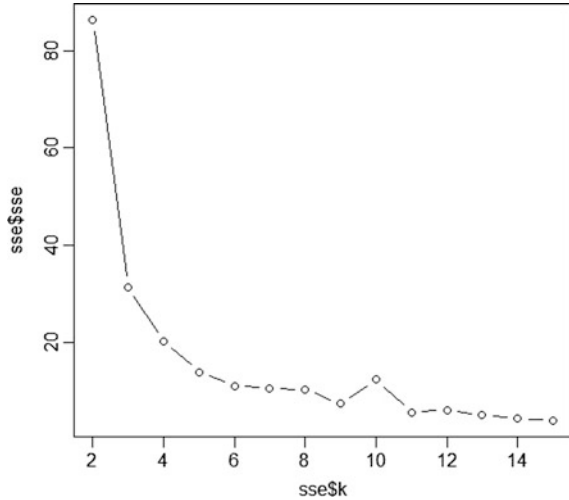
Event	Keywords	Word Frequency
#SaveFisherMan	Kanyakumari	4522
	Protest	1834
	Fishermen	1711
	Savetnfishermen	1699
#BusFareHike	Tamil Nadu	1352
	Busfarehike	1052
	Bus	1004
	Fare	891
#Jallikattu	Jallikattu	1619
	Protest	1461
	Students	1023
	Marina	989

When the clusters are well separated the “goodness” of the resulting clusters can be evaluated using Sum of Squared Error (SSE) to measure the compactness of the cluster. Sum of Squared Error (SSE) is calculated as,

$$SSE = \sum_{i=1}^n (x_i - \hat{x}_i)^2 \tag{29.5}$$

where each  $x_i$  is the actual value of observation, each  $\hat{x}_i$  is the estimated or forecast value of observation.

**Fig. 29.4** SSE of clusters formed using *K*-Mean clustering algorithm



By comparing the Sum of Squared Error (SSE) of the different number of clusters is one of the ways to determine the appropriate number of cluster. SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid. The plot of the SSE against the number of clusters *k* shown in Fig. 29.4 shows that as the *k*-value increases the SSE value decreases since clusters become smaller. In Fig. 29.4, the first elbow is found for the *k*-value 3. Thus the optimum number of clusters for the dataset is 3. To enable the detection and make the probability estimation feasible, we repeated the experiment using various datasets and the results were improved.

### 29.5 Conclusion

In this paper, we investigated existing text-mining methods for detecting civil unrest contents for preventing from the upcoming protest. Specifically, we proposed the Keyword-Based approach to detect civil unrest from social media before it may occur. We learned civil unrest keywords to train real-time tweets with clustering algorithm and tackled the problem of detecting civil unrest events. We integrated our ideas in a modular framework and experimentally demonstrated the validity and scalability of the method. The performance of the system can be improved, (1) to include location extraction method, by applying more advanced Geotagging scheme, using GPS signals, and by using information about the Twitter graph to estimate the location of a tweet from the location of related Twitter users, (2) multilingual text analysis can be applied to improve the clustering accuracy.

## References

1. <http://www.chinapost.com.tw/taiwan/national/national-news/2015/02/26/429715/Cabinet-on.html>
2. S. Wen, J. Jiang, X. Yang, S. Yu, To shut them up or to clarify: Restraining the spread of rumors in online social networks. *IEEE Trans. Parallel Distributed Syst.* **25**(12), 3306–3316 (2014)
3. E. Ferrara, Manipulation and abuse on social media. *SIGWEB Newsletter*, (Spring), 4 (2015)
4. J. Joslin Iyda, S. Visalaxi, G. Anitha, Discovering criminal communities from e-mails a graph-based approach. *J. Chem. Pharm. Sci. Spec. Iss.* **9**, 44–49 (2016)
5. M. Alzaabi, K. Taha, T.A. Martin, CISRI: A crime investigation system using the relative importance of information spreaders in networks depicting criminals communications. *IEEE Trans. Inf. Forensics Secur.* **10**(10), 2196–2212 (2015)
6. Charu C. Aggarwal IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, Counterterrorism, Social Network Analysis, in *Encyclopedia of Social Network Analysis and Mining* (Springer Book, 2014), pp. 285–289
7. Holly Paquette, Social media as a marketing tool: a literature review, Major Papers by Master of Science Students. Paper 2, 2013.
8. Xueyan Zhou, Jing Yang, Zehong Lin, Jianpei Zhang, “ITEPE: A source tracing algorithm for the microblog”, *PLoS ONE*, 9:e111380, 2014.
9. Mohammed Mahmood Ali, Khaja Moizuddin Mohammed, Lakshmi Rajamani, Framework for surveillance of instant messages in instant messengers and social networking sites using Data mining and Ontology, in *Proceeding of the 2014 IEEE Students’ Technology Symposium*, IIT Kharagpur, 2014
10. Z. Wang, W. Zhu, P. Cui, L. Sun, S. Yang, Social media recommendation, in *Social Media Retrieval, Computer Communications and Networks*, ed. by N. Ramzan et al., (Springer-Verlag, London, 2013). [https://doi.org/10.1007/978-1-4471-4555-4\\_2](https://doi.org/10.1007/978-1-4471-4555-4_2)
11. Yaniv Altshuler, Wei Pan, and Alex (Sandy) Pentland, Trends prediction using social diffusion models, in *Intl. Conf. on Social Computing, Behavioral-Cultural Modeling, and Prediction* (2012)
12. Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, Erel Uziel, IBM Research Lab, Social media recommendation based on people and tags, in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, July 19–23 July 2010
13. Mary McGlohon, Leman Akoglu, Christos Faloutsos, Statistical properties of social networks, in *Social Network Data Analytics* ed. by C.C. Aggarwal (Springer, Science Business Media, LLC, 2011). doi: 10.1007/978-1-4419-8462-3\_2
14. T. Hua, Z. Liang, F. Chen, C.-T. Lu, N. Ramakrishnan, How events unfold: Spatiotemporal mining in social media. *ACM Newslet. Sigspatial* **7**(3), 19–25 (2015)
15. Bo Hu, Mohsen Jamali, Martin Ester, Spatio-temporal topic modeling in mobile social media for location recommendation, in *IEEE 13th International Conference on Data Mining (ICDM)*, USA, 2013, ISSN: 1550-4786
16. S.C. de Andrade, C. Restrepo-Estrada A.C.B. Delbem, E.M. Mendiondo, J.P. de Albuquerque, Mining rainfall spatio-temporal patterns in twitter: a temporal approach, in *Societal Geo-innovation, 20th AGILE Conference on Geographic Information Science* (Springer, 2017)
17. Wei Wang, Yue Ning, Huzefa Rangwala, Naren Ramakrishnan, A multiple instance learning framework for identifying key sentences and detecting events, in *CIKM’16, ACM*, 24–28 October, 2016
18. Yiming Yang, Tom Pierce, Jaime Carbonell, A study on retrospective and online event detection, in *ACM Conference SIGR’98*, Melbourne, 1998
19. Fang Jin, Feng Chen, Rupinder Khandpur, Chang-Tien Lu, Naren Ramakrishnan. Absenteeism detection in social media, in *Proceedings of the SIAM International Conference on Data Mining (SDM’17)*, Houston, TX, Apr 2017

20. Minglai Shao, Jianxin Li, Feng Chen, Hongyi Huang, Shuai Zhang, Xunxun Chen, An efficient approach to event detection and forecasting in dynamic multivariate social media networks, in *ACM Conference WWW 2017*, Australia, 2017
21. M. Munezero, C.S. Montero, T. Kakkonen, E. Sutinen, Automatic detection of antisocial behaviour in texts. *J. Informatica* **38**, 3–10 (2014)
22. D. Choi, B. Ko, H. Kim, P. Kim, Text analysis for detecting terrorism-related articles on the web. *J. Netw. Comput. Appl.* **38**, 16–21 (2014)
23. P. Burnap, M.L. Williams, L. Sloan, Tweeting the terror: modeling the social media reaction to the Woolwich terrorist attack. *J. Soc. Netw. Anal. Mining* **4**, 206 (2014)
24. E. Ferrara, P. De Meo, S. Catanese, G. Fiumara, Detecting criminal organizations in mobile phone networks. *Int. J. Expert Syst. Appl.* **41**(13), 5733–5750 (2014)
25. A. Hoegh, S. Leman, P. Saraf, N. Ramakrishnan, Bayesian model fusion for forecasting civil unrest. *J. Technometrics* **57**, 332–340 (2015)
26. J. Xu, T.C. Lu, R. Compton, D. Allen, Civil unrest prediction: a tumblr-based exploration, in *Social Computing, Behavioral-Cultural Modeling and Prediction*, ed. by W. G. Kennedy, N. Agarwal, S. J. Yang, vol. 8393, (Springer, Cham, 2014). SBP 2014. Lecture Notes in Computer Science
27. C. Wu, M.S. Gerber, Forecasting civil unrest using social media and protest participation theory. *IEEE Trans. Comput. Soc. Syst.* **5**(1), 82–94

# Chapter 30

## Socioeconomic Status Classification of Geographic Regions in Sri Lanka Through Anonymized Call Detail Records



W. O. K. I. S. Wijesinghe, C. U. Kumarasinghe, J. Mannapperuma, and K. L. D. U. Liyanage

### 30.1 Introduction

In a comprehensive socioeconomic analysis, behavioural patterns are identified as a combination of economic activities, variations in mobility, and the networks of people who are associated with that particular individual. Therefore, this study is based on the argument that people of different SESs might exhibit different behavioural patterns, which are measured through income, occupation, and education [3]. The aggregation of these factors at a specific geographical region defines its SES standing.

#### 30.1.1 Problem

In order to identify the capability of people with respect to economic and social activities a socioeconomic status should be defined, as an index to measure the social and economic status of people.

The Census and Statistics Department in Sri Lanka has not yet defined an SES. Significant issues with results obtained from census data are: the time lag between the data acquisition and the result publication and this type of task is very costly. An up-to-date indicator of the behaviour of the people in a country can be taken as the CDRs, which include voice call, SMS, GPRS records along with recharge behaviour

---

W. O. K. I. S. Wijesinghe (✉) · C. U. Kumarasinghe · J. Mannapperuma · K. L. D. U. Liyanage  
Department of Computer Science and Engineering, University of Moratuwa, Bandaranayake  
Mawatha, Katubedda, Moratuwa, Sri Lanka  
e-mail: [isuru.10@cse.mrt.ac.lk](mailto:isuru.10@cse.mrt.ac.lk); [chamathk.10@cse.mrt.ac.lk](mailto:chamathk.10@cse.mrt.ac.lk); [jayaruwan.10@cse.mrt.ac.lk](mailto:jayaruwan.10@cse.mrt.ac.lk);  
[dananji.10@cse.mrt.ac.lk](mailto:dananji.10@cse.mrt.ac.lk)



of each mobile phone user. In the current context, the mobile phone usage in a developing country with respect to the population as a percentage, which is termed as the ‘mobile phone penetration rate’, is around 89%.

### 30.1.2 Purpose

The purpose of this study is to identify the socioeconomic status of geographic regions in Sri Lanka using anonymized CDR and using the census data processed and published by the Census and Statistics Department of Sri Lanka to validate our findings.

## 30.2 Methodology

### 30.2.1 Technology

The CDRs of millions of distinct users within a time period of 5 months generate a considerable amount of data, which falls into big data category. To solve this, the team read on technologies which can be used to handle the big data problem. Couple of them were MapReduce, Hadoop, PIG, Hive, Hortonworks, R, and RStudio. The overall high-level architecture of our approach is shown in Fig. 30.1.

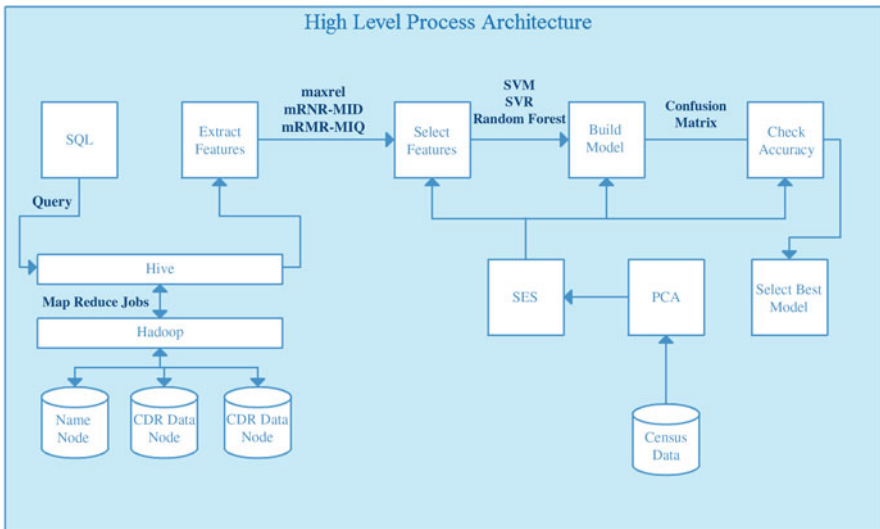


Fig. 30.1 High level process architecture

Even though Hortonworks provides an enterprise data platform, the problem with this is the Hortonworks Sandbox, the exact tool to practice Hadoop as a package, which enables a virtual environment. The main problem faced in setting up this was the high-end hardware requirements by the tool. Therefore, Hadoop was selected as the best tool to handle data [12].

Another reason for this choice is, the numerous other Hadoop related projects that can be used on top of Hadoop cluster. In this situation Hive that allows running queries against a Hadoop cluster has an SQL-like interface [12].

Since RStudio can't handle the initial data set, it was used to analysis and visualization tasks in the later stages of the process, where there is reduced amount of data.

### ***30.2.2 Cleaning and Preprocessing Data***

The following steps were taken to clean the raw data, which contained many duplicate records, invalid records, and missing values:

- Some of the attribute values have null values. For example, in voice call logs, some of the device names were null and therefore those entries were ignored.
- Some logs are repeated and therefore, all the duplicate entries were removed.
- All voice and GPRS logs are associated with a cell id in which the event has been occurred. But some of the logs contained cell ids which were not in the defined cell id list. Therefore, all logs containing non-relevant cell ids were removed.
- In refill data, there is an attribute to indicate whether the refill method is a top-up or a recharge card. But there was another category called 'other' and all the entries containing other as the refill method were removed.

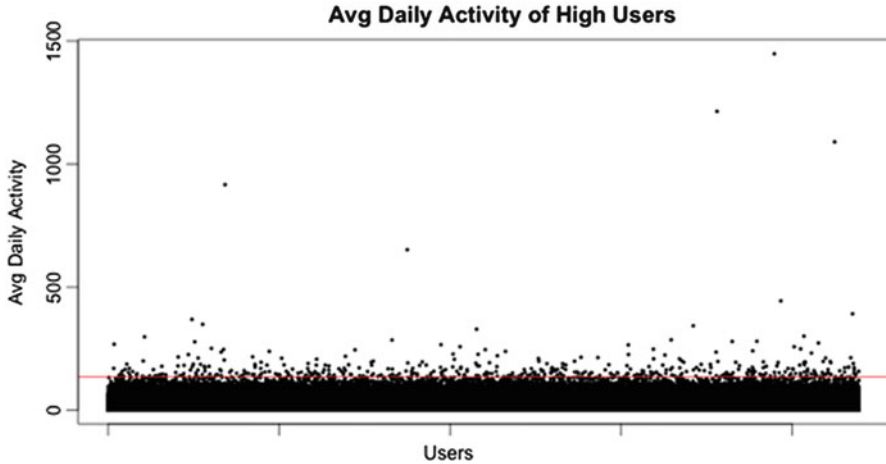
Besides cleaning the raw data set, all outlier entries were removed to reduce the noise in the data set. Outliers that are detected throughout the analysis process are listed as follows.

There were some individuals having a huge amount of daily activities when compared to other individuals. Therefore, an upper limit and a lower limit of daily activities were introduced to filter in normal users that were used to build the prediction model. As the lower limit, an average of 2 daily activity level has been selected [11]. After that filtration, 4.3 million users were categorized as high activity users.

According to Fig. 30.2, there exist a set of extremely high activity users which can be considered as outlier users. 99.99% of users out of high activity users have an average daily activity level less than 136.

### ***30.2.3 Feature Extraction***

We derived a set of features in order to classify each DSDs (330 DSDs in Sri Lanka) into socioeconomic levels. As the data set is based on individual user activities,



**Fig. 30.2** Distribution of average daily activity of high activity users

features were generated for individuals. It was important to extract as many features as possible. After extracting 107 features, we used feature selection techniques as described in Sect. 30.2.7 to obtain the best subset of relevant features for the target variable (socioeconomic status) from the entire feature set before feed into the prediction model. Features were broadly categorized into three main categories and they are as follows [4, 5].

### 30.2.3.1 Behavioural Features

These features imply the behavioural patterns of the users. We further divided this into multiple types such as call features (e.g. total outgoing calls, total incoming calls), refill features (e.g. total refill amount, total recharge amount, total top-up amount), device features (e.g. total unique devices used, voice unique devices used, GPRS unique devices used), call duration features (e.g. average outgoing call duration, average incoming call duration), IDD and local calling features (e.g. total outgoing calls IDD, total outgoing calls local), and GPRS features (e.g. total sessions for the whole period).

### 30.2.3.2 Mobility Related Features

Mobility features represent user's activity location wise distribution and movement. The number of distinct cells the user has made calls (user location varies with point of activity) is capable indication of mobility. Total distinct cells (voice and GPRS), total voice distinct cells, distinct cells for outgoing calls, distinct cells for incoming calls, total GPRS distinct cells, total distance travelled, diameter of the

area of influence, and radius of gyration are some mobility related features which we have extracted from the data set.

### 30.2.3.3 Social Network Related Features

This type of features indicates how users are connected to other users. For instance, distinct people outgoing calls (no of distinct people, called by a user), distinct people incoming calls (no of distinct people, called to a user), total distinct people connected, max percentage of outgoing calls, and max percentage of incoming calls were few of features which we have derived.

## 30.2.4 Home Location Detection

One major phase of this research was to locate the individual users to their home DSD locations. This had become more important because the prediction model is built for DSDs and not for individual users. To locate the individual users, the user activities are analysed over 5 months.

It is observed that for most individuals the work location and the home location are at different geographical locations. Thus, most of the individuals other than a smaller fraction of users who are staying at home throughout the day have to go to their workplace and come back home for each day. Therefore, most of the users leave their home for work in the morning and stay at the workplace during the day. Accordingly, there shouldn't be a drastic movement during work hours. At the end of the day, users are leaving their workplace to their home and they will stay at home at night until the morning of the next day. Therefore, we concluded that there should be two peaks of human movement for a given week day. During the study we used average movement of individual users for each hour of the day [7].

According to Fig. 30.3, the claim that there has to be two peak movements within a day can be proved. Therefore, work hours and home hours can be defined based on the movement patterns as follows:

$$\begin{aligned} \text{Work Hours} &= 10:00 \text{ AM} - 03:00 \text{ PM} \\ \text{Home Hours} &= 09:00 \text{ PM} - 05:00 \text{ AM.} \end{aligned}$$

On the results of the algorithm, the analysis has been done with the population density census in 2012 assuming that there is no significant density variation in population in 2012 and 2013. The statistics of the analysis is as follows:

$$\begin{aligned} \text{Correlation Coefficient} &= 0.8687 \\ R\text{-squared} &= 0.7545. \end{aligned}$$

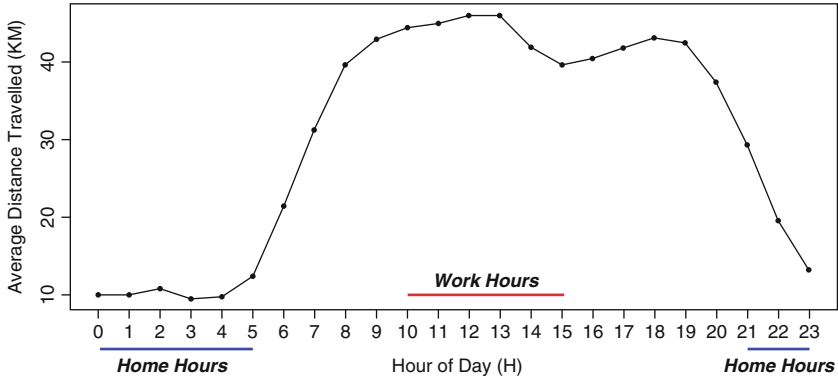
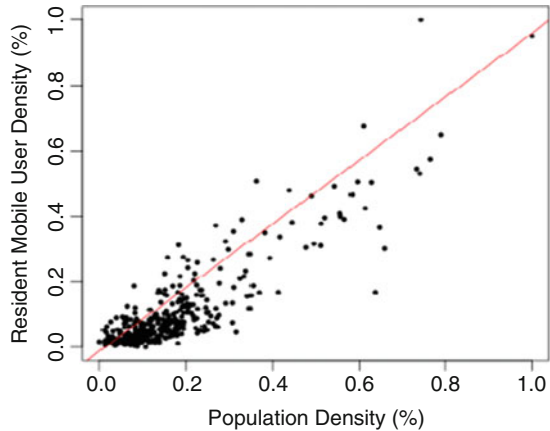


Fig. 30.3 Average movement by an individual in hours of the day

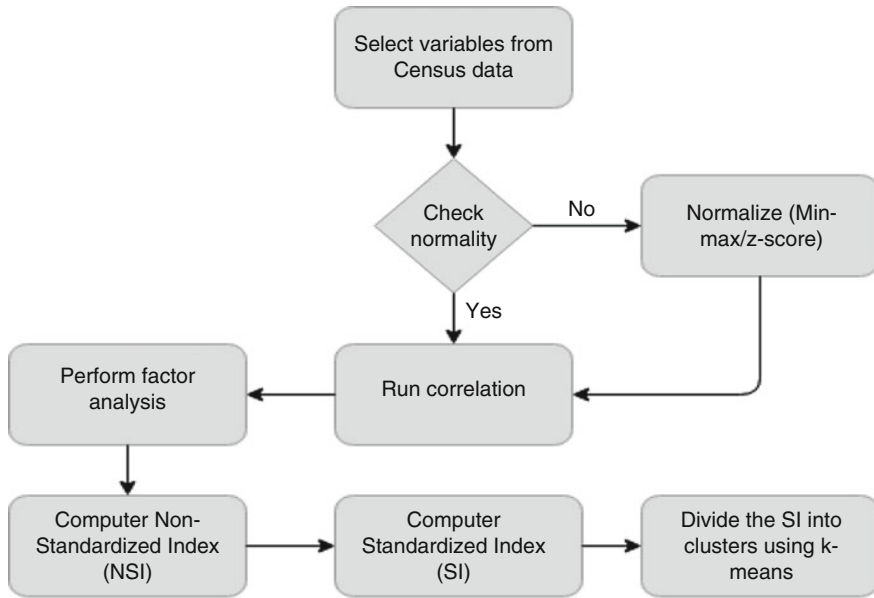
Fig. 30.4 Scatter plot for population and resident mobile user density



In Fig. 30.4, all the data points are gathered around the regression line having *R*-squared value of 0.7545 which is closer to 1. Therefore, the algorithm succeeds in locating users to their boarding places by analysing activities of 5 months. By comparing the density maps in Fig. 30.5, the correlation coefficient can be justified.

### 30.2.5 Socioeconomic Index Creation

The most important step is to develop a socioeconomic index for each DSD in Sri Lanka. Principal components analysis (PCA) is used to build a composite index using components and those components capture the largest possible variation in the original features. First we built an algorithm to calculate socioeconomic index from variable selection to index construction with special attention to the normalization (*z*-score normalization) procedures as well as factor analysis. Under



**Fig. 30.5** Socioeconomic index construction process

following sections, the computation procedure of the composite socioeconomic index is discussed within the perspective of PCA.

To provide a proxy measure of socioeconomic status at DSD level, we used several indicator types that were available in census data to build composite index for DSD level and then some feature types converted into headcount ratio, i.e. percentage of population and others converted into household ratio at each DSD. The indicator types we selected for the study of socioeconomic index creation are principal type of lighting, household type, principal materials of construction of roof, principal material of construction of wall, and sex and age [1, 10].

For the purpose of factor analysis, we used 36 features from census data. Since the features were not in the standardized form, the correlation matrix was used as an input to PCA to extract the feasible factors. SPSS statistical tool was used to extract certain factors with eigenvalue rule. We recognized ten factors by plotting the eigenvalue against the number of components.

The ten factors described 77.891% of the total variation and the percentage for each factor from the total variation is different with each other. Therefore, those factors should not impact socioeconomic condition in equal weights. Thus, using the proportion of variation with respect to the total variation for each factor as weights on the factor score coefficient, we developed non-standardized index (NSI) for each DSD using the following formula [8]:

$$\text{NSI} = \sum_{i=1}^{10} \left\{ \frac{\text{variance of } i\text{th component}}{\text{total variance}} \right\} * \text{factor score of } i\text{th component.}$$

The above non-standardized index measures the socioeconomic status of one DSD that is relative to the other on a linear scale. The interpretation is somehow difficult to us because this value gives either positive or negative value. Thus, we converted this non-standardized index to a standardized index (SI) using min–max normalization in the range of [0–100].

$$\text{SI} = \left\{ \frac{\text{NSI of DSD} - \text{MinNSI}}{\text{MaxNSI} - \text{MinNSI}} \right\} * 100$$

### ***30.2.6 Classification of DSDs into Socioeconomic Status Groups***

The socioeconomic index is not uniformly distributed among all DSDs in Sri Lanka. To identify the socioeconomic status of each DSD, first we grouped the SI values using  $k$ -means into clusters [2] and identified average values for each cluster. According to that we recognized five levels of socioeconomic status for Sri Lanka and assigned  $k = 5$  in  $k$ -means clustering algorithm to divide SI indexes into five levels or groups, namely very low, low, average, high, and very high.

### ***30.2.7 Feature Selection***

After the initial preprocessing stage, we built our training data set and it contains 330 rows one for each DSD. Each of these DSDs is composed of 107 features with its desired target class (socioeconomic status). In order to improve the accuracy of our model we selected the best set of features that are more relevant in our data set. For that purpose, we used two feature selection techniques called MaxRel algorithm and mRMR algorithm [6, 9].

### ***30.2.8 Model Building, Prediction, and Evaluation***

The classification problem could be regulated as assigning one of the socioeconomic levels from very low, low, average, high, very high to a given DSD. We used several classification methods; however, we only represent the results obtained from support

vector machines (SVM), which provided the best classification rates for our data set. We have verified the classification methods with the feature vectors ordered agreeing to each one of the three feature selection techniques called MaxRel, mRMR-MID, mRMR-MIQ described above in order to recognize which one produces better results.

We used non-linear SVM classification method with radial basis function (RBF) kernel. At this point for RBF kernel value we choose value '1'. For each feature selection order which was produced by MaxRel, mRMR-MID, and mRMR-MIQ and for each subset of ordered features in all 107-feature set, we recognized the optimum value for RBF kernel parameter that maximizes the accuracy using hold-out validation over the training data set.

We tested each SVM model using the test set which was produced by hold-out validation. The following figure shows the accuracy for each subset of ordered features for the three feature selection techniques that we used. We observed that mRMR-MID produces better accuracy than mRMR-MIQ or MaxRel. We found that when using the top 8 features the mRMR-MID is obtained as the best result with 81.65% accuracy rate. The confusion matrix when using top 8 features is as follows.

As seen in Fig. 30.6, the predictions in the diagonal are taken as correct and every other prediction is taken as incorrect. But as seen from the confusion matrix very little outliers can be seen and very high proportion of the predictions are close to the diagonal.

As seen in Fig. 30.7 the max accuracy can be seen 81.65% when the number of features considered is 8 and then it drops gradually with each feature addition or removal.

Figure 30.8 shows the maximum accuracy for MaxRel feature selection and as seen from the graph it is at 76% when the number of features considered is the top 20.

Figure 30.9 shows the maximum accuracy for mRMR-MIQ feature selection and looking at the graph, it is approximately 77% when the number of features considered is only the top 5.

		Actual	very-low	low	average	very-high	high
			1	2	3	4	5
Prediction							
very-low	1		4	2	0	0	0
low	2		0	26	6	1	0
average	3		1	2	35	2	1
very-high	4		0	0	4	24	0
high	5		0	0	0	1	0

Fig. 30.6 mRMR-MID accuracy confusion matrix for SVM



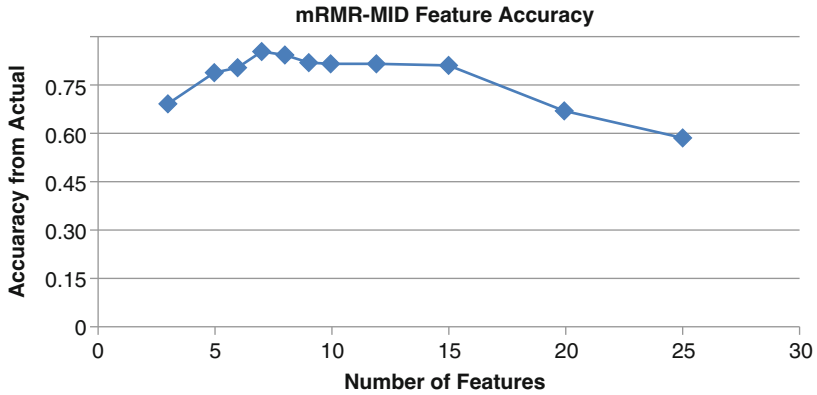


Fig. 30.7 mRMR-MID features accuracy for SVM graph

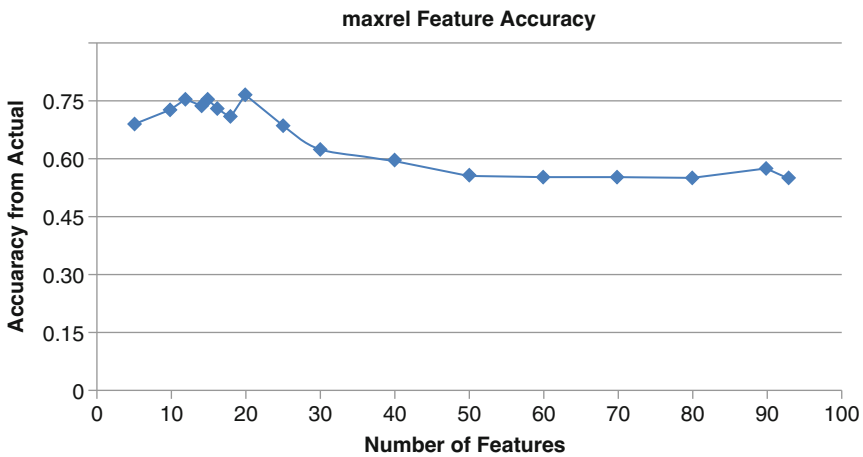


Fig. 30.8 MaxRel features accuracy for SVM graph

### 30.3 Discussion on Results

The accuracy of the prediction model is directly related with the accuracy of the home detection algorithm and SES calculation from PCA.

#### 30.3.1 Accuracy of the Home Detection

Misidentifying user home locations due to noisy users who doesn't represent general behavioural patterns can distort the results. If the user home location is distorted the aggregation of features to the DSD won't represent the real feature value of the DSD.

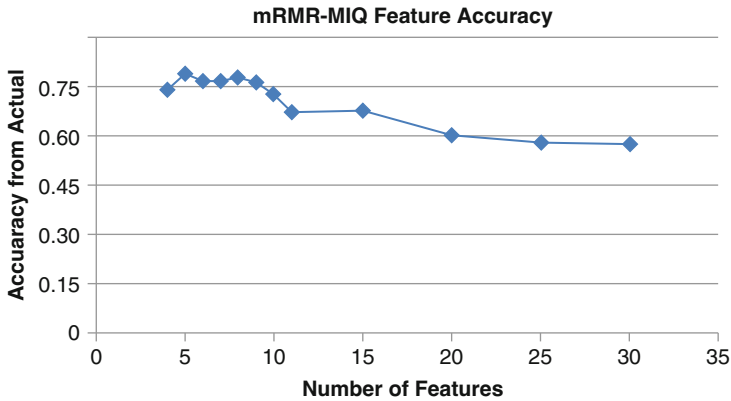


Fig. 30.9 mRMR-MIQ features accuracy for SVM graph

### 30.3.2 Accuracy of SES Calculation from PCA

For PCA to give SES, many SES features need to be considered. If the prediction of SES is different to real SES level of the region, then although CDR model gives accurate results the accuracy of the model will be seen as moderate.

## 30.4 Future Work

There are multiple factors that can be considered to classify SES of DSDs to improve the predictability of the model. These factors may include:

- Night-time Satellite Imagery taken by National Aeronautics and Space Administration (NASA).
- Electricity usage by DSD taken from the records of Ceylon Electricity Board in Sri Lanka.

## 30.5 Conclusion

Over 100 features were extracted from the data set and for that the data were pre-processed, filtered for outliers, and aggregated for DSDs. Using commonly referred feature selection algorithms in literature such as MaxRel, mRMR-MID, and mRMR-MIQ and using model building techniques such as SVM, random forest, neural networks, and advanced technique such as ensemble method, we were able to predict socioeconomic status (SES) of geographic region (GR) from call detail

record (CDR) with a high level of accuracy of 81.65%. These CDR classification models can be used to make a moderate estimation on SES of a region and whether it has increased or decreased given CDR data for short time period.

SES classification from CDR happened for the first time in Sri Lanka and using refill features for modelling is not seen in any literature to date.

This research can be used as a stepping stone for future researchers who wish to improve the prediction model to a very high level of accuracy which this modelling technique has the potential to be. The results and the methodology in the PCA section for SES classification can be used to classify divisional secretariat divisions (DSDs) in Sri Lanka by its SES, using the census data. SES classification on DSD using PCA result can be straightaway used in public policy making. The historical values of SES classification can be compared in the future for changes in SEL.

In conclusion it can be said that SES classification by CDR data is the way forward if policymakers are looking at saving money, time, and adapting to quick changes in conditions that cannot be taken from census information.

**Acknowledgements** The authors would like to express their heartiest gratitude towards Mr. Nisansa de Silva for giving us advice and helping us through the problems we faced during the research. Moreover, they would like to thank Dr. Shehan A. Perera for sharing his expertise knowledge in the field and spending his valuable time to guide us through the research.

## References

1. U. Amarasinghe, M. Samad, M. Anpuhas, Spatial clustering of rural poverty and food insecurity in Sri Lanka. *Food Policy* **30**, 493–509 (2005)
2. R.A. Becker, R. Caceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky, Clustering anonymized mobile call detail records to find usage groups, in *1st Workshop on Pervasive Urban Applications* (2011)
3. C. Cowan, R. Hauser, R. Kominski, H. Levin, S. Lucas, S. Morgan, M. Spencer, C. Chapman, Improving the measurement of socioeconomic status for the national assessment of educational progress: a theoretical foundation (recommendations for the National Center for Education Statistics), Washington (2012)
4. V. Frias-Martinez, J. Virseda, Cell phone analytics: scaling human behavior studies into the millions. *Inf. Technol. Int. Dev.* **19**, ICTD2012 (2013)
5. V. Frias-Martinez, J. Virseda-Jerez, E. Frias-Martinez, On the relation between socio-economic status and physical mobility. *Inf. Technol. Dev.* **18**, 91–106 (2012)
6. V. Frias-Martinez, V. Soto, J. Virseda, E. Frias-Martinez, Can cell phone traces measure social development? in *Third Conference on the Analysis of Mobile Phone Datasets*, (NetMob, 2013)
7. S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, Identifying important places in people's lives from cellular network data, in *Pervasive Computing* (Springer, Berlin, 2011)
8. V. Krishnan, Constructing an area-based socioeconomic index: a principal components analysis approach. Early Child Development Mapping Project, Edmonton, Alberta (2010)
9. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005)

10. A. Satharasinghe, Census department classifies GN divisions by poverty, Department of Census and Statistics Sri Lanka (2012)
11. V. Soto, V. Frias-Martinez, J. Virseda, E. Frias-Martinez, Prediction of socioeconomic levels using cell phone records, in *User Modeling, Adaption and Personalization* (Springer, Berlin, 2011)
12. H.R. Varian, Big data: new tricks for econometrics. *J. Econ. Perspect.* **28**, 3–27 (2014)

**Part II**  
**Workshop on the Analysis of Big Data**

# Chapter 31

## Hand Gesture Based Human-Computer Interaction Using Arduino



S. Shreevidya, N. Namratha, V. M. Nisha, and M. Dakshayini

### 31.1 Introduction

Many people use computer as one of the daily tools. Specially designed input and output devices are making the computer and human communication easier. The two known input and output devices are mouse and keyboard. In this era of evolving technologies, the latest hardware expedients have made the computer still more efficient and intelligent and thus enabling humans to be able to perform more complex interaction or operations with the computer. This has been made successful by creating human computer interfaces [1]. The communication between computer and human has been made successful by the system programmers. Its main aim is to reduce the complexity of operations accomplished with the advent of all new products in the market.

It has helped in facilitating many operations like tele-operation, robotics and human beings controlling the complex systems like monitoring systems, aeroplanes and cars for instance [2]. Previously, this type of intricate programs was avoided by the computer programmers. The major focus was the speed and flexibility than other features that could be modified. Nevertheless, a shift was made on people friendly Computer Human interfaces and to make computer to understand human language.

The non-verbally exchanged information is Gesture. Various expressions of human face and speech play a significant role in making a computer system to understand human gestures.

Enormous assortments of gestures can be performed by a man each one in turn. For PC vision analysts this has turned into an extraordinary enthusiasm, since the

---

S. Shreevidya (✉) · N. Namratha · V. M. Nisha · M. Dakshayini  
Department of ISE, BMS College of Engineering, Bangalore, India  
e-mail: [Shreevidya.scn17@bmsce.ac.in](mailto:Shreevidya.scn17@bmsce.ac.in); [namratham.scn17@bmsce.ac.in](mailto:namratham.scn17@bmsce.ac.in);  
[nishavm.scn17@bmsce.ac.in](mailto:nishavm.scn17@bmsce.ac.in); [dakshayini.ise@bmsce.ac.in](mailto:dakshayini.ise@bmsce.ac.in)

signals are seen through vision. The paper tries to determine the human gestures by creating an HCI [3]. A complex programming algorithm is made in coding of these gestures into machine language. To gain knowledge an abstract view of gesture recognition system is provided. Some of the applications of hand gesture operation are virtual reality, sign language, etc.

## 31.2 System Design

The proposed system aims at reducing usage of keyboard and mouse by implementing a reliable and simple hand gestures [4] controlling system for computers (Fig. 31.1).

### 31.2.1 Algorithm

#### Gestures used:

Gesture 1: (right sensor)

- When the elderly or blind people are watching any videos and listening to audios in the computer, if they want to increase or decrease volumes then they can use hand gestures which is used for volume increase and decrease operations

Gesture 2: (right sensor)

- When the elderly people are using web application in the computer, if they want to scroll up or down simultaneously then can use the hand gestures for those operations

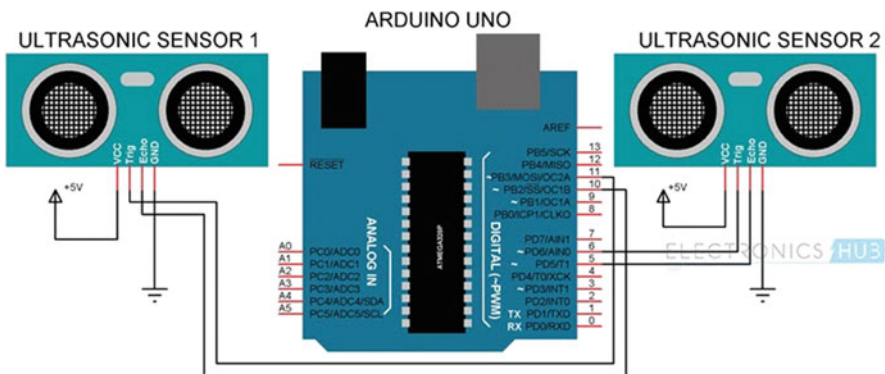


Fig. 31.1 System design of hand gestures based computer

#### Gesture 3: (right sensor)

- With the hand gestures on right sensor the elderly people can move to the next tab.

#### Gesture 4: (left sensor)

- With the hand gestures on right sensor the elderly people can move to the previous tab.

#### Gesture 5: (both sensor)

- With the hand gestures on both the sensors the elderly people can switch between the tasks.

This paper is useful for every aged people but mainly useful for the elderly and blind people since the usage of mouse and keyboard are complicated for such people in some operations.

### **31.2.2 System Model**

The proposed model of the hand gestures based Human Machine Interaction using Arduino is as shown in Fig. 31.2 and consists of the following:

1. User
2. Arduino board
3. Ultrasonic sensors
4. Laptop with internet connection
5. USB cable
6. Connecting wires

### **31.3 Implementation**

The connection setup for the hand gesture based computer is as shown in Fig. 31.3.

The circuit design of the system is made simple, but setting up the component is very crucial. The echo pins and trigger are connected to pins 10 and pins 11 of the Arduino which are placed on the left of the first ultrasonic sensor. The echo pins and trigger pins of the second Ultrasonic Sensor are connected to pins 5 and 6 of the Arduino.

Now, both the ultrasonic sensors are placed on top of the laptop screen, one Ultrasonic Sensor at the left end and the other Ultrasonic Sensor at right end. The double-sided tape is used to hold the sensors onto the screen if needed.



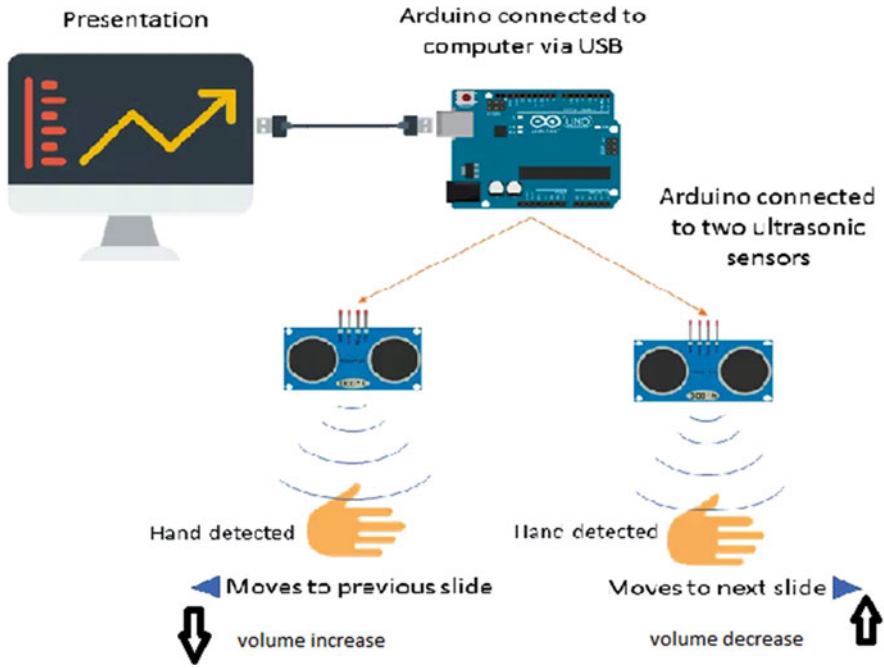
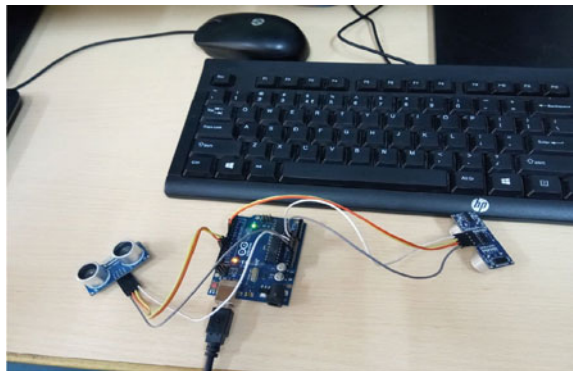


Fig. 31.2 System model of hand gestures based computer

Fig. 31.3 Connection setup for hand gesture based computer



The hand gestures of the ultrasonic sensors in front can be calibrated so that the gestures perform five different actions on the computer. Before taking a look at the gestures, the tasks that accomplished are to be seen

- Switching from the current tab to the next in an internet/web browser
- Scrolling downwards operation in a web/internet page
- Scrolling upwards operation in a web/internet page

- Switching between the two actions that are operating at a time (browser and media player)
- Playing/Pausing video or audio function in a VLC media player
- Increasing the volume
- Decreasing the volume

The below given are the five different actions of hand gestures and the actions that are programmed for the purpose of demonstration.

**Gesture 1:** The ultrasonic sensor on the right side of the computer or laptop is placed at a distance (between 15 and 35 cm), place the hand in front of the sensors, for a lesser duration, and move away the hand gradually from the sensor. The gesture performs the operation of scrolling downwards the web/internet page or decreases the volume.

**Gesture 2:** The ultrasonic sensor that is placed on the left side of the computer or laptop is placed at a distance (between 15 and 35 cm), place the hand in front of the sensors, for a lesser duration, and move the hand gradually towards the sensor. The gesture performs the operation of scrolling upwards the web/internet page or increases the volume.

**Gesture 3:** The ultrasonic sensor that is placed on the right side of the computer or laptop, swiping the hand in front of the ultrasonic sensors. The gesture performs the operation of moving to the next tab.

**Gesture 4:** The ultrasonic sensor that is placed on the left side of the computer or laptop, swiping the hand in front of Ultrasonic sensors. The gestures play out the activity of moving to the past tab or play/delay the video.

**Gesture 5:** Swiping the hand crosswise over both the sensors (left sensor first). The activity switches between the assignments.

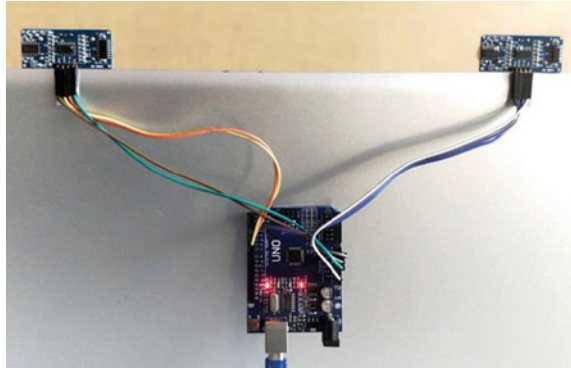
## 31.4 Result

The working model of the hand gesture based computer has successfully avoided the inconvenience caused due to usage of keyboard and mouse for the elderly and blind people. Using the hand gesture [5] operation it is implemented that the volume can be increased/decreased, move to the next tab or previous tab, or perform both the operation using both the ultra sonic sensors (Figs. 31.4 and 31.5).

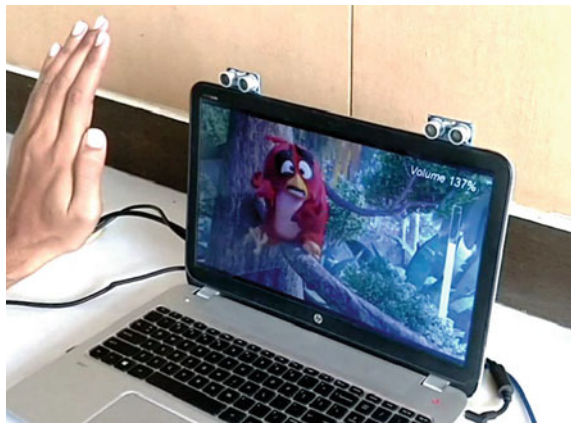
## 31.5 Conclusion

The usage of computer is growing in the present era. In recent days, as elderly and blind people are facing problems due to the usage of keyboard and mouse, this paper has tried to provide some solutions for such kind of people who are facing problems in operating the keyboard and mouse. All the above operations

**Fig. 31.4** Arduino and sensors are connected to the laptop



**Fig. 31.5** Working of the hand gestures



performed were implemented successfully. Hence one of the problems like blind people who cannot operate using the hands just by the movement and the knowledge of hand gesture this can be applied. In this paper, we have implemented Human Computer interaction using Arduino by making the hand gestures [6], where little hand gestures are made utilizing mouse or console before the PC which will play out specific undertakings in the PC without the utilization of mouse and console.

Those Gestures based Control of Computers are as of now present and an organization named Leap Motion has been actualizing the innovation in PCs.

### 31.6 Future Scope

The proposed system depends completely on the user to avoid the usage of mouse and keyboard. This sort of hand motion based control of PCs can be utilized for AR (Augmented Reality), 3D plan, VR (Virtual Reality), perusing communication through signing, and so forth.

## References

1. Y. Xu, C. Lee, Online, Interactive learning of gestures for Human/Robot interfaces, in *IEEE International Conference on Robotics and Automation*, Minneapolis, MN, USA, 2002
2. A. Pentland, T. Starner, J. Weaver, A wearable computer based American Sign Language (ASL) recognizer, Digest of Papers, in *First International Symposium on Wearable Computers*, Cambridge, MA, USA, 2002
3. H.-K. Lee, J.H. Kim, An HMM-based threshold model approach for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell* **21** (1999)
4. C. Nölker, H. Ritter, Detection of fingertips in human hand movement sequences, in *Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop, Bielefeld, Germany, September 17–19, 1997, Proceedings*, ed. by I. Wachsmuth, M. Fröhlich, (Springer, Berlin, 2001), pp. 209–218
5. E. Ueda, Y. Matsumoto, Individualization of voxel-based hand model, in *IEEE International Conference on Human-Robot Interaction (HRI)*, La Jolla, CA, USA, 2012
6. S. Ranganath, C.W. Ng, Real time Gesture recognition system and application. *Image Vis. Comput.* **20**, 993–1007 (2002)

# Chapter 32

## An Automatic Diabetes Risk Assessment System Using IoT Cloud Platform



M. Sujaritha, R. Sujatha, R. Anitha Nithya, A. Sunitha Nandhini,  
and N. Harsha

### 32.1 Introduction

Diabetes mellitus is a disease which occurs when there is uncontrolled blood sugar level in the body. The maintenance of blood sugar levels in the body is the duty of insulin, a hormone secreted by the pancreas. This chronic disease is characterized by the improper functioning of the pancreas, where it becomes incapable to produce sufficient insulin [1]. It may also happen due to the resistance of the body to insulin. Their symptoms of this disease are frequent urination (polyuria) and increased appetite (polyphagia) and consumption of water (polydipsia). The diabetes is categorized as Type 1, Type 2 [4], and gestational based on its cause and severity [15].

Type 1 diabetes is an auto-immune disease where the insulin secreting cells are attacked by the immune system of the body. This is treated by injecting enough insulin lifelong. This Type 1 is most common among children and young adults. People affected by this type of diabetes become insulin dependent. Type 2 is related to insufficient physical activity, poor diet, obesity or overweight, and hereditary issues. People affected by this type of diabetes can come out or manage by changing their lifestyles. Nowadays, it occurs at all age groups and is also becoming predominant in younger age groups. Since this type 2 diabetes can be prevented by creating awareness among younger generations and providing proper guidelines in

---

M. Sujaritha (✉)

Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India  
e-mail: [sujaritham@skcet.ac.in](mailto:sujaritham@skcet.ac.in)

R. Sujatha · R. A. Nithya · A. S. Nandhini · N. Harsha

Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

e-mail: [r.sujatha@skct.edu.in](mailto:r.sujatha@skct.edu.in); [r.anithanithya@skct.edu.in](mailto:r.anithanithya@skct.edu.in); [sunithanandhini.a@skct.edu.in](mailto:sunithanandhini.a@skct.edu.in);  
[17tpcs002@skct.edu.in](mailto:17tpcs002@skct.edu.in)

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_32](https://doi.org/10.1007/978-3-030-19562-5_32)

323

diet and exercises, the proposed system has been designed. Gestational diabetes happens for women during their pregnancy and it vanishes after the delivery of the child. It may sometimes lead to type 2 diabetes [14] later in their life. The gestational diabetes can also be managed by practicing healthy dietary and exercise habits.

Therefore the proposed system has been developed to provide adequate awareness to the younger generations through analyzing the diabetes risk factor and offering their dietary and exercise plans. Almost 108 million people over world suffered from diabetes in the year 1980 and it has been increased to 382 million people in the year 2013 and 500 million in the year 2018 [4]. Also the age structure of the worldwide population has been shifted gradually, such that the occurrence of diabetes mellitus is currently 10.5% among adults, nearly double the rate of 5.6% in 2000 and the type of diabetes is type 2 as well [9]. As per World Health Organization's (WHO) report, almost 1.5 million people over the globe died owing to diabetes in 2012, it became the eighth leading cause of death. Another 2.2 million deaths over globe which were applicable to high blood glucose and the increased risks of cardiovascular disease and other associated complications (e.g., kidney failure), which often lead to premature death [12]. People with diabetes often require medication regimes to control high blood glucose levels. In addition, affected persons may also require medications to reduce high blood pressure and/or cholesterol levels [6]. Persons affected by type 1 diabetes require regular injections of insulin (a protein that removes excess glucose from the blood) in order to regulate their blood glucose levels (and to survive), and some persons with type 2 diabetes also require insulin in cases where their diabetic condition is difficult to control [11]. Therefore a personalized system which regularly provides health guidelines to the person is required. This issue is addressed in this paper.

## 32.2 Diabetes Risk Assessment Tools

The assessment tools have been designed in each country to identify the diabetes risk factors and providing suggestions to control it.

The Madras Diabetes Research Foundation has devised the diabetes screening tool named Indian Diabetes Risk Score (IDRS). It analyzes family history, waist circumference, age, and physical activity of the individual [10]. Hence, individual with high risk of developing diabetes mellitus in near future can be identified and systematic counseling and further interventions can be applied in order to reduce diabetes related complications [13]

The Australian Type 2 Diabetes Risk Assessment Tool was developed by the Baker IDI Heart and Diabetes Institute on behalf of the Australian, State and Territory Governments as part of the COAG initiative to reduce the risk of type 2 diabetes (AUSDRISK) [3].

The European Agency for Health and Consumers (EAHC) and the IMAGE Group 2006309-IMAGE have developed the IMAGE—Toolkit [8], which consists of practice-oriented European guidelines for the prevention of type 2 diabetes

(T2DM) and step-by-step tips on how to initiate and manage a lifestyle intervention to prevent and type 2 diabetes. The American Diabetes association [2] and Korean associations [7] have also developed a tool to create awareness on diabetes among Americans.

In addition to these tools or toolkits, four mobile applications have been developed to guide people in maintaining their blood sugar levels. They are: (1) GoMeals, (2) Glucose Buddy Diabetes Tracker, (3) SparkRecipes, and (4) BlueStar Diabetes. However all these tools and Apps are not complete, since the data collected are not properly analyzed through machine learning tools [5] and stored in a common cloud platform. The proposed system uses ThingSpeak IOT based cloud platform and analyzes the data using machine learning techniques in MATLAB environment.

### 32.3 The Proposed System

The proposed system consists of Feet Pressure Sensor, Blood Pressure Sensor, and ECG Electrodes to get the input for the system and Raspberry PI board to analyze the data and store it in the cloud (Figs. 32.1 and 32.2).

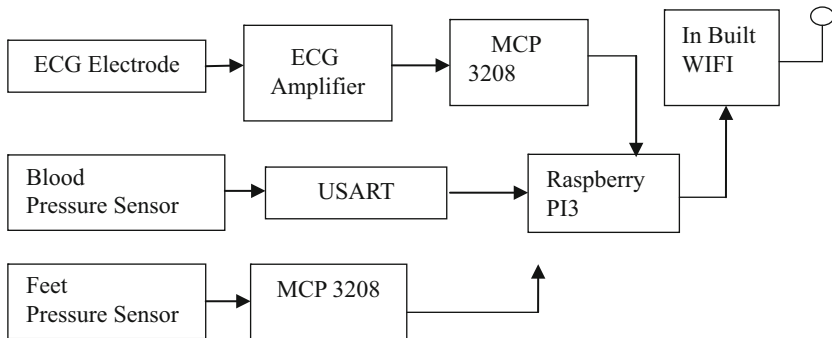


Fig. 32.1 The block diagram of the proposed system

Fig. 32.2 The proposed tool kit for diabetes risk assessment



## 32.4 Conclusion

Diabetes is an incurable disease which occurs in both adults and young aged people as their symptoms can cause variable reactions in body which might lead to chronic diseases. A risk model is designed to predict high and low risk levels in dataset retrieved from biosensors which are used for monitoring health status like electrocardiogram, glucometer, blood pressure sensor, and feet pressure sensor and all these data are stored in cloud and can be retrieved as comma separated form and analyzed using machine learning tools for providing personalized health care.

## References

1. A.A. Abdullah, M.G. Ahamad, M.K. Siddiqui, Application of data mining: Diabetes health care in young and old patients. *J. King Saud Univ.-Comput. Inform. Sci.* **25**(2) (2013)
2. American Diabetes Association, Standards of medical care in diabetes–2006. *Diabetes Care.* **29**(Suppl. 1), s4–s42 (2006)
3. D. Castro, Wearable-based human activity recognition using an IoT approach. *J. Sens. Actuator Netw.* **6**(4) (2017)
4. L. Chen, D. Magliano, B. Balkau, S. Stephen, P. Zimmet, A. Tonkin, P. Mitchell, P. Phillips, J. Shaw, AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures, 200. *MJA* **192**(4) (2010)
5. G.D. Kalyankar, R.P. Shivananda, V. D. Nagaraj, Predictive analysis of diabetic patient data using machine learning and Hadoop. I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), in *2017 International Conference on IEEE* (2017).
6. K. Kavitha, R.M. Sarojamma, Monitoring of diabetes with data mining via CART Method. *Int. J. Emerg. Technol. Adv. Eng.* **2**(11), 157–162 (2012)
7. Korean Diabetes Association, Diabetes Fact Sheet in Korea, (2016). Availableonline: [http://www.diabetes.or.kr/temp/KDA\\_fact\\_sheet%202016.pdf](http://www.diabetes.or.kr/temp/KDA_fact_sheet%202016.pdf) (accessed on 1 May 2018)
8. J. Lindstorm, A. Neumann, K.E. Sheppard, A. Gills-Januszewska, C.J. Greaves, U. Hande, Take action to prevent diabetes: the IMAGE Toolkit for the prevention of type 2 Diabetes in Europe. *Hormmetab. Res.* **42**, S37–S55 (2010)
9. S. Mall, G. Mansi, C. Rahul, Diet monitoring and management of diabetic patient using robot assistant based on Internet of Things, in *Emerging Trends in Computing and Communication Technologies (ICETCCT)*, *International Conference on IEEE* (2017)
10. V. Mohan, R. Deepa, M. Deepa, A simplified Indian Diabetes Risk Score for screening undiagnosed diabetic subjects. *J. Assoc. Physicians India* **53**, 759–763 (2005)
11. J.B. Pérez, Proposal of wearable sensor-based system for foot temperature monitoring. in *International Symposium on Distributed Computing and Artificial Intelligence* (Springer, Cham, 2017)
12. S. Poorejbari, N.H. Vahdat, W. Mansoor, Diabetes patients monitoring by cloud computing, in *Cloud Computing Systems and Applications in Healthcare. IGI Global*, 136, pp. 99–116 (2017)
13. A. Vardhan, A. Prabha, M.K. Shashidhar, N. Saxena, S. Gupta, A. Tripa, The value of the indian diabetes risk score as a tool for reducing the risk of diabetes among Indian medical students. *J. Clin. Diagnostic Res.* **5**(4), 718–720 (2011)



14. N. Wang, K. Guixia, A monitoring system for type 2 diabetes mellitus. in *e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on. IEEE (2012)*
15. World Health Organization, *Definition, Diagnosis, and Classification of Diabetes Mellitus and Its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus* (World Health Organization, Geneva, Switzerland, 1999)

# Chapter 33

## Message and Image Encryption Embedding Data to $GF(2^m)$ Elliptic Curve Point for Nodes in Wireless Sensor Networks



G. Leelavathi, K. Shaila, and K. R. Venugopal

### 33.1 Introduction

Wireless sensor networks (WSNs) form an ad hoc network, comprising multifunctional sensor nodes. Security procedures are generally the tasks that consumes most of the overall processing capacity of network devices in WSNs [3]. This work concentrated on the design of efficient hardware architectures for elliptic curve cryptography over binary Galois fields, which is mainly affected by the underlying arithmetic primitives.

#### 33.1.1 Problem Definition

From the literature survey the works discussed are majorly focused on the encryption and decryption of image with software implementation, using different programming languages on processors. Primary field is considered with mapping of message in some implementations. It is required to design a cryptoprocessor with low computational complexity to match the requirements of wireless sensor network nodes.

---

G. Leelavathi (✉)

Visvesvaraya Technological University, Belgaum, Karnataka, India

K. Shaila

VTU-Research Centre, Vivekananda Institute of Technology, Bengaluru, Karnataka, India

K. R. Venugopal

Bangalore University, Bengaluru, Karnataka, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_33](https://doi.org/10.1007/978-3-030-19562-5_33)

329

### 33.1.2 Motivation

Rami et al. [1] proposed a reconfigurable architecture using ECC for mapping and encrypting a message. The message considered is a 163 bit static data with sequential and parallel architecture. The basic idea is modified in our approach to reduce the computational complexity in elliptic curve cryptography. The data considered is a 9 bit hex value with continuous message and also image is encrypted to provide security of the data.

### 33.1.3 Contribution

The proposed work considers both data and image and aims to design cryptoprocessor with hardware implementations. The binary field  $GF(2^m)$  is chosen for design and embedding message on elliptic curve points. The variable text length is considered as input data and also different size images for image inputs. It is a stream oriented approach than static size data input.

## 33.2 Literature Survey

Sensor based and application domains of FPGAs are discussed in [3, 9, 12]. Block oriented key and data is considered in the design with binary field [8]. Mapping of text data to an elliptic curve over the Galois field is described in [7]. George Amalarethinam and Sai Geetha [4] add  $32 \times 48$ , magic rectangle to enhance randomness of the cipher text with RSA cryptosystem using Java. Soleymani et al. [14] introduced a map table technique over finite field  $GF(p)$  to transform an image pixel value to a point on a predefined elliptic curve. The design is implemented and analyzed by MATLAB with Intel Microprocessor.

Gupta and Silakari [5] and Gupta et al. [6] developed diffusion template with 3D standard map and Cat map. Astya et al. [2] implemented a technique for BMP images of different sizes using elliptic curve cryptography in C language. Singh and Singh [13] included digital signature with cipher image to afford authenticity and integrity by grouping the pixel according to the ECC parameters. In this design, pairing of the grouped pixels is mapped to elliptic curve coordinate, instead of mapping of values. Nagaraj et al. [11] and Nagaraj and Raju [10] proposed a magic matrix operations for protecting images and design is implemented using DOT-NET software. They apply the encryption methods only on grayscale images. Zhu et al. [15] scramble the image pixels with watermark, thus increasing the difficulty in decoding.

From the literature survey it is observed that most of the designs are software implementation with different programming languages like Java, C, C++, DOT-NET. This necessitates the hardware design of the proposed cryptosystem.

### 33.3 Model and Computational Details

The points of the elliptic curve ( $E_{a,b}$ ) are computed with Eq. (33.1)

$$(x, y)|y^2 + x * y = x^3 + ax^2 + b. \quad (33.1)$$

Equation (33.2) elliptic curve  $E_{a,b}$  is considered in this work.

$$y^2 + x * y = x^3 + x^2 + 1. \quad (33.2)$$

The additive abelian group elliptic curve  $E_{a,b}$  contains the points  $(x, y) \in GF(2^m)$  and fulfills Eq. (33.2). The points  $(x, y)$  are in affine coordinate system, where  $a$  and  $b$  are equal to 1. Core challenge is to map the input characters of message and image into  $(x, y)$  points on the  $E_{a,b}$  elliptic curve with Eq. (33.2). Finite multiplication and finite addition are used to encrypt the transmitted data using ElGamal method. Computation of  $\gamma$ , with chosen values of  $x_1$  and  $M$  is given in Eqs. (33.3) and (33.4).

$$\gamma = (M^2/x_1^2) + (M/x_1) + x_1 + a + (b^2/x_1^2), \quad (33.3)$$

where  $M, x_1 \in GF(2^m)$ ,  $M$  is the message or image data in hexadecimal value.

The point  $(x_1, M) \in E_{a+\gamma,b}$ , where  $E_{a+\gamma,b}$  denotes the whole points  $(x, y) \in GF(2^m)$  and fulfill Eq. (33.5)

$$y^2 + x * y = x^3 + (1 + \gamma)x^2 + 1 \quad (33.4)$$

$$\lambda^2 + \lambda = \gamma, \quad (33.5)$$

where  $a = 1 + \gamma, b = 1 \in GF(2^m)$ . For any  $\gamma \in GF(2^m)$ , Eq. (33.6) has a solution when,  $Tr(\gamma) = 0$ , the details are originally found in [1, 7] where mapping a message  $M$  to an EC point is explained mathematically. With solution of  $\lambda$ , the isomorphism can be implemented, which is represented as  $E_{a,b} \rightarrow E_{a+\gamma,b}$  and  $E_{a+\gamma,b} \rightarrow E_{a,b}$ . The  $\lambda$  is defined by the equation

$$\lambda = \sum_{i=0}^{n(n-1)/2} \gamma^{2^{2i}}. \quad (33.6)$$

### 33.4 Implementation

#### 33.4.1 Encryption Architecture

The data from the text file and image file is extracted and converted to hexadecimal values. These hexadecimal values are treated as message ( $M$ ) and using the same values the point is generated and mapped. The message ( $M$ ) is used in Eq. (33.3)

to calculate  $(\gamma)$ . Computation of  $P = (x, y) \in E_{a+\gamma, b}$  and mapping  $f : E_{a+\gamma, b} \rightarrow E_{a, b}$  is denoted by  $f(P) = f(x, y) = (x, M + x.\lambda) = (x^*, y^*) \in E_{a, b}$ . The implemented ElGamal public key encryption includes scheming Bob's public key via finite multiplication procedure under  $E_{a, b}$ . The first coordinate of Bob's public key i.e,  $qax$  or  $qay$  is multiplied by Bob's private key  $db$ . The public key is expected to be conveyed to Alice through a secured exchange protocol. The Alice's public key is computed  $(ka * db * x)$ , where  $ka$  is Alice's private random key. Encryption process with elliptic curve is shown in Fig. 33.1. The complete process of encryption is given in Fig. 33.2.

### 33.4.2 Decryption Method

The decryption is performed with Eq. (33.7).

$$[M + Ka * Qax] - [db(Ka * x)]. \tag{33.7}$$

With this equation  $-db(Ka * x)$  is obtained first. If  $P = (x, y)$  is a point on the  $E_{a, b}$ , then  $-P = (x, y + x) \in E_{a, b}$ . XOR operation calculates  $y + x$  of  $db(Ka * x)$  to obtain  $-db(Ka * x)$ . This is followed by a finite addition process under  $E_{a, b}$  between the received cipher message  $M + Ka * Qax$  and the decryption key  $-db(Ka * x)$  to get back the original message data.

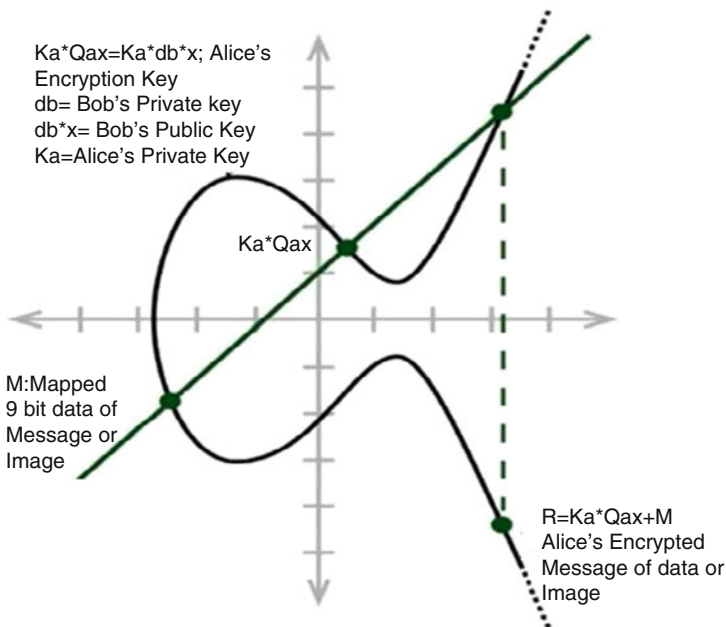


Fig. 33.1 Process of message mapping

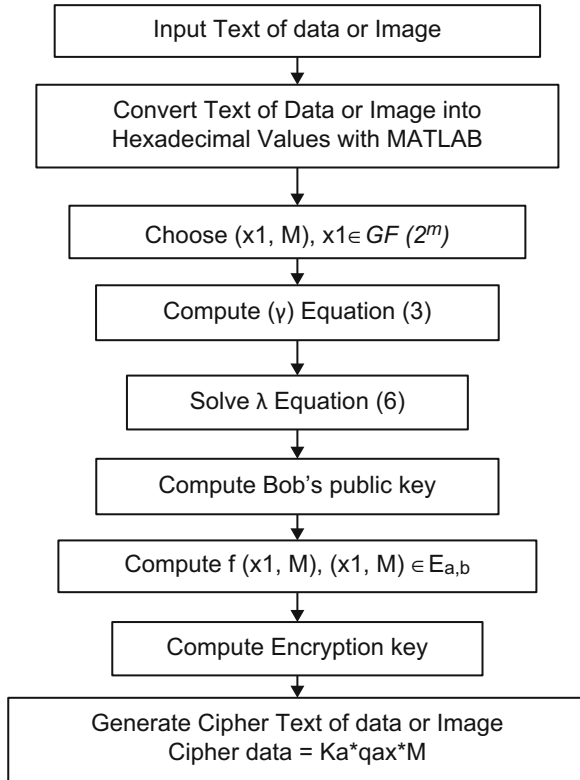


Fig. 33.2 Flowchart for image/message encryption

### 33.5 Results and Discussion

#### 33.5.1 Simulation Setup and Results

The Spartan FPGA device is used in this work for hardware implementation and it is a low cost FPGA. The same design on advanced FPGAs works expectedly faster satisfying all requirements of area and speed. The timing waveform for the image and message is shown in Fig. 33.3. The input values  $x$ ,  $db$ ,  $Ka$ , and  $M$  are of 9 bits. The 32 bit count value depends on the input message or image data. Different images are considered for execution and simulation results are shown in Figs. 33.5, 33.6, and 33.7. The  $M[8:0]$  is the input plain text data,  $y[8:0]$  cipher output, and  $out1[8:0]$  is decrypted output, i.e., original message. The text message with different sizes is considered for simulation as shown in Fig. 33.4 and image with different sizes is reshaped before assigning as input image.

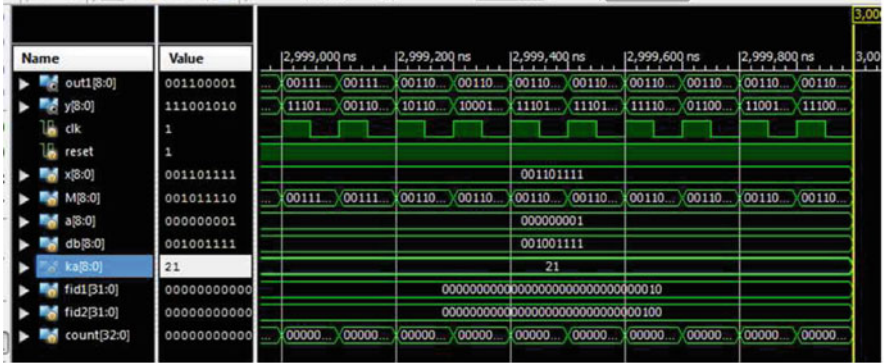


Fig. 33.3 Timing waveform for image input

```
Generated Cipher Text is:  
AQ ASBAPD(2kP AQKBAAL S0E AMZEA UKX0 00 AD 00t 1g 10M; A X00tP 60 +EAO g0 0tA 00k0 xEA ✓  
AMtj0 Ae*0z 0tA 0u0t0 0tAe s 0t0t0 s0M; A 0tA00t0 S0E AMZg* z. xit0tA ✓  
TIANJIA*.....  
Generated PlainText is:  
Do something because you really want to do it. If you're doing it just for the goal and  
don't enjoy the path, then I think you're cheating yourself,  
By Kalpana Chawla
```

Fig. 33.4 Cipher output of message

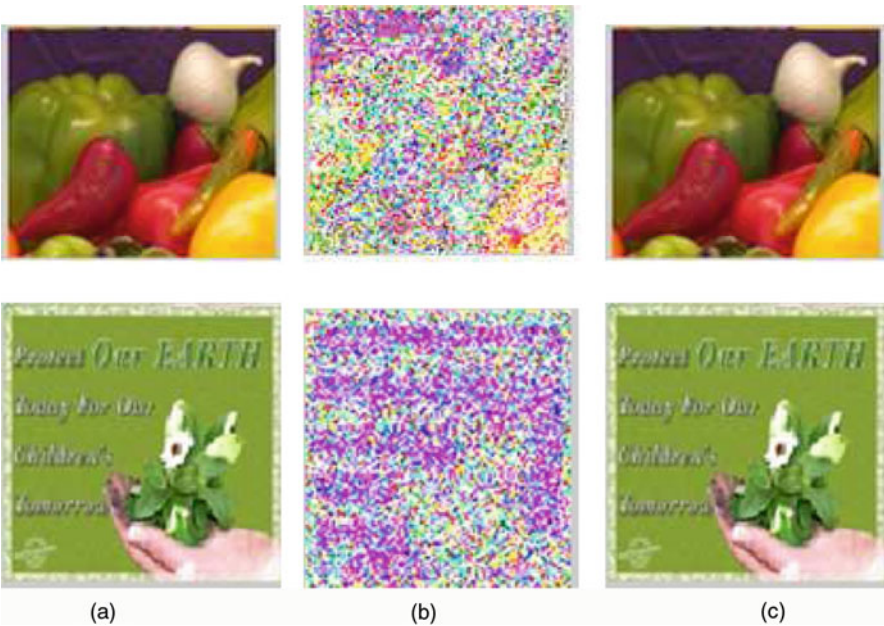
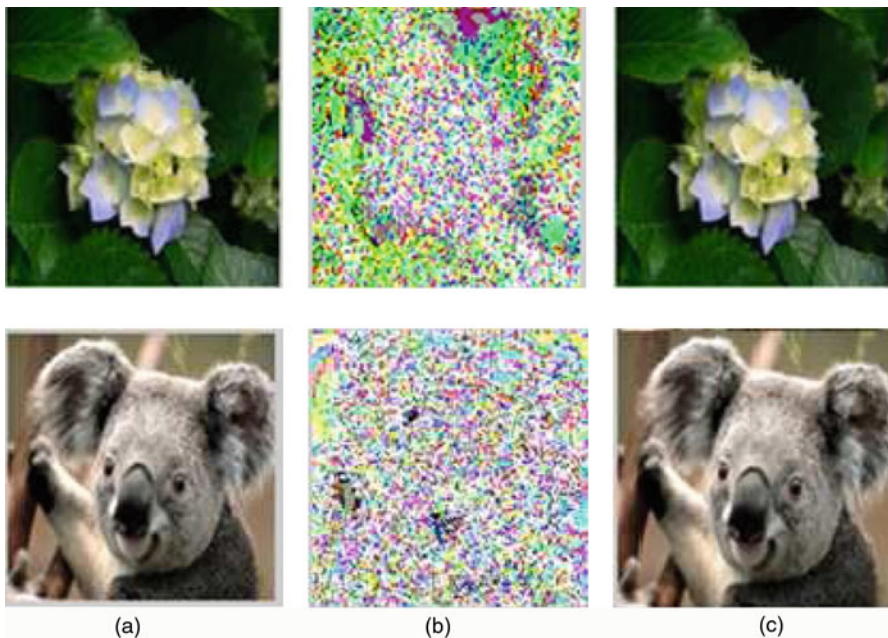


Fig. 33.5 Image—pepper and earth. (a) Original. (b) Encrypted. (c) Decrypted

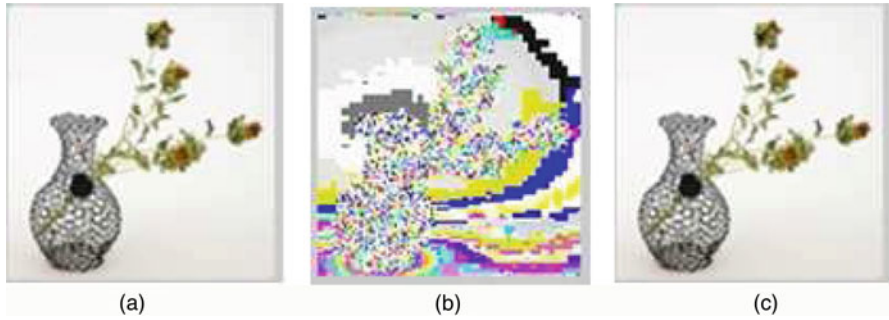
### 33.5.2 Performance Analysis

We have tailored this architecture for cryptographic applications specifically consisting of cost-efficient Xilinx Spartan-3 FPGA xc3s500e-4-fg320. The encryption and decryption time depends on the data and image to be processed. The time required is proportional to the count value, with the message length and image size. In this work a new embedding method is presented to convert text data and image pixel hexadecimal value to a point on a predefined elliptic curve over binary field  $GF(2^m)$  with ElGamal encryption. The output for text messages is shown in Fig. 33.4. The encryption process output is shown in Figs. 33.5, 33.6, and 33.7 for different images. The total encryption and decryption time results are around  $10.09021 \mu\text{s}$  for  $100 \times 100$  images and  $0.029 \mu\text{s}$  for a message. The static power utilization of  $0.034 \text{ W}$  is observed. To evaluate the strength of this algorithm an entropy statistical analysis is performed on plain and encrypted images shown in Table 33.2. The cipher image names are shown in bold. Table 33.1 shows the device utilization details. More number of LUTs is utilized thus reducing the computational delay. It is observed the setup and hold time are at 0 time, which indicates no combinational path delay is found in the design.



**Fig. 33.6** Image—flower and panda. (a) Original. (b) Encrypted. (c) Decrypted





**Fig. 33.7** Image—pot. (a) Original. (b) Encrypted. (c) Decrypted

**Table 33.1** Device utilization

Logic utilization	Used	Available	Percentage of usage
4 input LUTs	203	1920	10
Slices	124	960	12
Bonded IOBs	21	83	21

**Table 33.2** Entropy values of images

Images	Entropy
earthResize	7.2322
<b>earthenc</b>	7.2873
earthdec	7.2322
flowerResize	7.2218
<b>flowerenc</b>	7.6347
flowerdec	7.2218
pandaResize	7.8074
<b>pandaenc</b>	7.3050
pandadec	7.8216
potResize	6.2264
<b>potenc</b>	7.0926
potdec	6.2264

### 33.5.3 Entropy Analysis

Randomness in cipher image designates the proportion of security. A cipher image with high randomness is less vulnerable to the different attacks. To measure the randomness of an image a statistical scalar parameter entropy is used. To show that cipher image has a random texture [14] maximum value is around 8 and it is calculated by Eq. (33.8) and the result is shown in Table 33.2.

$$\text{Entropy} = \sum_{i=0}^n P_i \log_2 P_i. \tag{33.8}$$

When choosing the image, the irregular size of image is resized and given as input for conversion. The plain images with low entropy values are not suitable for this proposed work, as it is observed with encrypted image of Fig. 33.7.

This design is due to dynamic mapping resistant against brute force and side channel attacks. To the best of our knowledge till date, recent work on FPGA is concentrated only for the data of some static size of bits. In this work we have considered stream of input data and the work is also extended to encryption of images with different sizes.

### 33.6 Conclusions

The different level of security with different sizes of input data and image can be obtained to check suitability for WSNs. The static power utilization of 0.034 W is observed. To evaluate the strength of this algorithm an entropy statistical analysis is performed on plain and encrypted images whose values are found around eight to indicate randomness in cipher image. Total encryption and decryption time results are around 10.09021  $\mu$ s for  $100 \times 100$  images and 0.029  $\mu$ s for a message. Computational and combination path delay is not observed in any module design implementation. Further this work can be extended with advanced FPGAs that result in less area utilization and computational time.

The cryptosystem developed exhibits a shield against brute force attacks and thus the cryptosystem primarily focuses on increased level of security. The proposed architecture achieves a substantial reduction in area and time that makes it more appropriate for constrained implementations of cryptographic primitives in ultra-low power devices like WSN nodes.

### References

1. R. Amiri, O. Elkeelany, Concurrent reconfigurable architecture for mapping and encrypting a message in Elliptic Curve Cryptography, in *Proceedings of IEEE SoutheastCon* (2013), pp. 1–6
2. P.N. Astya, B. Singh, D. Chauhan, Image encryption and decryption using elliptic curve cryptography. *Int. J. Adv. Res. Sci. Eng.* **29**(3–10), 198–205 (2014). ISSN:2319-8354(E)
3. A. de la Piedra, A. Braeken, A. Touhafi, Sensor systems based on FPGAs and their applications: a survey. *J. Sens.* **12**, 12235–12264 (2012). <https://doi.org/10.3390/s120912235>
4. D.I. George Amalarethnam, J. Sai Geetha, Image encryption and decryption in public key cryptography based on MR, in *Proceedings of International Conference on Computing and Communications Technologies* (2015), pp. 133–138
5. K. Gupta, S. Silakari, Efficient hybrid image cryptosystem using ECC and chaotic map. *Int. J. Comput. Appl.* **29**(3), 1–13 (2011). ISSN:0975-8887
6. K. Gupta, S. Silakari, R. Gupta, S.A. Khan, An ethical way for image encryption using ECC, in *Proceedings of First International Conference on Computational Intelligence, Communication Systems and Networks* (2009), pp. 342–345

7. B. King, Mapping an arbitrary message to an elliptic curve when defined over  $GF(2^n)$ . *Int. J. Netw. Secur.* **8**(2), 169–176 (2009)
8. G. Leelavathi, K. Shaila, K.R. Venugopal, Implementation of ECC on FPGA using scalable architecture with equal data and key for WSN. *Int. J. Eng. Technol.* **9**(2), 773–796 (2017). ISSN: 2319-8613 (Print), ISSN: 0975-4024 (Online). <https://doi.org/10.21817/ijet/2017/v9i2/170902063>
9. D. Mahto, D.A. Khan, D.K. Yadav, Security analysis of elliptic curve cryptography and RSA, in *Proceedings of the World Congress on Engineering, London*, vol. I (2016), pp 1–4. ISBN: 978-988-19253-0-5, ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)
10. S. Nagaraj, G.S.V.P. Raju, Image security using ECC approach. *Indian J. Sci. Technol.* **8**(26), 1–5 (2015). ISSN: 0974-6846 (Print); ISSN: 0974-5645 (Online). <https://doi.org/10.17485/ijst/2015/v8i26/81185>
11. S. Nagaraj, G.S.V.P. Raju, K.K. Rao, Image encryption using elliptic curve cryptography and matrix, in *Science Direct Elsevier Proceedings of International Conference on Computer, Communication and Convergence* (2015), pp. 276–281
12. J.J. Rodríguez-Andina, M.J. Moure, M.D. Valdes, Features, design tools, and application domains of FPGAs. *IEEE Trans. Ind. Electron.* **54**(4), 1810–1823 (2007). <https://doi.org/10.1109/TIE.2007.898279>
13. L.D. Singh, K.M. Singh, Image encryption using elliptic curve cryptography, in *Eleventh International Multi-Conference on Information Processing-2015, Procedia Computer Science* (2015), pp. 472–481
14. A. Soleymani, M.J. Nordin, Z.M. Ali, Novel public key image encryption based on elliptic curves over prime group field. *J. Image Graph.* **1**(1), 43–49 (2013)
15. G. Zhu, W. Wang, X. Zhang, M. Wang, Digital image encryption algorithm based on pixels, in *IEEE International Conference on Intelligent Computing and Intelligent Systems* (2010), pp. 769–772

# Chapter 34

## Crack Detection in Welded Images: A Comprehensive Survey



L. Mohanasundari and P. Sivakumar

### 34.1 Introduction

Cracks are the surface fissures that develop in a material. Various kinds of cracks such as heat treatment, grinding, stress corrosion and cooling exist in metals. Crack grows and lead threats to component life. It has become essential to identify the crack and eradicate the defects. The process of the crack detection is shown in the Fig. 34.1.

#### 34.1.1 Crack Image Acquisition

It is termed as digital imaging. Photographic images are acquired such as physical scene or any interior structure of an object. It includes processing, storage, compression, etc.

---

L. Mohanasundari (✉)

Department of Electronics and Communication Engineering, Kingston Engineering College, Vellore, Tamil Nadu, India

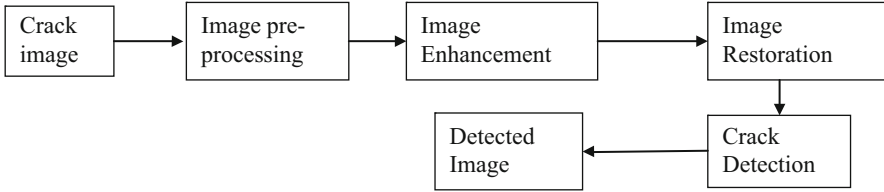
P. Sivakumar

Department of Electronics and Communication Engineering, Dr. N.G.P Institute of Technology, Coimbatore, Tamil Nadu, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_34](https://doi.org/10.1007/978-3-030-19562-5_34)

339



**Fig. 34.1** Block diagram of crack detection using image processing

### ***34.1.2 Preprocessing***

It deals with the images at the level of abstraction. The basic operations in preprocessing is suppressing the unwanted distortions and enhancing the important features of the input images for further processing.

### ***34.1.3 Image Enhancement***

In enhancement, the input image is adjusted for image analysis and for clear display. It can be done either by removing noise or by brightening the image to identify the important key features.

### ***34.1.4 Image Restoration***

It is the process of removing the corruption in the noisy images. The corruption is of any forms such as motion blur, noise mis-focus, etc. Image processing techniques takes place either in time domain or in frequency domain.

The parameters to be identified in the detection process are accuracy, sensitivity and specificity.

### ***34.1.5 Welding***

Welding is a monumental process which joins metals or thermoplastics. Fusion is used where the base metal is not melted. The various types of arc welding are gas tungsten, flux cored, plasma, submerged, shielded arc welding, etc. Various types of welding involves different elements like gas, electricity and laser beams.

### 34.1.5.1 Non-destructive Testing Methods

Non-destructive testing methods are used to detect cracks in metals. It has been widely used in product evaluation, troubleshooting, and research since it does not cause any harm to the product under test. The most popular among the different methods are eddy current, penetrant and magnetic particle testing. These methods are commonly used in various discipline of engineering such as civil engineering, mechanical engineering, system engineering, aeronautical engineering, medicine and art. It has a good impact on medical imaging like radiography, cardiography and ultrasonography. Non-destructive testing methods find application in industries. The failure of a component may lead to major hazard or monetary loss in transportation, hoisting equipment, piping, and building structures.

## 34.2 Literature Survey

The survey provides a detailed description of various techniques used in detecting the cracks in welded images. The performance metrics like accuracy, sensitivity and specificity were analysed using different methods or algorithms.

Sylvie Legendre (2001) [1] has proposed an efficient method for testing the weld. To test the metallic welds, electromagnetic acoustic transducer was used to generate the lamb waves. The wavelet transform and the neural networks are utilized in feature extraction and classification, respectively. The use of these techniques is justified by their integration in specialized processors. Accuracy of about 90% is obtained using these methods.

T. Warren Liao (2003) [2] has proposed pattern classification methods for welding crack detection. They are MLP neural networks, case based reasoning, combination of multiple classifier and MLP NN based attribute weighting. The three major functions in the computer-aided weld quality interpretation system are segmentation, detection and classification of welding flaw types. Preprocessing in increasing noise and contrast, Subtracting the image under test by standard welds and post-processing to combine the detected flaws are the most important steps in the detection of cracks. Using these methodology low false positive rate and high accuracy rate can be obtained.

Yongjoon Cho (2004) [3] has used a pattern recognition method to estimate the quality in resistance spot welding. The weld quality is determined by the primary circuit dynamic resistance. It includes information of the nugget formation. The Hopfield neural network is used to classify the pattern vectors.

Jurg Neuenschwander (2004) [4] has utilized ultrasonic phased-array technology. The quality of the Electron beam welding is controlled by this technology. The general purpose detector, CMS (Compact Muon Solenoid) in which an electron beam welding process is used to reinforce the conductor in the solenoid using alloy profiles. Different types of welding flaws are dealt and the performance metric of ultrasonic assessment system and indicators are obtained.

Asa Prateepasen (2004) [5] has proposed a fuzzy ARTMAP under resistance spot welding to classify the worth of nugget development in DC spot welding process. AE parameters which is extracted online is used. Automatic selection of network structure, convergence properties and online learning are the appealing properties of fuzzy over feed forward neural networks. To identify the quality of nugget formation, Peel and metallographic test plus spatter exploding observation were used. If spatter is used in combination with the AE parameters, the performance can be enhanced.

Yuan Li (2010) [6] has presented a structure light based vision inspection system for monitoring, crack detection and profile measurement. For high quality welding, weld bead inspection is important. The dimensional parameters (groove width, filling depth, etc.) and weld defects are measured and detected respectively during multilayer welding processes. The configuration of the sensor is analysed. Without the requirement of complicated 3D reconstruction of the weldment, the vision inspection system can be easily calibrated. It has been proposed to increase the resolution of the sensor in future.

Dong Du (2010) [7] has dealt with the image registration techniques. Real-time X-ray inspection is used to inspect long consecutive weld. Registration forms basis of spatio-temporal image processing methods, where filtering, blur removing and defect tracking are needed in the image sequences. Here, the image registration processing is separated into rotation registration and translation registration. The rotation registration is implemented by the calculation of the incline position of weld. Phase correlation obtained by extraction the featured from different images is used to implement the translation registration. It is found to be an efficient way to index the X-ray weld image sequences. The slope angle calculation is found to be accurate. The feature extraction is feasible and the image registration is found to be stable and precise.

Wei Wang (2012) [8] has investigated the acoustic emission (AE) measurements to detect the welding crack defects in the specimen which is made of HG70 steel. The structural component is manufactured. Some of the structural component of track crane are outriggers, crane boom, turntable and vehicle frame. Three-point bending test on the standard specimen is conducted. AE source location and AE source characteristics are analysed. For AE source characteristics, procedural map of amplitude, RMS or energy rate, energy accumulation map are used to analyse the welding crack defect. There is a sudden energy leap because of the welding crack. The influence of machine vibration and noise arising due to the engine of track crane is proposed to be done in future.

Wafaa Al-Hameed (2013) [9] has introduced different segmentation methods, known as “data-driven.” Segmentation of the defects is found to play a vital role in classification and detection of the imperfections. The gray level data is used to recognize a region of the image containing a defect. It relies on the thickness and width of the material under test. The morphology process yields better results than the gradient based edge detection methods (Sobel and Canny filters).

Giuseppe Casalino (2013) [10] has analysed various weld imperfections as tolerable defects. They are surplus weld metal, extreme penetration, imperfect filled

groove, deficient penetration, etc. It was based on multiple neural network mapping and fuzzy logic clustering of the process parameters and the tolerable levels for the weld imperfections. Feed forward neural network is used to connect the weld imperfections to the weld parameters. C-means fuzzy clustering algorithm is used. Single metric evaluation criterion was used to provide a better solution. It is best suited for the aeronautical and modern automotive industries.

Jayendrakumar (2014) [11] had proposed a novel approach for weld flaw classification. It includes texture feature extraction techniques and measurement of geometrical feature using Artificial Neural Network (ANN) classifier. To recognize the weld regions and to detect the weld flaws, noise reduction and contrast enhancement techniques are used. Various segmentation techniques (edge base, region growing and watershed) have been applied on images.

S.P. Srivastava (2014) [12] has presented a novel approach for weld flaw classification using Gray level co-occurrence matrix (GLCM). Artificial neural network (cascade-forward back propagation neural network) is used for classification. RGB to gray conversion, region of interest (ROI) selection, noise reduction and contrast enhancement were the image processing techniques used. From each digitized weld images 8, 64, and 44 texture feature vectors have been obtained. 86.1% accuracy is obtained in the classification of radiographic images.

Faiza Mekhalfa (2014) [13] has utilized support vector machine classification in radiographic images. It deals with one versus one methodology in multiclass classification. Different types of imperfections of radiographic images are considered. SVM is an efficient algorithm for automatic classification and high accuracy percent could be achieved using this algorithm. It is faster than multilayer perceptron artificial neural network (MLP-ANN). Large data set of images are proposed to be dealt in future.

Amit Kumar (2014) [14] has analysed the three input parameters (welding voltage, current, and travel speed of welding) and one output parameter (ultimate tensile strength) in MIG welding process. Classification is done using an artificial neural network and optimization by genetic algorithm. Better weld quality is achieved using these optimized weld parameters. These gases may react with the work piece and may cause metal oxides. Hence the work is found to be satisfactory.

Kamran Ali (2015) [15] has presented an automated weld defect recognition framework. Image processing and pattern recognition methods were used on weld radiographs. Undesired distortions are suppressed in preprocessing and image features are enhanced. Support vector machine and artificial neural network are used for the classification. Genetic algorithm is utilized for optimal solution. Defect is accepted as per the image calibration and acceptance criterion. It is proposed to use various feature extraction techniques and feature selection methods for robust and adaptive algorithm.

Changying Dang (2015) [16] has proposed multi-step radiographic image enhancement algorithm (MSRE) and fuzzy enhancement algorithm for segmentation. The enhancement deals the weighting among an original radiographic image and an equalization image. The weighted image is smoothed by an isotropic



diffusion filtering method. Various algorithms were compared and the segmentation accuracy of about 95% is obtained.

Rajesh Ranjan (2015) [17] has presented multivariate method for data analysis. Certain parameters define the quality of the weld. Most of the parameters are to be evaluated to have a complete metallographic analysis to deal with the quality of the weld. It is found that there is no modification is required in the methodology but separate models are required for different wire materials and types. Johannes Gunther (2015) [18] has proposed the use of machine intelligent architecture. It has three machine learning techniques to prevent the defects in laser welding. Low level features are extracted from high level data by auto encoding neural network. The temporal difference algorithm is used to acquire the information regarding the process of laser welding. It has been used in production line in industrial applications.

### **34.3 Comparative Study on Crack Detection in Welded Images**

Table 34.1 compares the different techniques and their salient features in improving the accuracy and reducing the error rate to the minimum. It is found that the combination techniques have improved the performance metric.

### **34.4 Gaps in Literature Survey**

From the above results it has been concluded that the following points were considered to be the problem statement in the literature survey.

1. The accuracy is found to be increased when the combination of techniques are taken into account. But combining any two techniques increases the complexity of the extraction process.
2. Spatter exploding observation has to be used before AE classification to have a dramatically improved performance.
3. The resolution of the vision sensor has to be increased.
4. Various input feature vectors have to be considered in the image acquisition process and algorithms have to be proposed to obtain better sensitivity, specificity and accuracy rate.
5. Different methods could be investigated to get better classification accuracy.
6. Defect segmentation, classification and feature extraction can be made better by various edge assessment techniques.

**Table 34.1** Comparative study on crack detection in welded images

S. No	Authors	Techniques	Features	Parameters (accuracy)
1	Sylvie Legendre, Daniel Massicotte, Jacques Goyette Tapan K. Bose (2001)	Wavelet Transform and Neural network	Ultrasonic NDT method is used to test metallic welds	>92%
2	Warren Liao E. Triantaphyllou (2003)	Combination of MLP NN and CBR	Euclidean distance Hamming distance were used	94%(normalized) 94.5%(normalized) Detection accuracy—97.6%
3	Yongjoon cho Sehun Rhee (2004)	Pattern vector classification	Iteration VS squared error were found	—
4	Asa prateepasen Pakorn kaewtrakulpong Chalermkiat/jirarungsatean (2004)	Fuzzy ARTMAP	Quality of nugget formation in DC microspot welding	85.42%
5	Yuan Li, You Fu Li, Qing Lin Wang, Min Tan (2010)	Visual inspection	Weld bead profile measurement, monitoring and defect detection	—
6	Dong Du, Runshi Hou Jiaxin Shao, Baohua Chang, Li Wang (2010)	Real-time X-ray inspection	Slope angle of weld is found	Proposed method is found to be stable and precise
7	Giuseppe Casalino, Sabina Luisa Campanelli, Fabrizio Memola Capece Minitolo (2013)	Neuro fuzzy model-c means fuzzy clustering	Tolerable defects were analysed	Standard deviation is found

(continued)

Table 34.1 (continued)

S. No	Authors	Techniques	Features	Parameters (accuracy)
8	Jayendra Kumar, R.S. Anand, S.P. Srivastava (2014)	Artificial Neural Networks (Cascade—Forward Back Propagation neural network)	Texture feature extraction Techniques and enhancement techniques were used in the recognition	87.34%
9	Faiza Mekhalifa Nafaa Nacereddine (2014)	SVM (one vs. all) MLP ANN SVM (one vs. one)	SVM is found to be faster than MLP	95.4% 92.64% 96.96%
10	Alireza Azari Moghaddan (2015)	Radiography	Three types of Defects 1. Lack of penetration 2. Incomplete fusion 3. External undercut	98.07% 94.58% 96.28%
11	Kamran Ali Majid Awar Abdul Jalil (2015)	SVM with GA ANN with GA	Radiographic images were used	93.7% 88%
12	Johannes, Patrick, Gerhard, Hao Shen, Klano Diepold (2015)	SVM RBF kernel	Laser welding with deep learning and reinforcement learning were used	Classification Accuracy—82% Predictor Accuracy—93.1%

## **34.5 Types of Welding**

### ***34.5.1 Arc Welding***

It was invented during 1802. Electric current is provided by a flux coated core wire which acts as an electrode. When the metal being welded is brought in contact, an electric arc is created at the junction generating high temperature. The advantage of this type is that shielding gas is not required and found to be effective on rusty metals. The requirement of skilled operator is one of the drawbacks of this type. Arc welding is the best one of all and mostly used in manufacturing and construction industry.

### ***34.5.2 Metal Inert Gas (MIG) Welding***

It was introduced in the year 1960. It uses a gun which is fed with an electrode. It shields the welded metals from environmental factors and makes the process to be continuous and quick. It has good electrode efficiency, has less welding fumes, learning process is easy and requires less heat input. It requires an external shielding gas, not effective on thick materials and the equipment is found to be costly. It can be used with various alloys like aluminium, copper, silicon bronze, magnesium and nickel. It finds application in robotics, plumbing, automotive repairs, construction, fabrication and maritime repairs.

### ***34.5.3 Tungsten Inert Gas (TIG) Welding***

It was introduced in the year 1941. Here without the filler metal, two pieces of metals are heated together to form the weld. High quality welds are produced whereas it requires highly qualified skilled operators. It does not work on rusty materials. The welding process is found to be difficult and time consuming. It has been used in aerospace welding, automobile manufacturing and in high precision welds.

### ***34.5.4 Flux-Cored Arc Welding***

It uses solid wire instead of wire filled flux, whereas the other processes remain the same as ARC welding. With external gas it can be self shielded. It is used in heavy equipment repair and in welding thick materials.

## **34.6 Types of Non-destructive Methods**

### ***34.6.1 Visual Inspection***

It detects macroscopic flaws such as incomplete penetration welds, crater cracking, slag inclusion and likewise. It detects flaws in various types of piping and in composite structures. The other testing includes wrong dimensions, awful welds or joints, deprived fits, omitted fasteners or components, inappropriate surface end, outsized cracks, cavities, delaminations in coatings, insufficient size, and deficient code approval stamps.

### ***34.6.2 Radiography***

The main advantage of radiography is that it provides an enduring reference for the interior soundness of the object. When the materials with more void is exposed to the rays, it penetrates much in the void and the area could be easily recognized. Hence it found to be an effective method but since handling the radioactive materials is hazardous, it has become least importance among the other methodology.

### ***34.6.3 Liquid (Dye) Penetrant Method***

The attracting features of this method are its flexibility and ease of use. It is best suited for surface defect inspection. It is based on the concept on flow of liquid by capillary action. Hence it is found to be an inexpensive and a user friendly technique. It is not been suited for inspecting subsurface flaws. It is used on non-magnetic materials and to detect certain flaws including exhaustion cracks, over load and bang fractures, slake cracks, pin holes in welds, grind cracks, laps seams, deficient of fusion, porosity, etc.

### ***34.6.4 Magnetic Particles***

It uses minute magnetic particles to detect the imperfections. Ferromagnetic materials like cobalt, iron, nickel and their alloys are used to make the elements to be inspected. At the surface to be examined, an electromagnetic yoke is placed. The electromagnet is keyed up when kerosene-iron satisfying deferral is poured on the plane. Magnetic flux will be broken when there is any flaw on the surface and a new north and south pole were created. The iron particles scattered on the surface will be attracted towards the poles. Compared to the actual crack, it is more convenient to

see these clustered particles and this forms the basis for the inspection. To achieve improved sensitivity, the magnetic lines should be vertical to the imperfection.

### ***34.6.5 Eddy Current Testing***

Electromagnetic induction is used in this method to create Eddy currents. The current carrying conductor gets induced by the magnetic field around it. The generated current will be inclined by the temperament of the material used. A probe is placed on the surface to be inspected and Eddy current on work piece is monitored by electronic equipment by placing a probe on the plane that was inspected. The advantages of eddy current inspection are aptness for different applications, sensitivity to minute imperfections, ability to scrutinize composite shapes and sizes of conductive elements, ability to identify planar defects, instantaneous results, transferable equipment, minimum measurement preparation and no need to have the part together under inspection. Some of the limitations of eddy current inspections are importance of an available plane to the probe, limited depth of penetration, skilful and trained personal, hard to detect the flaws which lie parallel to the winding of the coil, possible interference of surface finish and roughness, probe scan direction and necessity for reference standards for setup. It is used for the material width measurements, material identification measurements for conductivity, coating width measurements, case deepness measurements, heat management measurements and heat damage detection.

### ***34.6.6 Acoustic Method***

Acoustic emission and acoustic impact technique are the two kinds of acoustic methods.

#### **34.6.6.1 Acoustic Emission**

It is performed by stressing the part of the structure like bending a beam, applying force to beam, pressurizing the container, measuring the acoustic responses etc. Sensors consisting of piezo electric ceramic elements are used to detect acoustic emissions. This method is effectively used in the observation of various structures.

#### **34.6.6.2 Acoustic Impact Technique**

The surface of an object is tapped by these techniques and the signals are analysed to detect the flaws. This technique is automated and can be instrumented and it has

ease of performance. A reference standard is required for identifying flaws and the results depend on geometry of the parts of the material to be detected.

### **34.6.7 Ultrasonic Inspection**

Measurement is made using the sound energy. It can be used for estimation, substance characterization, feature measurements and evaluation. The functional units of UT inspection system are display devices, transmitter and receiver and various transducers. Electronic devices generate high voltage electrical pulses. Ultrasonic energy is generated by the transducer of various types. Typical couplants like water, grease, oil and glycerine are used to transfer the ultrasonic waves to the test piece. A portion of the ultrasonic wave is interrupted and back propagated by the imperfections in the weld. The amplitude and the time period of the returning wave decide the presence and location of the flaws. It has high sensitivity and penetrating power. It inspects defects in larger parts. The advantages of this method are sensitivity to the planar discontinuities, instantaneous results with electronic equipment, higher depth of diffusion for imperfection detection or measurement, detailed imaging with automated systems, skill to single-sided access for pulse-echo technique, high accuracy in shaping reflector position and examining shape and size, least element preparation, option for other uses such as width measurements. It has certain limitations such as the need for an accessible plane to transmit ultrasound, need for reference standards for equipment calibration, wide knowledge, skill and training, necessity for a coupling medium that transfers the sound energy into test specimen, thickness or not homogeneity, restrictions for roughness, compactness, complexity in inspecting coarse grained elements due to small sound communication and large signal noise, shape indiscretion, need for the proper orientation of the defects with respect to the sound beam and characterization of defects.

## **34.7 Defects in Welds**

Each welding processes has its own characteristics defects. The various welding processes are arc welding, electron beam, resistance, etc. The weld defects in arc welding are misalignment, cluster of porosity, lack of penetration, root undercut, lack of side wall fusion. Electron beam also has the same defects as found in arc welding processes, the surface defects include weld root and weld cap undercut. Due to overheating and inadequate pressure at the joint these defects are encountered. Various other processes are used for joining the metals are diffusion bonding, soldering, brazing and friction welding.

## 34.8 Conclusion

Crack is detected in the welded images using various non-destructive testing methods. The parameters to be found are sensitivity, specificity and accuracy. Certain authors have introduced mean square error rate, standard deviation and analysis based on various iterations. Different techniques have been combined together to improve the performance metric. Based on this survey, a novel algorithm can be identified to increase the detection accuracy and overcome all the gaps found in the literature survey. Recent developments for longitudinal crack detection can be made using continuous Density hidden Markov Model, Kalman filtering and Dynamic programming.

## References

1. S. Legendre, D. Massicotte, J. Goyette, T.K. Bose, Neural classification of lamb wave ultrasonic weld testing signals using wavelet coefficients. *IEEE Trans. Instrum. Meas.* **50**(3), 672–678 (2001)
2. T. Warren Liao, E. Triantaphyllou, P.C. Chang, Detection of welding flaws with MLP neural network and case based reasoning. *Intell. Autom. Soft Comput.* **9**(4), 259–267 (2003)
3. Y. Cho, S. Rhee, Quality estimation of resistance spot welding by using pattern recognition with neural networks. *IEEE Trans. Instrum. Meas.* **53**(2), 330–334 (2004)
4. J. Neuenschwander, B. Blau, R. Christin, T. Lüthi, G. Rössler, Quality monitoring of the electron beam welding of the CMS conductor using ultrasonics. *IEEE Trans. Appl. Supercond.* **14**(2), 526–529 (2004)
5. A. Prateepasen, P. Kaewtrakulpong, C. Jiarrungsatean, Classification of DC microspot welding quality using fuzzy armap on acoustic emission monitoring. *IEEE Trans.* **500**, 649–652
6. Y. Li, Y.F. Li, Q.L. Wang, X. De, M. Tan, Measurement and defect detection of the weld bead based on online vision inspection. *IEEE Trans. Instrum. Meas.* **59**(7), 1841–1849 (2010)
7. D. Du, R. Hou, J. Shao, B. Chang, L. Wang, Registration of real-time X-ray image sequences for weld inspection. *Nondest. Test. Evalu.* **25**(2), 153–159 (2010)
8. W. Wang, H. Wei, Y. Zhen, Y. Dou, H. Heng, Nondestructive evaluation of welding crack defects in structural component of track crane using acoustic emission technique. *Intell. Autom. Soft Comput.* **18**(5), 513–523 (2012)
9. Wafaa Al-Hameed, Yahya Mayali and Phil Picton, Segmentation of radiographic images of weld defect, *J. Global Res. Comput. Sci.*, 4, 7, July (2013)
10. G. Casalino, S.L. Campanelli, F.M.C. Minutolo, Neuro-fuzzy model for the prediction and classification of the fused zone levels of imperfections in Ti6Al4V alloy butt weld. *Adv. Mater. Sci. Eng.* **2013**, 7 (2013)
11. Jayendra Kumar, R.S. Anand, S.P. Srivastava, Flaws classification using ANN for radiographic weld images, 978-1-4799-2866-8/14/\$31.00 ©2014 IEEE
12. Jayendra Kumar, R.S. Anand, S.P. Srivastava, Multi-class welding flaws classification using texture feature for radiographic images, in *2014 International Conference on Advances in Electrical Engineering (ICAEE)* (IEEE, 2014), pp. 1–4
13. Faiza Mekhalifa, Nafaa Nacereddine, Multiclass classification of weld defects in radiographic images based on support vector machines, in *Tenth International Conference on Signal-Image Technology & Internet-Based Systems* (2014)



14. Amit Kumar, R.S. Jadoun, Ankur Singh Bist, Optimization of MIG welding parameters using Artificial Neural Network (ANN) and Genetic Algorithm (GA), *Int. J. Eng. Sci. Res. Technol.*, 614–620 (2014). ISSN: 2277-9655
15. Kamran Ali, Majid Awan, Abdul Jalil, Fiaz Mustansar, Localization and classification of welding defects using genetic algorithm based optimal feature set, in *2015 International Conference on Information and Communication Technologies (ICICT)* (IEEE, 2015), pp. 1–6
16. C. Dang, J. Gao, Z. Wang, F. Chen, Y. Xiao, Multi-step radiographic image enhancement conforming to weld defect segmentation. *IET Image Process.* **9**(11), 943–950 (2015)
17. R. Ranjan, A. Talati, M. Ho, H. Bharmal, V.A. Bavdekar, V. Prasad, P. Mendez, Multivariate data analysis of gas-metal arc welding process, in *The International Federation of Automatic Control* June 7-10 (Whistler, British Columbia, Canada, 2015)
18. J. Gunther, P.M. Pilarski, G. Helfrich, H. Shen, K. Diepold, Intelligent laser welding through representation, prediction, and control learning: an architecture with deep neural networks and reinforcement learning. *Mechatronics* **34**, 1–11. <https://doi.org/10.1016/j.mechatronics.2015.09.004>

# Chapter 35

## An Effective Hybridized Classifier Integrated with Homomorphic Encryption to Enhance Big Data Security



R. Udendhran and M. Balamurgan

### 35.1 Introduction

As a whole, researchers have given significant contribution by proposing algorithms for classification (i.e. here decision tree) and prediction, and they also have proposed different approaches for merging the local decision trees. From the state of the art it has been observed that many of the algorithms are limiting their performance due to they are not good enough with small memory size of RAM (i.e. memory resident), mainly working for a small data size, not domain-free [1], static in nature, less efficient in terms of processing and communication overhead. None of the research has focused on scalable and dynamic classification and prediction process of data mining in distributed environment [2]. At present enough research work is going on such work. The researchers are working for developing scalable and dynamic techniques for classification and prediction which may be able to handle large dataset in distributed environment. In distributed environment, a series of challenges have been emerged in the field of data mining, triggered in different real life applications. The first issue tackled is that it is neither feasible nor desirable for gathering all of the data in a centralized location as because it may need high Internet bandwidth and storage space requirements. For such kind of application domain, it should be advisable and feasible is that to develop the systems for acquiring the knowledge and performing the effective analysis at local sites where the data and other computing resources are present, then transmit the results/models to the needed sites. But this also causes the data privacy and security to share the data of autonomous organizations. In such kind of situations, the knowledge acquisition techniques to be developed which may learn from the statistical summaries and

---

R. Udendhran (✉) · M. Balamurgan

Department of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

© Springer Nature Switzerland AG 2020

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_35](https://doi.org/10.1007/978-3-030-19562-5_35)

353

these can be supplied whenever required. One of the data mining function is said to be known as the classification, which allocate items/instances from the data set in a gathering of data [3]. The main aim of ordering is to exactly forecast the target class from the data in each case. There are lots of applications of classification with data mining. For instance, a classification model can be employed to detect loan applicants. Every Classifier process includes two steps: Classifier Building stage: This method is to build a learning phase. The classifier, which evolved from training certain databases instances/tuples, is constructed by the classification algorithms [4]. Individual instance/tuple that is composed of the preparation group is mentioned as a group. These tuples can also be mentioned to as data points. Usage of Classifier stage—The training model/classifier generated employing the data set will classify test data set objects/tuples. The chief concern is for the preparation of data for classification and prediction. There are certain activities in preparing the data:

- Cleaning the data: This means removal of noise which is made possible by applying the smoothing techniques and treating the missed values which could be replaced by certain commonly occurred values from that attribute.
- Relevance analysis: Database could consist of some unrelated attributes. Correlation analysis will identify any two given attributes are associated.
- Data transformation and lessening: The following methods help in data transformation:
  - Normalization: This transformation occupies scaling every value that will make them descend within a specific range. This method is mainly used in the learning step when the neural networks or the methods evolving measurements are employed.
  - Simplification: This is a major concept of transforming the data. Here we can use the hierarchy concept. Data mining also involves clustering, classification, regression, frequent pattern generation and many more analysis and processing facilities. Thus wider range of data can be processed by classification technique than either regression or correlation. This is the main reason why the popularity of classification is increasing day to day. Data mining is the most important, significant and accurate machine learning [5] application. It allows very large volume of day-to-day data to be processed effectively and to generate useful analysis which may further extend to help the prediction for decision making. There are high chances of mistakes during analysis of large volume of data specially for finding the correlation among the different features of the data sets. Due to the above-mentioned mistake some time it's difficult to find solutions and take decision. These problems can be easily resolved by machine learning which improves the efficiency of the systems. Classification technique is capable of processing/analysing wider range of data for decision-making. There are numerous techniques available such as neural network, Naive Bayesian, support vector machine (SVM), K-nearest neighbour classifier (kNN), instance based learning (IBL), rule based classification and decision tree.

## **35.2 Comparison of Algorithms**

### ***35.2.1 Neural Network***

Neural network is said to be the collection of neurons similar to the human brain system. This neural network stands intermediate among the artificial intelligence (AI) and the approximation algorithm. The neural network is also known as non-linear predictive model since it learns through training resemble structured biological neuron networks. Neural networks are best suited for detecting the pattern, making predictions and it also learns from the past.

### ***35.2.2 Convolutional Neural Network***

Convolutional networks have been used to tackle many practical object recognition applications. For instance, they have been successful applied to recognizing digits and characters in documents and house numbers. Recently, fast computation libraries on GPUs have enabled large-scale training of convolutional networks. Large convolutional networks have been used to produce impressive gains over state of the art computer vision techniques in the ImageNet challenge [6]. CNN consists of layers which can be employed for training. Generally, convolutional layer is formed as the first layer and then an activation layer is included which returns a convolutional and activation block and after employing this block again which constitutes a series of block. An activation layer is a non-linear function which consumes a single number and performs fixed mathematical operation over it. ReLU, Sigmoid and Tanh are employed since we can exchange these functions with addition and multiplication operations. After creation of series of block, an average pooling layer is applied. A pooling layer reduces the size of data. Some well-known pooling layers are average pooling and max pooling. Max pooling is avoided since it does not possess max operation over encrypted data. Instead max pooling, average pooling is employed since it can determine summation of values as well as addition and does not affect the algorithm.

### ***35.2.3 Instance Based Learning***

k-NN does not have training part, it looks for k-nearest neighbours for new sample for the selection of the class to which it belongs. k-NN is also incremental classification in which the algorithm uses indexing concept to finding the neighbours efficiently. The comparison is an exhaustive if it finds the nearest neighbour for the new instance with storing all instances in memory which may also lead to lots of memory usage. The solution of the above problem is the instance based learning

(IBL) which is an enhancement of the k-NN classification algorithm [7]. The k-NN algorithm requires large space while IBL does not need to maintain model abstractions. The k-NN does not more suitable with noisy data, while the IBL supports noisy data and hence it may be applied on lots of real-life datasets.

### 35.2.4 Rule Based Learning

The rule based classification is the systematic selection of a small number of features used for the decision making. It Increases the comprehensibility of the knowledge patterns. The useful if-then rules have been extracted from the dataset on statistical significance. IF-THEN rules can be extracted using the sequential covering algorithm (SCA) such as AQ, CN2 and RIPPER from the training set of data. In this algorithm it is not needed to generate the decision tree (DT). Many of the tuples of the given class are covered by each rule. In this category of the classification the rules are learned all at a time. Every time learning of rules is performed followed by removing the tuple covered by that rule and thus this process continues for all the set of tuples. The path from root to leaf in a decision tree represents a rule [8].

## 35.3 Proposed Classifier

Classification technique has a major trouble in minimizing the feature space dimension by identifying constricted amount of features, which yields to a better performance for classification [9]. Feature selection has undertaken heavy work in these recent years. The time for the pattern classification could be completely reduced by the feature elimination [10]. This also enables the categorizing routine with the privilege of machine learning. At last a list could be created that shows the feature (feature contribution will be more) to be enabled and the features to be eliminated (feature contribution will be less) [11]. The data set  $D$  as a whole separated across different data set  $S_i$  where  $i = 1, 2, 3, \dots, d$ , each data  $S_i$  now process the dataset  $D_i$  to produce the decision tree. The produced dataset utilizes weka tool with the help of J48 algorithm. The training data contains the set of already classified samples  $S = S_1, S_2, \dots, S_n$ . Here each sample  $X_i = X_1, X_2, \dots, X_m$  is nothing but the vector and  $X_1, X_2, \dots, X_m$  represents the features or attributes of the sample. Where  $C_1, C_2, \dots$  etc. represent the classes in which the samples belong. The training data is generally augmented with this vector  $C$ . The most contributing attribute which effectively splits the set of samples is selected at each node of the tree of J48. Using the normalized information gain (i.e. entropy difference) the attribute is selected to split the data. Highest normalized information gain attribute is selected to make the splitting decision. This process will continue and will form the decision tree. It is necessary to convert the decision trees into

decision tables, since the decision tree occupies more space. Due to this procedure, minimum overhead at the network side takes place. The implementation of J48 needs a scan among the whole training set for every nodes of the tree, which shows a split on an attribute.  $O(n)$  is denoted as the amount of nodes in the tree based on the data source, where  $n$  denotes the training instances that makes the time intricacy for the part  $O(n^2)$ . It is difficult and very much complex to merge the different local decision trees to form the global one.

At the left side of each rule the multiple predicates have been formed. From the observations of the decision rules many of the rules are complex, overlapping and somewhat differing with single attribute. In distributed environment the decision rules derived from the local decision trees need to be transferred to merging process. Scanning all the decision rules of different network is somewhat complex and time consuming process. As a whole reducing the network overhead and complexity the decision rules are converted into the decision table form. For the efficient merging process the decision tree rules have been converted into the simple decision rules. Using the J48 parser the decision rules have been derived from the decision tree.

### 35.3.1 Homomorphic Encryption

Homomorphic encryption (HE) method enables operations like addition and multiplication over the cipher text while maintaining its original structure of the message. Generally homomorphic encryption consists of four major functions, namely, generation which is employed for key generation, encryption, decryption and a special function known as evaluation. After the advent of homomorphic encryption, many researchers proposed several other homomorphic encryption methods. But most of these methods encountered certain drawbacks For example, Paillier cryptosystem was able to hold only one addition operation. This kind of homomorphic encryption schemes are known as Somewhat Homomorphic Encryption (SHE). The working principle of homomorphic encryption is to include a small noise which represents a value for encrypting. Hence a small amount of noise is generated in each ciphertext. Therefore, the noise increases when two ciphertexts are added. The decryption process is executed successfully if the amount of noise is less than a threshold. The threshold represents a bound on the number of computations operated encrypted data. In order to reduce the noise first an entity should decrypt and encrypt the ciphertext. However, this is not possible, for many years, the researchers have tried to reduce the noise excluding the secret key [12].

## 35.4 Evaluation

We employed Naïve Bayes as well as decision trees algorithms for comparing the performance of our proposed classifier. Sensitivity, accuracy and specificity are considered as important metrics in determining the best algorithm. We also

		Predicted Class	
		YES	NO
Actual Class	YES	<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
	NO	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

**Fig. 35.1** Confusion matrix

employed confusion matrix of two rows and two columns which is represented by true positives, false positives, false negatives, and true negatives. The confusion matrix leads to a detailed investigation rather than incompetent accuracy. The row in the confusion matrix represents experimental class while expected class is represented by column whereas the cell tallies the number of samples in connection of the matrix. The true positives and true negatives classifications reveal accurate classifications that reside along the diagonal line in the confusion matrix. The model errors represent the remaining fields. We can derive performance metrics from the confusion matrix. We can determine the accuracy by the percentage of the total number of predictions which were correct. The true positive (TP) rate is the percentage of positive cases that were appropriately recognized. The confusion matrix is given Fig. 35.1.

In test stage, we employed threefold cross validation method and average values were calculated. We also compared our performance with KNN algorithm. From Fig. 35.2, it is proven that our proposed classifier achieves accuracy rate of 96 when compared with KNN, SVM, Naïve Bayes algorithms. The best classification performance of our proposed classifier was obtained with eight inputs and its correct classification rate is 97.4%. The correct classification rate of Naïve Bayes (NB) with nine inputs is 95.2% and the correct classification rate of SVM is 95.6%. Therefore, the proposed classifier is best suited for the classification performance with the minimum number of input parameters.

## 35.5 Conclusion and Future Work

The trending technology of big data applications used by wireless sensor network is widely employed in this modern world for a maximum range of application for the best quality of human life but due to their energy constraints and consumption, these wireless sensor network has their limitation in their application and their functionality. Sometimes Big data are gathered from wireless sensor network enabled applications faces main threats which are already prevailing in wireless sensor network. Enhancing big data security with real time intrusion detection will be conferred as future work.

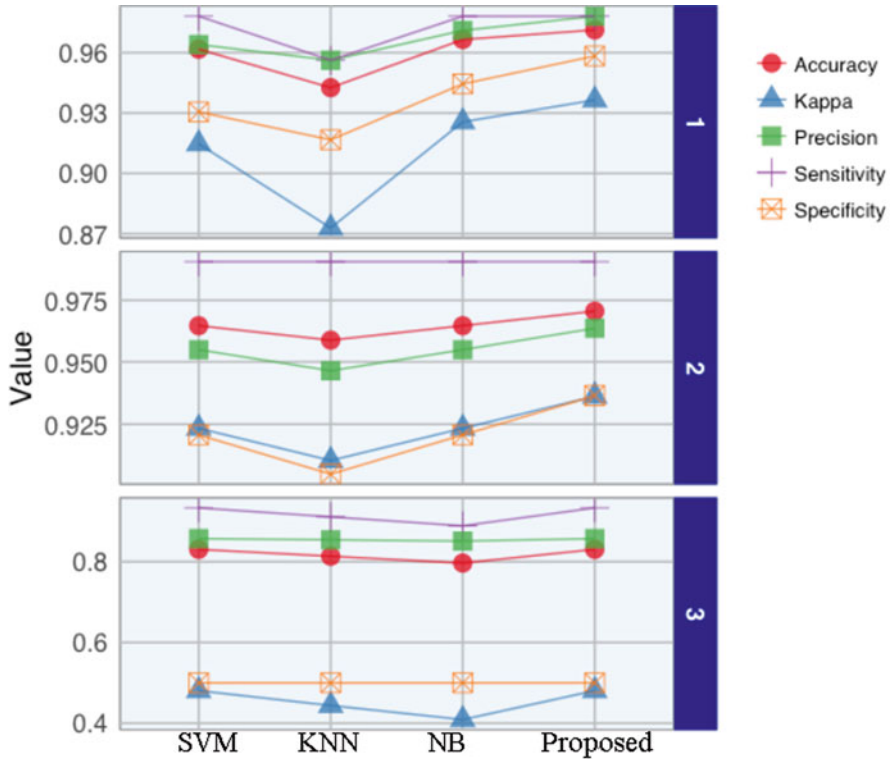


Fig. 35.2 Comparison of algorithms in terms of accuracy

## References

1. M. Fouada, N.E. Oweis, T. Gaberb, M. Ahmedd, V. Snasel, Data mining and fusion techniques for WSNs as a source of the big data, in *International Conference on Communication, Management and Information Technology*, 2015
2. M.S.V. Halde, S.T. Khot, Big data in wireless sensor network: issues & challenges. *Int. J. Adv. Eng. Manag. Sci. (IJAEMS)* 2(9) (2016)
3. H.S. Gol, Integration of wireless sensor network (WSN) and Internet of things (IOT), investigation of its security challenges and risks. *Int. J. Adv. Res. Comp. Sci. Softw. Eng.* 6(1), 37–40 (2016)
4. E.J. Cho, C.S. Hong, S. Lee, S. Jeon, A partially distributed intrusion detection system for wireless sensor networks. *Sensors* 13(12), 15,863–15,879 (2013)
5. C.L. Schuba et al. Analysis of a denial of service attack on TCP, in *1997 IEEE Symposium on Security and Privacy*, pp. 208, May 1997
6. R. Mitchell, R. Chen, A survey of intrusion detection in wireless network applications. *Comput. Commun.* 42, 1–23 (2014)
7. Boon Ping Lim, Safi Uddin. Statistical-based SYN-flooding detection using programmable network processor, in *ICITA '05*, 2005
8. I.F. Akyildiz, E.P. Stuntebeck, Wireless underground sensor networks: research challenges. *Ad Hoc Netw.* 4(6), 669–686 (2006)



9. I. Butun, S.D. Morgera, R. Sankar, A survey of intrusion detection systems in wireless sensor networks. *IEEE Commun. Surv. Tutorials* **16**(1), 266–282 (2014)
10. V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey. *ACM Comput. Surv. (CSUR)* **41**(3), 15 (2009)
11. Q. Mamun, A qualitative comparison of different logical topologies for wireless sensor networks. *Sensors* **12**(11), 14,887–14,913 (2012)
12. R. Udendhran, A hybrid approach to enhance data security in cloud storage, ICC Cambridge, United Kingdom, March 2017 ACM New York, NY, USA ©2017 Computing

# Chapter 36

## AI Powered Analytics App for Visualizing Accident-Prone Areas



Preethi Harris, Rajesh Nambiar, Anand Rajasekharan, and Bhavesh Gupta

### 36.1 Introduction

Accident is an event, occurring suddenly, unexpectedly, and inadvertently under unforeseen circumstances. Each day thousands of road accident victims succumb to injuries and death globally. At present India accounts to 10% of road fatalities which is the highest across the globe [1]. Nearly three-quarter of deaths resulting from motor vehicle crashes occur in developing country [2] and nearly half in the Asia-Pacific region. Road traffic injuries are predicted to rise from ninth place in 2004 to fifth place by 2030 as a contributor to the global burden.

Today the rapid advancement in technology has a profound impact in our life style. AI makes our daily experiences smarter, by embedding predictive intelligence into apps [3]. There is a common intuition that getting this level of analytical power at one's fingertips might be a difficult and time-consuming proposition. In fact, this functionality can be enabled in minutes by simply enabling Einstein in Lightning User Interface (UI) on [Force.com](#) platform [4].

With the advancement of intelligence being embedded in vehicular systems, a lot of work is still in progress pertaining to crash collision based on statistical and real-time data. Some of the notable contribution in literature is summarized as follows:

- In the work of Abdel et al. [5] a comprehensive overview of real-time traffic safety improvement on freeways has been discussed. Loop detector data were used to identify crash-prone conditions on the freeway mainline and ramps. Then Intelligent Transportation Systems (ITS) strategies to reduce the crash risk in real

---

P. Harris (✉) · A. Rajasekharan · B. Gupta  
Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

R. Nambiar  
UiPath, Bengaluru, Karnataka, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_36](https://doi.org/10.1007/978-3-030-19562-5_36)

361

time has been proposed. The results show that, variable speed limit strategies reduced the crash potential under moderate-to-high speed conditions.

- Freeway locations with high crash potential has been examined by Ahmed and Abdel [6] using real-time speed data collected from automatic Vehicle Identification (AVI). Crash data on the expressway network in Orland in 2008 were also collected. The results showed that the possibility of a crash is statistically related to speed data obtained from AVI segments within an average length of 1.5 mile are classified with about 70% accuracy.
- A research was done based on 400 sets of accidents data collected from 10 major roads in Beijing city [7]. Through the statistics of the typical factors and the Logistic regression analysis, the relationships between the traffic accident road type, the vehicle type, driver state, weather and other related data were also studied. The results showed that the location of car in road transects, the road safety grade, the road surface condition, the visual condition, the vehicle condition and the driver state were the most significant factors that primarily lead to traffic accident.
- A complete dataset of hourly aggregated traffic data such as flow, occupancy, mean time speed and percentage of trucks were collected from three random loop detectors in the Attica Tollway located in Greater Athens Area in Greece for the 2008–2011 period [8]. The modeling results showed an adequate statistical fit and revealed a negative relationship between accident occurrence and the natural logarithm of speed in the accident location.
- Abdel et al. [9] examined the relationships between Visibility Related (VR) crash risk and real-time traffic data collected from Loop/radar Detector (LD) data. The model developed based on AVI data, the coefficient of variation in speed observed at the crash segment, at 5–10 min prior to the crash time, affected the likelihood of VR crash occurrence. The results showed that both LDs and AVI systems could be used for safety application.

The literature review reveals that the basic approach for predicting accident-prone areas had been proposed using various regression analysis, statistical approaches, time series method, neural networks, theoretical methods, etc. Such approaches require large samples and obtaining it from government/online databases is usually a tough task. At the outset, predicting the future is also tricky with the increasing pace of technology changes that continue to grow each day. However Salesforce's Einstein promises analytics without any models, algorithms, or cumbersome coding [10]. Hence the proposed work has used the Einstein's AI power to visualize and share crash collision zones through an App.

The next section describes the methodology of the proposed prototype, followed by the results in Sect. 36.3. The final section gives the concluding remarks.

## 36.2 Project Description

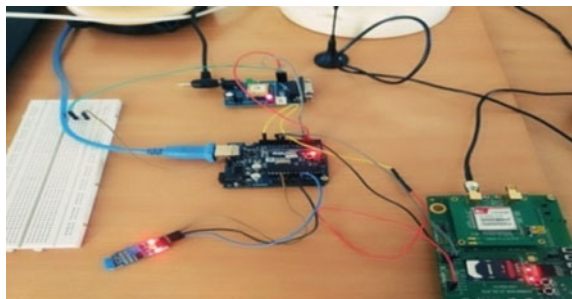
To automate the identification of major accident-prone areas and notification, a low-cost prototype has been developed using Arduino, Piezo vibration sensor, and Wireless Communication technology [11].

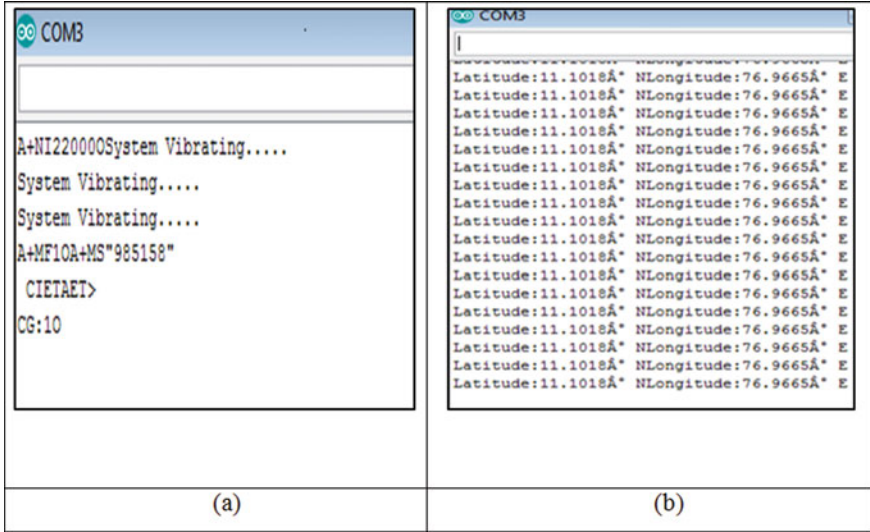
### 36.2.1 Accident Detection Prototype

The experimental setup of the proposed prototype is depicted in Fig. 36.1 where the piezo vibration sensor is the driving force of the experimentation [12, 13]. It comprises three pins, the first pin is connected to the ground, second pin to 5 V power supply, and the third pin to the Arduino. When a thrust is sensed by the sensor, its in-built amplifier transmits the amplified value to the Arduino board. The location is also tracked by the GPS device connected to the Arduino board [14], where the Tx pin of GPS is connected to the fourth pin of Arduino including the power supply and ground pins. The sensor and GPS values are transmitted to the PC via serial port.

When a vibration is induced on the piezoelectric vibration sensor, the analog value generated is converted into digital value by the Arduino microcontroller. Then the vibration values are displayed in the serial monitor as depicted in Fig. 36.2a when the piezo sensor is physically deformed. The threshold value for the vibration sensor is set as 100 units based on the simulation carried out. If the vibration value is less than 100 units it is considered as a “major” accident and “minor” otherwise. Then the GPS module detects the latitude and longitude coordinates of the crash site as depicted in Fig. 36.2b. The sensor values and GPS triggered values are stored in a Salesforce custom object with the latitude and longitude values converted to the names of cities that they represent.

**Fig. 36.1** Experimental setup of accident detection prototype





**Fig. 36.2** Snapshot of the serial monitor output. (a) Detecting vibration; (b) location of the crash

### 36.2.2 *Salesforce Einstein AI*

Einstein Analytics is a self-service application in Salesforce that enables you to explore and provide insights into large amounts of data [10]. The data can be instantly visualized to infer invaluable information through dashboards to continually monitor key metrics based on the latest data uploaded in Salesforce Cloud.

**Connect Data:** The sensor and GPS data collected by simulating major and minor accidents from different locations in and around Coimbatore is stored in a CSV (Comma Separated Values) file. Salesforce has an online dataloader called Dataloader.io [15] which is a web-based application that works on all major browsers. Through the online dataloader, the accident data collected can be imported in Salesforce Org for analytics using the Analytics Studio.

**Explore and Visualize:** The custom object that holds the contents of the CSV file uploaded in Salesforce Cloud is grouped based on cities. Then in the Analytics Studio the app launcher is invoked to create a shared App. This is followed by the creation of a dashboard where the following operations are performed [16]:

- **Aggregation operation:** It determines the number of accidents recorded and loaded in the custom object for further manipulation.
- **Filter:** Then the filter operation is invoked on the dimension- accident type which is a Picklist comprising major or minor accidents. This is followed by tapping the Drill-in filter option to specify the date.

- **Visualization:** Then the stacked bar is used for visualization of the accident data. Notifications are also set for the registered contacts using the widgets in the dashboard [17]. This notification warns registered drivers of the accident-prone regions.
- **Share the story:** After the accident data is explored and visualized, further action of the analytics data is initiated by publishing a snapshot of the dashboard through Chatter, which is a powerful collaboration tool embedded in the App developed. When the registered driver has the Salesforce mobile app downloaded in his device, logs in and taps the Post icon the visualization of the accident-prone region can be viewed along with statistics, that includes annotations also.

### 36.3 Results

The accident detection prototype was subjected to 512 different vibrations at different locations in Coimbatore to simulate the accident scenario. The vibrations values that deformed the piezo sensor were analyzed, based on which the threshold level was set as 100 units. Vibration values generated by sensor which is less than 100 units denoted major accident and minor accident otherwise. During this experimentation the minimum vibration value generated was 78 units and the maximum vibration value was 108 units in the digital representation. The simulation generated 86 major accidents (vibration values falling within the threshold value) and 426 minor accidents (vibration values exceeding the threshold level). For visualization of accident zones at the macro level, the latitude and longitude has been converted into location names they represent in Coimbatore city. Through the Einstein Analytics studio the accident-prone locations are visualized in the App dashboard to thereby caution drivers while passing these areas. Registered contacts are also notified via chatter posts.

The accident details collected from the sensor values were uploaded into Salesforce Cloud for analytics. The steps for analytics in the Salesforce Org are summarized as follows for the simulation carried out:

- **Aggregation:** The count of the rows of accident data is selected as the measure through the operator “#.”
- **Grouping:** The accident data is grouped based on the cities from where the sensor data was collected.
- **Filtering:** The data collected is further narrowed down based on the type of accident. Then the accident details are further drilled down based on start and end dates entered for summarization of the incident.
- **View:** The simulated accident statistics is depicted as a Donut chart as in the left-hand side of Fig. 36.3. It includes both minor and major accidents simulated during the experimentation carried out. The second figure gives details of the major accidents filtered for a particular duration as a timeline chart.



Fig. 36.3 Snapshot of exploration posted in salesforce chatter

- **Share the story:** With the world moving at a fast pace and travelling becoming a daily routine, the snapshot of the proposed exploration can be shared as a post in a matter of a few taps through Chatter option. Figure 36.3 represents posts to the registered drivers warning them of the accident-prone zones through the App developed in Salesforce platform.

### 36.4 Conclusions

Most of the developing nations do not maintain information pertaining to the reported accident statistics base which invariably leads to anecdotes and case studies for crash collision details. In this paper a low-cost prototype has been modeled to collect accident details and thereby notify accident zones to create awareness to drivers. As the entire AI based analytics is integrated into an App, it enables registered contacts to visualize accident-prone areas from anywhere, anytime in

their mobile devices through Salesforce Chatter. This work can be further extended for predicting accident-prone regions using Einstein prediction builder.

## References

1. India ranks first in road deaths in the world, <https://auto.economicstimes.indiatimes.com/news/industry/india-ranks-first-in-roaddeaths-in-the-world/56221070>
2. Road accidents are becoming more deadly in developing countries, <https://www.economist.com/graphic-detail/2017/09/29/roads-are-becoming-more-deadly-in-developing-countries>
3. Salesforce Einstein is artificial intelligence in business technology, <https://www.salesforce.com/in/products/einstein/overview/>
4. Einstein AI drills deeper into Salesforce Clouds, <https://searchsalesforce.techtarget.com/tip/Einstein-AI-drills-deeper-into-Salesforce-clouds>
5. M. Abdel Aty, A. Pande, C. Lee, V. Gayah, C. Dos Santos, Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. *J. Intell. Transp Syst. Technol. Plann Oper* **11**, 107–120 (2007)
6. M. Ahmed, M. Abdel-Aty, The viability of using Automatic Vehicle Identification data for real-time crash predictions. *IEEE Trans Intell Transp Syst* **13**, 459–468 (2012)
7. Tao Lu, Yan Lixin, Zhu Dunyao, Zhang Pan, The traffic accident hotspot prediction: based on the logistic regression method, in *The 3rd International Conference on Transportation Information and Safety*, pp. 107–110, P. R. China, 2015
8. A. Theofilatos, G. Yannis, P. Kopelias, F. Papadimitriou, Predicting road accidents: a rare-events modeling approach. *Transp. Res. Proc.* **14**, 3399–3405 (2016)
9. M.A. Abdel-Aty, H.M. Hassan, M. Ahmed, A.S. Al-Ghamdi, Real-time prediction of visibility related crashes. *Transport. Res. C* **24**, 288–298 (2012)
10. Analytics Basics, [https://trailhead.salesforce.com/modules/wave\\_analytics\\_basics/units/wave\\_start\\_surfing\\_the\\_wave](https://trailhead.salesforce.com/modules/wave_analytics_basics/units/wave_start_surfing_the_wave)
11. S. Amol, J. Monish, W. Akshay, A system for car accident sensing, indication and security. *Int. J. Adv. Res. Comput. Sci. Software Eng.* **5**, 290–294 (2015)
12. Connecting Arduino with vibration sensors, <http://m.youtube.com/result?q=arduino%20connection%20with%20vibration%20sensor&sm=12>
13. How to work with Arduino, <http://m.youtube.com/results?q=how%20to%20work%20with%20arduino>
14. Connecting Arduino with GPS and GSM, <http://m.youtube.com/result?q=arduino%20connection%20with%20vibration%20sensor&sm=12>
15. Import and Export with Data Management Tools, <https://trailhead.salesforce.com/projects/import-and-export-with-data-management-tools>
16. Mobile analytics exploration. [https://trailhead.salesforce.com/en/modules/wave\\_mobile\\_exploration](https://trailhead.salesforce.com/en/modules/wave_mobile_exploration)
17. Analytics Dashboard Navigation. [https://trailhead.salesforce.com/modules/wave\\_exploration\\_dashboard\\_navigation](https://trailhead.salesforce.com/modules/wave_exploration_dashboard_navigation)



**Part III**  
**Workshop on Big Data and Society**

# Chapter 37

## IOT Based Autonomous Inventory Management for Warehouses



A. Madhu Vamsi, P. Deepalakshmi, P. Nagaraj, Akash Awasthi, and Anup Raj

### 37.1 Introduction

Many of the warehouses are being managed manually. The inventories are being sent by the company to the warehouses and they are being recorded manually and after a certain time they will check the number of missing inventories and inform the company to send the missing inventories or if there is any mistake related to inventories (i.e., inventory may change or instead of one inventory they may receive another inventory) [1]. But it is already being more time from being error or mistake. The error is being identified too late. In some industries, these record checking will be done on approximately of one time per 4–6 months or once annually. As even there is a mistake now, there is no use of identifying the mistake as has been committed before 4–5 months. While counting the inventories manually, they have to halt the dispatching and receiving process which results in degradation of the process. So, we are introducing an autonomous device in order to reduce this problem and to increase the efficiency of the warehouses [2].

#### 37.1.1 Related Work

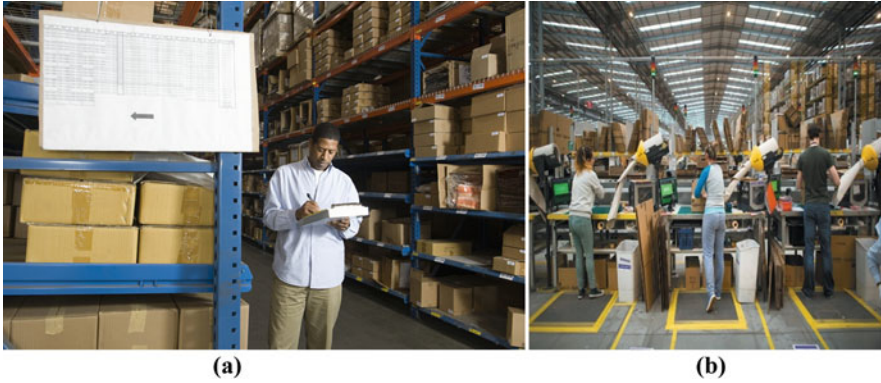
In many warehouses, the inventories are being calculated and managed manually as in Fig. 37.1a which are being imported to warehouses from various companies. These are being stored and counted manually. Even though there are few semiau-

---

A. M. Vamsi (✉) · P. Deepalakshmi · P. Nagaraj · A. Awasthi · A. Raj  
Department of Computer Science and Engineering, School of Computing, Kalasalingam  
Academy of Research and Education, Srivilliputhur, Tamil Nadu, India  
e-mail: [deepa.kumar@klu.ac.in](mailto:deepa.kumar@klu.ac.in)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_37](https://doi.org/10.1007/978-3-030-19562-5_37)

371



**Fig. 37.1** (a, b) Warehouse inventory management in Flipkart



**Fig. 37.2** (a, b) MSM solutions automated inventory management

tomated systems present in warehouses to reduce the human work, still they have scope for full automation as in Fig. 37.1b. It is difficult to keep tracing the entire inventories manually. If there are any errors or any inventories missing, they will inform to the shipping companies [3].

MSM solutions have been developed as an automated inventory management system [4] recently which uses RFID tracking system [5]. They are using portal tracking systems which track the inventory and send the signal to the warehouse management workers. They will find out the inventory and they will scan and they will get the inventory related information and count of inventories (as in Fig. 37.2). All these information is being stored in database. Even though this process is quite good enough there should be human interaction in order to complete this process efficiently [6].

Amazon has been using Unmanned Aerial Vehicles (UAV) [7], for warehouse inventory management [8]. This approach has been quite impressive but there are many restrictions and this approach was not suitable in each and every environment. Also there are many restrictions and rules to be followed to use an UAV but we cannot deploy this approach everywhere [9]. Besides, this approach is highly expensive while compared to other approaches [10]. Camera based (optical) sensors

are also used to scan RFID codes in the range of 10 m and they will store the information in inventory database [11].

In general, all these warehouses will use RFCycle Store by dividing all inventories into three categories as A, B, C. A category inventories will receive more priority than other category inventories (i.e., B and C category inventories) [12]. In this approach all inventories won't get equal priority and requires more human interaction which is tedious work [13].

## 37.2 System Architecture

### 37.2.1 Proposed System

The inventories are being carried out into the warehouses by a means of electronic trolley or other mobility gadgets. These mirrors can be fixed at a particular position where all these inventories are being passed by (such as entry point or check point). These mirrors will scan the inventories being passed in by and they will check with their database and if it is present in database it will forward a signal to the barcode scanner which is movable. It will find out the inventory and it will scan the barcode of the specific inventory and increases the count of the inventory in its database.

- Mirrors equipped with digital camera and movable rod with large field view barcode scanners can read the barcodes of the inventories from top to bottom [14].
- Mirrors fixed at the entry point of the warehouse scan the inventories from top to bottom when it enters in the warehouses and increases the countX variable in database for X kind inventory.
- It will notify the database which automatically check the total inventory count of the particular inventory to see the location of barcode X stored in database.
- If they notice that X inventory from the inventories, then there is a shortage
- If the count of inventory is less than desired, a signal will have generated by signal generator and message will be automatically sent to Shipment Company for upcoming shipments to fulfill the shortage.
- The next date from the shipment company will be scheduled to fulfill the shortage of the inventory.
- After being the data analyzed, user can then approve/decline an automated message to the shipment company for purchasing more inventories.

Here it's all a matter of efficient procedure implementation where the devices are automated and communicated with each other and the user acts only as a supervisor.

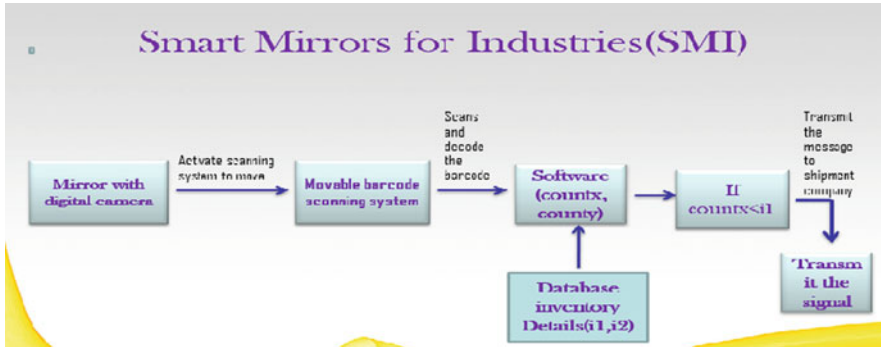


Fig. 37.3 Architecture of proposed system

### 37.2.1.1 System Reliability

We can extend our system to include the model of fog computing in our proposed model to store the data which is sensed by the barcode scanner. If there is lack of inventory, then this signal will be processed by the fog nodes which is available at the fog layer nearer to this sensing layer. It increases the reliability of the system, because if the local system gets failed then data can have transferred to these virtualized fog nodes and it can processed by these fog nodes.

## 37.2.2 Architecture of Proposed System

Image sensor is used to recognize the image when an inventory comes in front of the mirror if the image matches with the vehicle with the inventory then it activate our movable barcode scanner [15], to move from top to bottom and scan the whole inventory from both the sides and it stores the value into the database it compares the total count with stored value if it is less than a message is transmitted to the supplier end [16].

## 37.3 Benefits and Problem Solved

- Our proposed movable barcode scanner reduces the time to scan the barcodes of the inventories automatically when it enters in the warehouse.
- Human labor is reduced by automating scanning process.
- It makes the industry faster and automated.
- It reduces human errors while scanning the barcodes manually.

- Industry can concentrate on the production rather than hiring lot of workers for the inventory management.
- It reduces the effort of counting the inventories while loading into warehouses and updating the databases manually.
- It is of low cost and efficient than present day costly technologies.

## 37.4 Hardware Setups

In our proposed system we have used very high-speed barcode scanner which can scan more than one inventory at a time with kinetic camera with CCD image sensor which uses the image recognition algorithms to recognize the image of the inventory car. If there is a lack of inventory, the signal transmitter will send a signal to the supplier end to schedule more inventory next day [17].

## 37.5 Conclusion

Our proposed system reduces the manual interaction and makes the industry faster. We have used the concept of fog computing which makes the system reliable and speed up the processing. These virtualized fog nodes appear nearer to the sensing and processing unit of the data without any communication charges. We have used the movable barcode scanner which scans the inventory from top to bottom without any human interaction which is better than the existing system.

## References

1. S. Jayanth, M.B. Poorvi, M.P. Sunil, Inventory management system using IOT, in *Proceedings of the First International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing*, ed. by S. Satapathy, V. Prasad, B. Rani, S. Udgata, K. Raju, vol. 507, (Springer, Singapore, 2017), pp. 1–6
2. M.N. Abdelkrim, E. Bajic, A. Zouinkhi, H. Chekir, S. Trab, IoT-based risk monitoring system for safety management in warehouses. *Int. J. Inform. Commun. Technol.* **13**, 424–425 (2018)
3. M. Riad, A. Elgammal, D. Elzanfaly, Efficient management of perishable inventory by utilizing IoT, in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (IEEE, Stuttgart, 2018)
4. E & S diversified Homepage. <http://eandsdiversified.com/>, last accessed 2018/11/21
5. P.R. Wurman, R. D'Andrea, M. Mountz, Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine* **29**(1), 9–10 (2008)
6. Y. Song, D. Han, Exception specification and handling in workflow systems, in *Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications* (Springer, Xian, 2003), pp. 495–506

7. E. Lee, K. Farahmand, Simulation of a base stock inventory management system integrated with transportation strategies of a logistics network, in *Proceedings of the 2010 Winter Simulation Conference* (IEEE, 2010), pp. 1934–1945
8. O. Jukic, I. Hedi, *Inventory management system for water supply network*, MIPRO, Opatija, Croatia, 26–30 May 2014
9. Z. Li, L. Jialing, *Supply chain management*, vol 3 (China Central Radio and TV University Press, Beijing, 2006), pp. 129–132
10. D. Long-gang, Based on RFID, Wi-Fi, Bluetooth, ZigBee of things of electromagnetic compatibility and interference coordination. *Internet Things Technol.* **1**, 59–61 (2011)
11. M.K. Lim, W. Bahr, S.C.H. Leung, RFID in the warehouse: a literature analysis (1995–2010) of its applications, benefits, challenges and future trends. *Int. J. Prod. Res.* **145**, 409–430 (2013)
12. M. Karkkainen, Increasing efficiency in the supply chain for short shelf life goods using RFID tagging. *Int. J. Retail Distrib. Manag.* **31**, 529–536 (2003)
13. Y. She, R. Ehsani, J. Robbins, J. Qwen, J.N. Levia, Applications of small UAV systems for tree and nursery inventory management, in *13th International Conference on Precision Agriculture*, St. Louis, USA, 2016
14. J.H. Ong, A. Sanchez, V. Williams. Multi-UAV system for inventory automation, in *1st Annual RFID Eurasia*, Istanbul, Turkey (IEEE Press, 2007), pp. 1–6
15. A. Buffi, P. Nepa, R. Cioni, SARFID on drone: drone-based UHF-RFID tag localization, in *2017 IEEE International Conference on RFID Technology & Application (RFID-TA)*, Warsaw, Poland (IEEE Press, 2017), pp. 40–44
16. F.J. Valente, A.C. Neto, Intelligent steel inventory tracking with IoT/RFID, in *2017 IEEE International Conference on RFID Technology & Application (RFID-TA)*, Warsaw, Poland (IEEE Press, 2017), pp. 158–163
17. D. Roy, A. Krishnamurthy, S.S. Heragu, C.J. Malmborg, Blocking effects in warehouse systems with autonomous vehicles. *IEEE Trans. Automat. Sci. Eng.* **11**, 439–451 (2014)

# Chapter 38

## Internal Repeats of Human Organs



B. Ramya and E. S. Samundeeswari

### 38.1 Introduction

#### 38.1.1 *Bioinformatics*

Bioinformatics is an interdisciplinary field of science for analyzing and interpreting vast biological data using computational techniques. IRHO database gives a walk-through of the major aspects of bioinformatics such as the development of databases and computationally derived hypothesis. In bioinformatics a large part of proteins are contained in the sequence of repeat that are created by internal duplication and repeatedly relate to structural functional units of proteins. Internal Repeats in proteins are difficult to find in the sequence, and in order to avoid the problem IRHO database is created. More number of repeats is helpful to find the data in families of different biological format. Such mutation in sequence can be useful for structure prediction in bioinformatics.

#### 38.1.2 *Repeats of Sequences*

Sequences are arranged with the set of amino acid. Protein sequences are stored in multiple databases. Each database is having classification among protein sequences. Major role of protein sequences is to find the motif of all the repeat sequences in protein databases. Domain of biological data is to consider all the amino acid for further purposes in order to find the structure analysis of data. In human organs the protein of 99.9% play a same role of all the human body. Every amino acid is having the 2D and 3D view of structure. Main role of cloud computing is repeats are stored

---

B. Ramya (✉) · E. S. Samundeeswari  
Vellalar College for Women, Erode, Tamil Nadu, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_38](https://doi.org/10.1007/978-3-030-19562-5_38)

377



in server for future references. Repeats in amino acid contains all the alphabet for string calculation or mining various tools are implemented. FAIR tool is used to find all the repeated sequences in protein data. Protein Data Bank provides the sequences as wells as the structure. Mining the repeats is used to find similarity and identical repeats of all the sequence of human organs. The study and analysis of protein data have become very important research direction and content in biological domain. Mutation of DNA and Protein sequences A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y and 'A', 'G', 'C' and 'T' is similarity of character are totally different. The fundamental mining can be carried out with the help of amino acid.

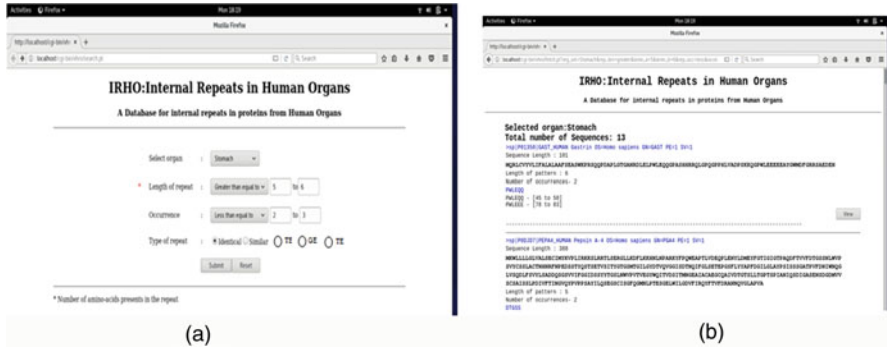
## 38.2 Literature Survey

Sequence predicting the repeats can be gathered as protein repeats, frequently with a length between 12 and 40 residues. Many repeats concentrate on amino acid rates which provide high risk [1]. Repeats can be extracted with certain rates by several algorithm or tools. Predicting carried for more number of molecules of structure and functions [2]. ARM and HEAT can share the similarity of proteins and sequence motif. In DNA sequences there may have the slight changes in length of the repeats. It is difficult for the user to find the structure [3]. The protein kinases are evaluated with the protein sequences. But rarely it reaches the data which it found. Single-amino-acid tandem repeats are very common in mammalian proteins but their function and evolution are still poorly understood. Here we investigate how the variability and prevalence of amino acid repeats are related to the evolutionary constraints operating on the proteins [4]. We find a significant positive correlation between repeat size difference and protein non-synonymous substitution rate in human and mouse orthologous genes. All the common acid individuality of, involving both trinucleotide slippage and nucleotide substitutions, preferentially occurs in proteins subject to low selective constraints [5]. Negative value of repeats has a contradiction in data. Many tools find the alignment of sequences data but it finds all the single repeats only, so it is difficult for many sequences particularly for glutamine, glycine, and alanine repeats. Tools can have the single or multiple repeats but mining takes more time.

## 38.3 Experiment Analysis

Human organs have genes which are classified by tissues. Genes are expressed and separated by their biological function of human body. Analysis of tissue can be evaluated by the function of respective tissues. The main aim of protein sequence is to find the internal repeats of the protein sequences. The 90% of repeated sequence are find out by using FAIR tool analysis and these pattern are separated entirely stored inside the databases for further reference. The data can be used for the unknown amino acid, which can be added in the predicted sequences. Example of protein sequence,





**Fig. 38.2** (a) IRHO tool sequences of protein internal repeats are displayed in webpage by using several queries. (b) Display all the protein sequences repeat with position. If user wants to view in separate window, it can be viewed by clicking view button in web page

## 38.4 Conclusion

A fast and robust toolkit, IRHO, is proposed to identify the major repeat based on tissue analysis sequence available in the UniProt databases. It mainly detects repeats based on criteria and choice chosen by the user for easy access of the data in a defined space interval. Based on organs category user can choose and view the results accurate which is compared to all the other tools kit for multiple repeats can be shown easily.

## References

1. J.C. Wootton, Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285 (1994)
2. J. Jorda, A.V. Kajava, Protein homorepeats sequences, structures, evolution, and functions. *Adv. Protein Chem. Struct. Biol.* **79**, 59–88 (2010)
3. A. Biegert, J. Söding, De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24**, 807–814 (2008)
4. M.A. Andrade, C. Perez-Iratxeta, C.P. Ponting, Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131 (2001)
5. L. Marsella, F. Sirocco, A. Trovato, F. Seno, S.C. Tosatto, REPETITA: Detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics* **25**, i289–i295 (2009)

# Chapter 39

## Bitcoin Prediction and Time Series Analysis



Krishna Chakravarty, Manjusha Pandey, and Siddharth Routaray

### 39.1 Introduction

The most important part of any research is data collection and its analysis. The collected data is summarized and interpreted by statistical and logical methods to identify patterns which can predict relationships or trends. Time series prediction has been in use to predict stable financial markets like the stock market and in-depth research is ongoing in this field. Python & R are the main technologies for the daily data processing chores for today's data scientist.

For controlling the creation of additional units, verifying assets transaction and securing all transactions, cryptography is used and Cryptocurrency, which can be referred as digital or virtual asset work as the medium of exchange. The control is decentralized here in comparison with centralized economy like centralized banking systems. A public transaction database through which the decentralized control works for each cryptocurrency is termed as *Blockchain*, which functions as distributed ledger. Some of the transactional properties are: (1) *Irreversible*: transaction cannot be reversed after confirmation, (2) *Pseudonymous*: real-world identities are not connected to accounts or transactions, (3) *Fast and global*: instantaneous transactions, (4) *Secure*: a public key cryptography system is in place where cryptography funds are locked and (5) *Permission-less*: no permission is required to use cryptocurrency. Adoption of bitcoin as a leading cryptocurrency is growing consistently in the world at present. Bitcoin presents an interesting platform in time series prediction problem as it is still in its transient stage. As a result, the market is highly volatile and thus there is an opportunity in terms of prediction. At present, the factors affecting bitcoin price are (1) bitcoin supply and

---

K. Chakravarty (✉) · M. Pandey · S. Routaray  
KIIT Deemed to be University, Patia, Bhubaneswar, Odisha, India  
e-mail: [manjushafcs@kiit.ac.in](mailto:manjushafcs@kiit.ac.in); [siddharthfcs@kiit.ac.in](mailto:siddharthfcs@kiit.ac.in)

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_39](https://doi.org/10.1007/978-3-030-19562-5_39)

381

increasing/decreasing demand, (2) regulations enforced by governments on bitcoin transactions, (3) bitcoin users and developers influence the rise and fall of price, (4) bitcoin in news, the influence of media on garnering negative and positive publicity and (5) new technological changes to bitcoin. Bitcoin has open nature and it operates on a decentralized, peer-to-peer system which is termed as trust-less as all transactions are irreversible in nature and recorded on an open ledger called blockchain. Transparency of this level is not common in other financial markets. Ethereum (ETH) is a decentralized software platform, which was launched in 2015, helps to create Distributed Applications and smart contracts. These applications can run without downtime and will not be interfered by fraud or any third-party control.

*Cryptocurrency Analysis:* Fundamental analysis implicates financial health evaluation and the chance of survival of a company, which is an essential input for stock investments. This is mainly done by analysing company's financial statements. Good numbers give us confidence that the company has good fundamentals and we tend to invest. But in case of cryptocurrency fundamental analysis, absence of financial statements has made the situation radically different. There are no financial statements because: (1) Financial statements are applicable for corporations whereas cryptocurrencies are actually representations of assets within a network; they are not related to any corporations. Its value directly depends on the community participants like developers, users and miners and not impacted by any revenue generating system.

Through different applications of Blockchain technology, these decentralized cryptocurrencies are manifested. (2) The present time can be considered as infant stage of cryptocurrency which is mainly the development stage and thus limits the real-world use cases. There is a lack of track record and we need a different methodology for fundamental analysis. The current situation and the complex nature of crypto technology demands more work in research field to evaluate the viability and potential of any cryptocurrency. More understanding and investigation in this area will ensure more informed investment decisions. The in-depth knowledge of a currency's fundamentals will also help us to form our own opinion which is not common in the complex crypto world. Creating our own stand will definitely be of unique phenomenon.

While still in its beginning stages, big data analytics is starting to be used to analyse bitcoin and other cryptocurrencies. While many may decry the potential uses of big data analytics for cryptocurrency like identifying users, saying that such uses undermine the spirit of cryptocurrency itself, there are still ways in which big data analytics can legitimately benefit cryptocurrency, such as by identifying fake or dangerous users, preventing theft, and predicting trends. Cryptocurrency analysis is applied in the predicting trends of cryptocurrency prices. Throughout bitcoin's short history, it has been affected many times by world events and overall bitcoin community sentiment. For example, analysts analysed social media after the shutdown of Mt. Gox, once the largest bitcoin exchange, in early 2014. Social media trends were used to identify community sentiment, key voices, and stakeholders, and then tie this information to things like the currency's price performance, which is similar to other financial assets that can be affected by major events.

## 39.2 Literature Survey (Table 39.1)

### 39.3 Proposed Cryptocurrency Prediction Analysis

A series of data points indexed (or listed or graphed) in the order of time is termed as time series. Usually successive equally spaced points in time are recorded to form a time series sequence. Meaningful statistics and other characteristics of time series data are extracted by methods which fall under the umbrella of time series analysis. A model is used to predict future values based on previously recorded values, this process is termed as time series forecasting. The regression analysis which aims at value comparison of a single time series or multiple dependent time series at different points in time cannot be considered as “time series analysis”. For our data we are required to predict the high, low, or close values of the bitcoin. In such

**Table 39.1** Details of literature survey

Author, Year	Title	Purpose
Tian Guo and Nino Antulov-Fantulin February 2018	“Predicting short-term Bitcoin price fluctuations from buy and sell orders”	Analysis of short-term fluctuation of bitcoin market mixture model was proposed to capture the dynamic effect of order book features and to provide interpretable results
Jeffrey Chu, Saralees Nadarajah, and Stephen Chan July, 2015	“Statistical analysis of the exchange rate of bitcoin”	The log-returns of the exchange rate of bitcoin as compared to the United States Dollar have been statistically analysed
Evita Stenqvist and Jacob Lonno, 2017	“Predicting bitcoin price fluctuation with twitter sentiment analysis”	Twitter data related to bitcoin is studied and sentiment analysis is performed. This study shows how twitter analysis can help in prediction of probable future variation of bitcoin price
Abhyudit Bisht, Puru Agarwal, 2017	“Analysis of bitcoin using linear regression and data mining techniques”	Cryptocurrency historical data is used with the help of data mining techniques and regression algorithms to predict different attributed like volume, market cap, etc.
Nakamoto, S Google Scholar, 2008	“Bitcoin: a peer-to-peer electronic cash system”	Instead of trust, the need for an cryptographic proof based electronic payment system is investigated. This allows willing parties to transact directly without any trusted third party interference

cases there is no independent variable like in multivariate systems where we have a dependent variable and an independent variable. It's just the values that come up for particular attributes based on the dates and other upcoming dates. Our data is a data properly curated with all the attributes for a particular day. Moreover, our data is consistent and every record is for a particular day which is constant, hence it is a time series based data. So for such kind of problems we either prefer time series based algorithms such as ARIMA or MA or AR or ARMA and deep learning based models such as RNN and LSTM. Regression and tree-based algorithms are perfectly supervised learning where we have an output variable against every record or feature set. But our dataset comes under semi-supervised learning where we particularly don't have any output variable rather we have a certain trend on particular features which is dependent on time. Hence the liner regression model's basic assumption that observations are independent is true in this scenario. Most time series show increasing or decreasing seasonality trends which means for a particular time frame, variations can be observed. For example, if we observe the sales of winter jackets over time, data will show higher sales value in winter seasons because of obvious reasons.

Essentially when we model a time series we decompose the series into four components: **trend, seasonal, cyclical, and random**. The random component is called the residual or error. It is simply the difference between our predicted value(s) and the observed value(s). Serial correlation is when the residuals (errors) of our TS models are correlated with each other. In layman's terms, ignoring autocorrelation means our model predictions will be bunk, and we're likely to draw incorrect conclusions about the impact of the independent variables in our model.

Terms relating to serial correlation:

*Expectation:* The expected value  $E(x)$  of a random variable  $x$  is its mean average value in the population. We denote the expectation of  $x$  by  $\mu$ , such that  $E(x) = \mu$ .

*Variance:* The variance of a random variable is the expectation of the squared deviations of the variable from the mean, denoted by  $\sigma^2(x) = E[(x - \mu)^2]$ .

*Standard deviation:* The standard deviation of a random variable  $x$ ,  $\sigma(x)$ , is the square root of the variance of  $x$ .

Co-variance tells us how linearly related any two variables are. The co-variance of two random variables  $x$  and  $y$ , each having respective expectations  $\mu_x$  and  $\mu_y$ , is given by  $\sigma(x,y) = E[(x - \mu_x)(y - \mu_y)]$ . Co-variance tells us how two variables move together.

Correlation—It is a dimensionless measure of how two variables vary together, or “co-vary”. In essence, it is the co-variance of two random variables normalized by their respective spreads. The (population) correlation between two variables is often denoted by

$$\rho(x, y) : \rho(x, y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\sigma(x, y)}{\sigma_x \sigma_y}$$

If a TS's statistical properties such as variance and mean do not vary over time, the TS can be referred as a stationary time series. Very strict parameters are used to define stationarity. To assume a TS to be stationary in practical situations, we can consider some factors like (1) variance and (2) mean are constant over time, and (3) contrivance is independent of time. The dataset can be observed to show overall increasing trend along with some seasonal variation. But it is not always possible to infer from the visual presentations. Hence we will be using the following statistical methods.

**Plotting Rolling Statistics:** The moving average or moving variance can be plotted and can be observed if it varies with time. 'Moving average/variance' means that we'll take the average/variance of the last year (last 12 months) at any instant time ' $t$ '.

**Dickey-Fuller Test:** Time series stationarity can be checked by this test. TS is considered as non-stationary as part of the null hypothesis here. A test Statistic and some critical values for difference confidence levels can be furnished as test results. If the 'Critical Value' is greater than the 'Test Statistic', the null hypothesis can be rejected and the time series can be called as stationary.

In real-world situations almost no time series can be considered as stationary. Therefore we can take help of statistics to make a series perfectly stationary. Although this job is almost impossible in reality but we can try to make it as close as possible. Two factors are contributing to the non-stationarity nature of a TS: (1) *Trend*: over a period of time, the mean varies. For example, population in an area is growing over time. (2) *Seasonality*: variations according to specific time frames, e.g. people may buy cars or home appliances in a particular month because of festival deals or pay increment.

First method to estimate and eliminate trend is to reduce trend through transformation. For example, for the significant positive trend a transformation can be applied which penalize higher values more than smaller values. A log, square root, cube root, etc. can help here. To make the TS stationary, different statistical techniques work well where we have to forecast a time series. We can create a model on TS through this most used technique. Once the exercise is completed, it is relatively easy to add noise and seasonality back into predicted residuals in this situation. In these estimation techniques to test trend and seasonality, we can observe two cases: (1) A rare case where the values are independent to each other and we can term this as a strictly stationary series. (2) When the values are significantly dependent on each other. In this case ARIMA, a statistical model, is used to forecast the data.

The full form of ARIMA is Auto-Regressive Integrated Moving Averages. A linear equation (e.g. linear regression) is used to forecast the time series in this model. The parameters of this model are  $(p, d, q)$ : (1) Number of AR (Auto-Regressive) terms ( $p$ ): this is the number of lag observations present in the model, also called the lag order. For example, if  $p$  is 6, the predictors for  $x(t)$  will be  $x(t - 1) \dots x(t - 6)$ . (2) Number of MA (Moving Average) terms ( $q$ ): this is the



size of the moving average window, also called the order of moving average. These are lagged forecast errors in prediction equation. For example if  $q$  is 6, the predictors for  $x(t)$  will be  $e(t-1) \dots e(t-6)$ . (3) Number of Differences ( $d$ ): number of non-seasonal differences are referred here, also called the degree of differencing. We can pass this variable and put  $d = 0$  or pass the original one and make  $d = 1$ . Both the cases yield same results.

Our dataset is split into train and test data beforehand only, i.e. we store our train dataset as bitcoin\_train.csv file and test dataset as bitcoin\_test.csv file. The model trains itself then tests itself on the test data (Table 39.2).

**Table 39.2** Pseudo-code used in ARIMA prediction

---

**Algorithm:** Predictions using ARIMA (Auto-Regressive Integrated Moving Averages) Mod

---

**Preprocessing:** Import( statsmodels.tsa.arima\_model)

**Input:** Ds= Data Set (bitcoin\_test.csv)  
 Dst= Training Data Set  
 Dss=Testing Data Set

**Step1:** *Work on training data, grouping by time series, remove trend and seasonality*  
 1.1 dataset['Close'].plot() #Plotting the Yearwise trend of dataset  
 1.2 dataset = dataset['Close'] ##As here we want to predict the Closing Value of the bitcoin for every day  
 1.3 #Rounding up of data to a weekly, monthly, yearly, quarterly basis  
 weekly=dataset.resample('W').sum(). weekly.plot()  
 monthly=dataset.resample('M').sum(). monthly.plot()  
 year=dataset.resample('Y').mean(). year.plot()  
 quarter=dataset.resample('Q').mean(). quarter.plot()  
 1.4 #Perform Dickey-Fuller test:  
 dfctest = adfuller(timeseries, autolag='AIC')  
 dfcoutput = pd.Series(dfctest[0:4], index=['Test Statistic', 'p-value', '#Lags Used', 'Number of Observations Used'])  
 1.5 #remove trend and seasonality  
 transform\_dataset\_log\_t = np.log(transform\_dataset)  
 seasonality expwighted\_avg =  
 transform\_dataset\_log\_t.ewm(halflife=7,min\_periods=0,adjust=True,ignore\_na=False)  
 .mean()

**Step2:** #Fit to ARIMA model  
 model = ARIMA(transform\_dataset\_log\_t, order=(8, 1, 18)) results\_ARIMA = model.fit(dispatch=-1)

**Step3:** #Prediction and forecast  
 predictions\_ARIMA\_diff=pd.Series(results\_ARIMA.fittedvalues, copy=True)  
 predictions\_ARIMA\_diff\_cumsum = predictions\_ARIMA\_diff.cumsum()  
 predictions\_ARIMA\_log = pd.Series(transform\_dataset\_log\_t.iloc[0]  
 forecast = pd.Series(results\_ARIMA.forecast(steps=7)[0],dates)  
 forecast = np.exp(forecast)

**Step4:** Output  
 print(predictions\_ARIMA\_diff\_cumsum.head()).print(forecast)  
 #Plotting the Actual and The PRedicted With the RMSE  
 plt.plot(forecast, 'Predicted rates')  
 plt.plot(test, 'Observed from test data')

---

### 39.4 Analysis and Result: Time Series Analysis

This section presents the results of time series analysis for bitcoin using python and the result generated for predicted values of the cryptocurrency were very much similar to the original values as depicted by the graphs below. The following Fig. 39.1 represents a graph of weekly, monthly, yearly and quarterly variation of bitcoin values which presents a steep increase between 2017 and 2018. The steep identified symbolizes the acceptance of bitcoin as preferred cryptocurrency by many people throughout the world.

The below code snippet represents logic used for checking the stationary characteristics of data. The characteristics are checked with the help of dickey fuller test that considers rolling mean and rolling standard deviation (Fig. 39.2).

The following figures present the code snippet and the results achieved after removing the trend and seasonalities using log transformations. This is done to make our data set stationary for further experiments (Figs. 39.3 and 39.4).

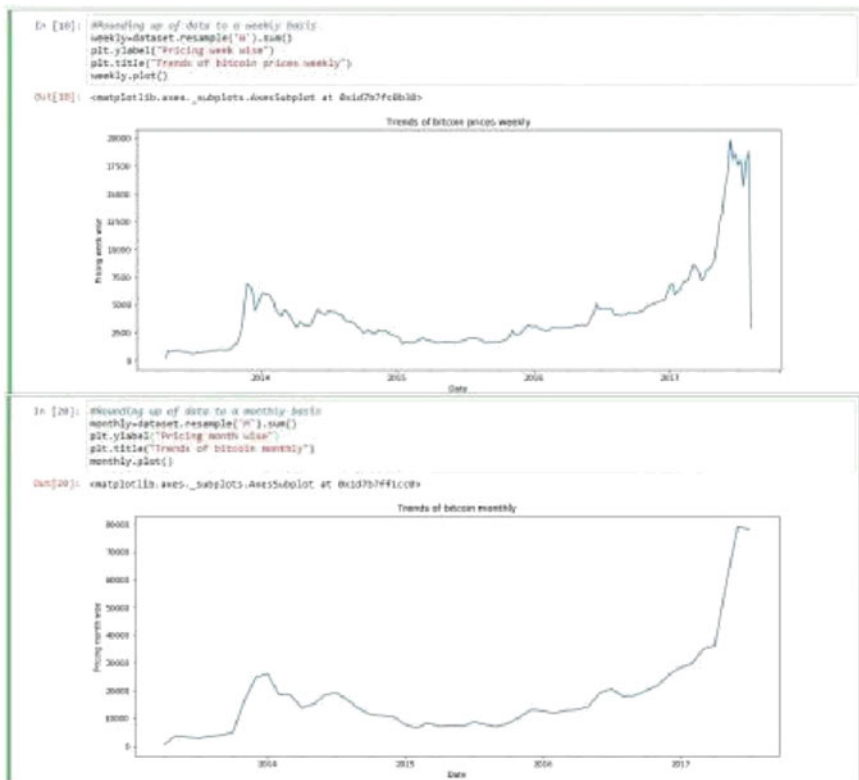


Fig. 39.1 Plotting the graph on a weekly, monthly, yearly and quarterly

```
In [14]: def test_stationarity(timeseries):  
    #Determining rolling statistics  
    rolmean = pd.rolling_mean(timeseries, window=12)  
    rolstd = pd.rolling_std(timeseries, window=12)  
  
    #Plot rolling statistics:  
    orig = plt.plot(timeseries, color='blue',label='Original')  
    mean = plt.plot(rolmean, color='red', label='Rolling Mean')  
    std = plt.plot(rolstd, color='black', label = 'Rolling Std')  
    plt.legend(loc='best')  
    plt.title('Rolling Mean & Standard Deviation')  
    plt.show(block=False)  
  
    #Perform Dickey-Fuller test:  
    print ('Results of Dickey-Fuller Test:')  
    dftest = adfuller(timeseries, autolag='AIC')  
    @output = pd.Series(dftest[0:4], index=['Test Statistic','p-value','lags used','Number of Observations used'])  
    for key,value in dftest[0:4].items():  
        @output[key] = dftest[key] + value  
    print (@output)  
  
In [10]: #Dickey Fuller Test on The Data  
test_stationarity(transform_dataset)
```

Fig. 39.2 Checking stationarity of data by dickey fuller test for rolling mean, standard deviation

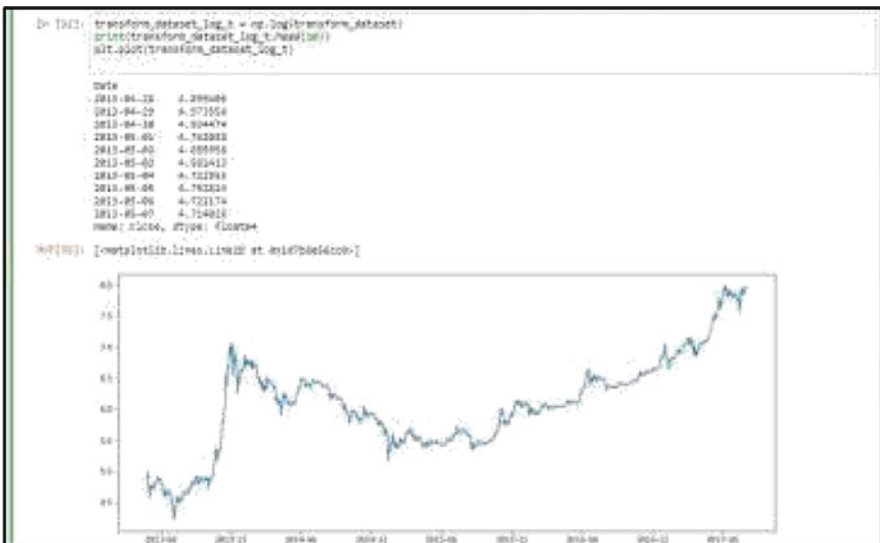


Fig. 39.3 Removing the trends and seasonality by performing log transformation to make our dataset stationary

The following figure is the most important result generated throughout our experiments for Time Series Analysis of change in values for bitcoin using ARIMA model. The ARIMA model is one of the most preferred models for analysis of dynamically changing datasets as of cryptocurrencies. The result thus obtained was used for further forecasting for open and closed values of bitcoin for coming days which was then compared with the actual values. The results thus obtained were encouraging and presented an accuracy of 99% (Figs. 39.5 and 39.6).



Fig. 39.4 Performing exponential weighted to remove trends and seasonality which still exist

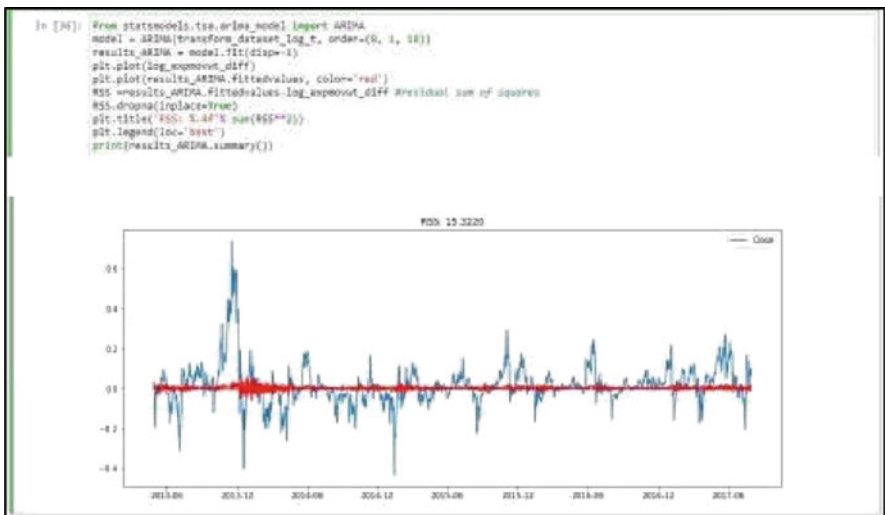
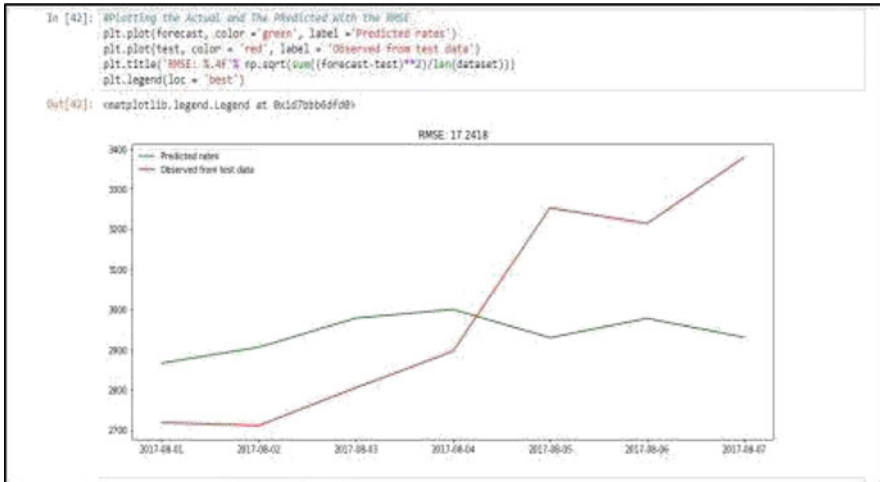


Fig. 39.5 Using the ARIMA model as our time series based algorithm

### 39.5 Conclusion and Future Work

Prediction of bitcoin rates is an important aspect in today’s competitive world of market analysis. We had tried to visualize our analysis report through various graphs and also made conclusions about the fluctuations of bitcoin over time of 28th April 2013 to 1st August 2017. We have also made a detailed analysis and predictions of the future close value of bitcoin with respect to other factors using two machine



**Fig. 39.6** Graph depicting the actual and the predicted

learning algorithms, i.e. Linear Regression and Polynomial Regression. Since the accuracy of Linear Regression was higher than Polynomial Regression we prefer the former model. Using time series analysis, we plotted the variations of close values over weekly, monthly, quarter-yearly and yearly basis.

Through the predictive analytics, graphical visualizations and machine learning, we could estimate the trends of bitcoin.

We would like to implement complex algorithms like RNN and CNN of deep learning. We would also try to look into methods or modify our existing algorithms which can give more accuracy percentage for prediction. We would also try to make a complete graphical user interface for administrator to view analysis and predict more efficiently by entering a specific date.

## References

1. Muhammad Amjad, Devavrat Shah, Trading bitcoin and online time series prediction, in *NIPS 2016 Time Series Workshop* (2017)
2. T.G. Andersen, T. Bollerslev, F.X. Diebold, P. Labys, Modeling and forecasting realized volatility. *Econometrica* **71**, 2 (2003)
3. F. Black, M. Scholes, The pricing of options and corporate liabilities. *J. Polit. Econ.* **81**, 3 (1973)
4. W. Bolt, On the value of virtual currencies. *SSRN Electronic J.* (2016)
5. S. Brahim-Belhouari, A. Bermak, Gaussian process for non-stationary time series prediction. *Comput. Stat. Data Anal.* **47**, 4 (2004)
6. Tianqi Chen, Carlos Guestrin, Xgboost: A scalable tree boosting system, in *SIGKDD (ACM, 2016)*

7. D.L.K. Chuen, *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data* (Academic Press, 2015)
8. J. Civitarese, Volatility and correlation-based systemic risk measures in the US market. *Physica A* **459**, 55–67 (2016)
9. Jonathan Donier, Jean-Philippe Bouchaud, Why do markets crash? Bitcoin data offers unprecedented insights (2015)
10. H.N. Duong, P.S. Kalev, C. Krishnamurti, Order aggressiveness of institutional and individual investors. *Pacific-Basin Finance J.* **17**, 5 (2009)
11. S. Nakamoto, *Bitcoin: A peer-to-peer electronic cash system*, in *Google Scholar*, (2008)
12. F. Mosteller, J.W. Tukey, *Data analysis and regression: A second course in statistics* (Addison-Wesley, Reading)
13. A. Gelman, J. Hill, *Data analysis using regression and multilevel/hierarchical models* (Cambridge University Press, Cambridge)
14. <https://www.coindesk.com/information/understanding-bitcoin-price-charts/>
15. <https://www.cs.waikato.ac.nz/ml/weka/>
16. Kenshi Itaoka, Regression and interpretation low R-squared
17. S.L. Nelson, E.C. Nelson, *Excel-data-analysis-for-dummies* (Wiley, Weinheim)

# Chapter 40

## Smart Active Helmet



W. Gracy Theresa and A. Gayathri

### 40.1 Introduction

IoT, the rapidly growing technology, have gained its scenario in modern wireless communication for sensing, networking and robotics. Its main idea is just connecting devices to the internet. It connects various sensors and interfaces, utility and industrial components, vehicles, etc. with data analytics capabilities [2]. The key features of IoT include artificial intelligence, connectivity, small devices, sensors and active engagement. IoT has its own applications from industry to markets, especially in smart devices, smart technologies and smart cities [15]. Smart device is one among them, which has some level of automation and local computing with user interface for specific use. They are fast and efficient in doing what they expected to do [7]. The main theme of this proposed work is to ensure the lifetime safety during the travel in the vehicle. Mostly, the recent survey by the Times of India on April 2018 forecasted 92% of roadside accident is due to drink and drive. The main theme of this research work is to design a special device named “Smart Active Helmet” for roadside safety from accidents. Keeping this in mind a smart helmet has been designed which mainly keeps the motor cyclists safe in the event of an accident [6]. The head and the brain are the most vulnerable to injury in a motor cycle accident. To avoid this situation drivers wearing helmet increases the chance of survival. Mostly accidents occur due to drunken driven, so there is need to find the alcohol testing for the motorists who consumes alcohol [5]. Thereby our proposed system in smart active helmet is embedded with a fuel cell sensor at the mouth of the helmet.

---

W. G. Theresa  
Department of Computer Science, Loyola Institute of Technology, Chennai, Tamil Nadu, India

A. Gayathri (✉)  
School of Information and Technology Information Science and Technology, VIT, Vellore, Tamil Nadu, India

© Springer Nature Switzerland AG 2020  
A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, [https://doi.org/10.1007/978-3-030-19562-5\\_40](https://doi.org/10.1007/978-3-030-19562-5_40)

393

This sensor has the capability to determine the suspected victim by measuring the blood alcohol level. Now through IOT the sensor in helmet communicates to the ignition interlock device (IID) in motor to lock the motor from further operation. So by using smart active helmet we can avoid the cause of accidents and provide a safe zone for the motorist.

## **40.2 Breathe Alcohol Testing Equipment**

### **40.2.1 Drunkometer**

It is the first roadside practical breathe testing device, which collects the breathe sample of the motorist. The collected sample is passed into a balloon inside a machine where it is pumped through an acidified potassium permanganate solution [19]. The presence of alcohol will change the solution color, the greater the color change, the more the alcohol present in the sample.

### **40.2.2 Breathalyser**

It is a device to check the blood alcohol content from the breathe sample. It was invented at 1927, where the samples of breathe are moved through chemicals in water that would change the color [22]. Later this device uses chemical oxidation and photometry to determine the alcohol concentrations. Subsequently changes happen and converted primarily to infrared spectroscopy. It falls into three category, Personal breathalysers which habits semiconductor oxide sensors, Professional breathalysers, such as [BACtrack Mobile](#), [BACtrack S80](#), [BACtrack Trace](#), [BACtrack S75](#), [BACtrack Element](#), and [Lifeloc FC-10](#), which trains using fuel cell sensors and finally the Spectrophometer breathalysers which uses infrared light to detect drunken drivers. This device becomes a valuable purchase for personal safety and safety of others.

### **40.2.3 Brethometer**

It is a small device that wads into our smart gadgets to measure the quality of alcohol in ours breathe. It is similar to the breathalyses, where it uses the help of an app in our gadgets to display the blood alcohol concentration level (BAC). It is so compact, so that it can even attach to the key chain or purse. It also reminds to check the BAC before we take a drive [17]. This device apart of checking BAC also analyses our



breathe for other chemical compounds that exist [1]. So the policemen are extremely convenient of using smart gadgets for detecting the presence of alcohol in our blood.

#### **40.2.4 Alco Blow**

A very simple device and a rapid response instrument to detect the breathe samples of the person for the presence of alcohol. It uses the fuel cell sensor, so there is no physical contact between the person and instrument. Device in active mode, the victim blows into the sampling cone to get the sample in the instrument for the alcohol analysis [20]. It is fully automatic, where within seconds a coloured light appears in the device with a beep tone indicating the presence of alcohol content in the blood level. In passive mode the operator presses the passive button to get the sample from around the suspected person. It also responds quickly in the display system whether the suspect consumes alcohol or not. Finally this device is suited to a condition where there are more number of people to be tested and provides us the result with minimum of instruction.

### **40.3 Related Work**

In [3] it discloses the breathe alcohol detection systems that detects the identity of the operator using skin sensor. It determines the skin characteristics by sensing the skin of the victim face or mouth. This detecting device detects the breathe alcohol concentration and permits igniting the vehicle on if below the predetermined threshold level, otherwise the vehicle is locked [9]. An infrared alcohol testing using differential absorption has been proposed in [14]. It determines that the alcohol gas has well-defined absorption characteristics inside infrared region of electromagnetic spectrum [13]. A non-dispersive infrared gas detection principle has been used to detect the presence and concentration of alcohol gas within a consumed victim [17]. Using the concepts of infrared transmission spectroscopy both alcohol and CO<sub>2</sub> can be measured in small handheld unit in [4]. Recently Breathalyzer integrated with ignition locking system to prevent the intoxicated drivers to start their vehicle [16]. This device uses an ethanol-specific fuel cell for a sensor to response to the initial breathes data. Later it uses a relay box to relay the information from the handheld unit to a command station for a warning alarm either through lights flashing, horn honking or both. In [18] the author installs either ignition interlock device or a breathe alcohol ignition interlock device to the dashboard of the vehicle. It prevents the driver from starting the vehicle only if he/she successfully passes the blood alcohol concentration (BAC) test. Here the driver has to blow to the BAC tester before he starts [10]. If the test results within the limit he/she can start and operate normally, otherwise vehicle will be locked and they can't use the vehicle. A novel approach of detecting intoxication from the motion differences was obtained

by the Wearable sensor devices [1]. The problem of drunkenness was obtained through supervised machine learning for both binary classification problem (drunk or sober) and a regression problem (the breathe alcohol content level) [13]. The author collected from 30 subjects using Google glass and LG-G watch, Microsoft Band and Samsung galaxy S4 and validated the result using Breathalyzer used by the police. The system successfully detected intoxication and avoid causing fatal accidents. The concept of Breathalyzer also takes it's another role in vehicle by embedding it in the steer of the car and senses the victim when he/she touches the car steering [8]. Using the sensor, the position of the finger is discovered and the force is calculated which is exerted by the finger on the gear. The alcohol level is indicated through different LED lights (high, medium, low) and locks the car from further operation.

#### **40.4 Fuel Cell Sensor**

Fuel cell sensor is the only reliable sensor for alcohol testing. Normally when a person consumes alcoholic beverages, the alcohol that is consumed is absorbed by the mouth, throat and later to the intestine and at last reaches the blood. When blood flow travels to the lungs, a fraction of the blood reaches alveoli, which becomes volatile. So this area and the area around it is actually the last portion where the air exhaled during an alcohol level test. The concentration of alcohol contained in alveolar air is related to blood alcohol concentration levels. The fuel cell sensor collects the breathe sample for the victim for the alcohol testing. This consists of a porous layer with electrolyte that is very delicate to ethanol. This layer is implanted between two platinum levels that act as electrolytes. The breathe air sample passes to the porous layer and the presence of ethanol allows the electrolyte to oxidize. During oxidation, various acetic acids, protons and electrons are created. The electric current that is generated has the value related to the blood alcohol content level and that is processed by the microprocessor. This sensor is extremely selective and sensitive for ethanol, a substance for measuring the alcohol level. These sensors are used in Breathalyzer for alcohol testing which are very accurate, precise, provide repetitive values and finally used to maintain these characteristics for long time comparison. It has the capability to perform better even when alcohol level is high and also test the alcohol close to the legal limits [23].

#### **40.5 Ignition Interlock Device (IID)**

IID is the newest device for the vehicles which prevents the cause of road accidents by the drunken drivers. This device is incorporated in the glove compartment of the vehicle and hard wired to the engines ignition system. Using the computer chip this device records the alcohol content of the driver, so when he tries to start the vehicle, the alcohol content will be downloaded from the IoT hub. If the Blood

Level Alcohol content is above the threshold limits, then this device triggers to lock the vehicle from starting. So this device prevents the driver from consuming alcoholic beverages before starting his vehicle and provides a safe environment from the terrible accidents.

## 40.6 Methodology

IoT takes the responsibility for controlling all the devices through Internet. The smart active helmet of the motorists and the ignition interlock device (IID) in the motors get connected through Internet by IoT. The IoT manage and process these devices without the need for any personalized computers. The main controller or the hub which is also referred as the gateway is the essential part of the smart devices. This central controller is connected both to the helmet of the victim and to the IID. So both the devices transmit or receive the commands using this centralized hub. Once the victim wears his smart helmet, the Fuel Cell Sensor collects victims breathe sample and calculates the blood alcohol content (BAC) of the victim. This sensor then reports the BAC value to the central hub. The hub after receiving the commands it communicates its response to the Cloud network located over the network. This is possible only if there is reliable internet connection between hub and smart phone that helps to connect to the cloud network. The storage and maintenance of data over internet can be done via cloud-based networking. Therefore the sensor can send over the cloud to the hub for the response. The hub responses by reading the signal from the intended sensor and analyses the BAC level of the signal and send the reply back to the Fuel Cell Sensor. If BAC level exceeds the threshold limit, then the hub immediately prompt the sensor to trigger the IID to lock the vehicle from starting. Once the required action is performed then the hub updates the information to cloud. This is to control and monitor every smart devices connected to internet. Wi-Fi communication protocols will help to provide the communication between the devices that is connected through internet.

Figure 40.1 represents the procedures to be followed for smart active helmet.

**Step 1:** Connect the Smart Phone with Mobile Data enabled option for IoT service.

**Step 2:** Internet cloud is activated by interconnecting all its active devices such as Smart Phone, Hub (Gateway), Smart Helmet and IID.

**Step 3:** The central controller (hub/gateway) enables the Fuel Cell Sensor in Smart Helmet when the victim wears his helmet.

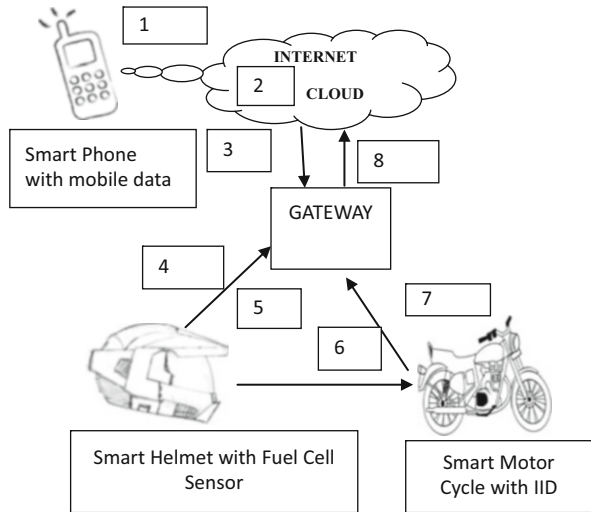
**Step 4:** The Sensor detects the breathe sample and calculates the Blood Alcohol Content level. It sends the signal to the Hub for the further processing.

**Step 5:** The hub on receiving the signal analyses the result in two cases:

If BAC level is below the threshold level then Hub responses to Fuel Cell Sensor for normal working of the vehicle.

Else if BAC level is above the threshold then Hub prompt sensor to trigger IID to lock the vehicle immediately from starting.

**Fig. 40.1** Procedure for smart active helmet



- Step 6:** On consumption of alcohol, the Fuel cell sensor activates the IID in motor cycle to lock.
- Step 7:** IID on receiving the signal from the sensor activates the ignition to lock and response to hub for further updating.
- Step 8:** Finally the Hub updates the information to cloud for storing and maintaining the entire device set-up for future use.

### 40.7 Significance of Smart Active Helmet

Incorporation of liquor-free driving helps to prevent from accidents, which in turn protects the life. Safest mode of travel is ensured which leads to happy journey in vehicle. Ensuring the traveller at instant checking can be ensured with the help of central hub technique in associated with IoT.

Accessing all the provisions with technology usage in just one click is the best way for human life travel. Easy way of liquor testing with simplified procedure can help to avoid unnecessary accidents and safe us from hazards.

### 40.8 Conclusion

As the population span increases in wider space, space to travel becomes highly tedious. The main root cause for accidents is liquor driving with negligence thoughts. To protect the human life, a new smart active helmet helps driving for

hassle free travelling. Convenient driving leads to pleasure journey. Life is protected with safest mode of travel involved in it. This is achieved with the help of smart active helmet.

## 40.9 Future Research Directions

This smart active helmet can be incorporated in urban areas, especially where the population of the people is more. This is an efficient way to control accidents in school zone, play area and at civilized areas.

## References

1. B. Nassi, L. Rokach, Y. Elovici, Virtual Breathalyzer” arXiv:1612.05083v1 [cs.Hc] 14 December (2016)
2. L. Biljana, R. Stojkoska, K.V. Trivodaliev, A review of Internet of things for smart home: challenges and solutions. *J Cleaner Prod* **140**, 1454–1464 (2017). <https://doi.org/10.1016/j.jclepro.2016.10.006>
3. J. Dai, J. Teng, X. Bai, Z. Shen, D. Xuan, Mobile phone based drunk driving detection, in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on-NO PERMISSIONS* (IEEE, 2010), pp. 1–8
4. Davin E. Lopez, Luis M. Ortiz, Beverages disguise for handheld breathalyzer interface of ignition interlock device, United States Patent. Patent No. US 8590364B2
5. A.J. Fernandez-Ares, A.M. Mora, S.M. Odeh, P. Garcia-Sanchez, Wireless monitoring and tracking system for vehicles: a study case in an urban scenario. *Simul. Model. Pract. Theory* **73**, 22–42
6. G. Loukas, Y. Yoon, G. Sakellari, T. Vuong, R. Heartfield, Computation offloading of a vehicle continuous intrusion detection workload for energy efficiency and performance. *Simul. Model. Pract. Theory* **73**, 83–94 (2017)
7. N. Gershenfeld, R. Krikorian, D. Cohen, The Internet of things. *Sci. Am.* **291**(4), 76–81 (2004)
8. N. Gomathi, S. Kumar, Internet of things based accident detection and prevention. *IJMET* **8**(9), 196–204 (2017)
9. T.D. Goswami, S.R. Zanwar, Z.U. Hasan, Android based rush and drunk driver alerting system. *Int. J. Eng. Res. Appl.* **4**, 1–4 (2014)
10. R. Hardy, E. Rukzio, Touch & interact: touch-based interaction of mobile phones with displays, in *Proceedings of ACM MobileHCI '08*, Amsterdam, The Netherlands, September (2008)
11. J. Hernandez, D. McDuff, R.W. Picard, Biowatch: estimation of heart and breathing rates from wrist motions, in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2015), pp. 169–176
12. J. Xu, J.J. Han, Y. Zhang, YU.'a. Sun, B. Xie, Studies on alcohol sensing mechanism of ZnO based gas sensors. *Sens. Actuators B* **132**, 334–339 (2008)
13. T. Jinkawa, G. Sakai, J. Tamaki, N. Miura, N. Yamazoe, Relationship between alcohol gas sensitivity and surface catalytic property of tin oxide sensors modified with acidic or basic oxides. *J. Mol. Catal. A* **155**, 193–200 (2000)
14. K.I. Ozoemena, S. Musa, R. Modise, Fuel cell-based breath-alcohol sensors: innovation hungry old electrochemistry. *Curr Opin Electrochem* (2018). [10. 10.1016/j.coelec.2018.05.007](https://doi.org/10.1016/j.coelec.2018.05.007)

15. L. Atzori, A. Iera, G. Morabito, The Internet things: a survey. *Computer Networks* **54**, 2787–2805 (2010)
16. Michael W. Walter, Douglas E. Devries, 2011 Ignition Interlock Breathalyzer, United states Patent, Patent No: US 7934577B2, Date of Patent: May 3 (2011)
17. M.R. Rahman Jesse, T.S. Allan, M.Z. Ghavidel, L.E. Prest, F.S. Saleh, E.B. Easton, The application of power-generating fuel cell electrode materials and monitoring methods to breath alcohol sensors. *Sens. Actuators B* **228**, 448–457 (2016)
18. R.M. Altarawneh, P. Majidi, P.G. Pickup, Determination of the efficiency of ethanoloxidation in a proton exchange membrane electrolysis cell. *J. Power Sources* **351**, 106–114 (2017)
19. B. Sterling, *Shaping Things—Mediawork Pamphlets* (MIT Press, 2005)
20. W. ding, S. Zhang, Z. Zhao, A collaboration calculation on real time stream in smart cities. *Simul. Model. Pract. Theory* **10**, 1016 (2017)
21. W. Xu, M.-C. Huang, N. Amini, J. J. Liu, L. He, and M. Sarrafzadeh, Smart insole: a wearable system for gait analysis, in *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments. ACM* (2012), p. 18.
22. N. Yamazoe, G. Sakai, K. Shimano, Oxide semiconductor gas sensors. *Catal. Surv. Asia* **7**, 63–75 (2003)
23. N. Yamazoe, New approaches for improving semiconductor gas sensors. *Sens. Actuators B* **5**, 7–19 (1991)
24. Z.H. Mir, F. Filali, Large scale simulations and performance evaluation of connected cars—A V2V communication perspective. *Simul. Model. Pract. Theory* **73**, 55–71 (2017)

# Index

## A

- Accident-prone areas
  - experimental setup, 363–364
  - Salesforce Einstein AI
    - data uploaded, 364–365
    - registered contacts, 365
    - for simulation, 365–366
    - vibration values, 365
  - statistical and real-time data, 361–362
- Acoustic method, 349–350
- Active appearance model (AAM), 271–272
- Adaptive network fuzzy inference system (ANFIS), 32, 33
  - degrees of membership, 34–35
  - membership function and gradient descent, 34
  - sigmoidal functions, 34
  - supervised learning, 34
- Adaptive neuro-fuzzy inference system, 32
- Ad hoc on-demand Distance Vector Routing (AODV) protocol, 63, 89
- AFC algorithm, *see* Ant–fuzzy clustering algorithm
- Agent-based model (ABM)
  - agent interactions, 283–284
  - behaviour of customers, 283
  - customer variable features, 282
  - percentage of customers, 284
  - rules of, 284
  - system performance, 283
- Agglomeration, 10
- Amazon Simple Notification Service, 167
- American Diabetes association, 325
- Amplify and Forward (AF) relaying protocol, 249
- Analog to Digital converters (ADC), 88
- Analysis of Deviance, 147
- Analysis of Variance (ANOVA), 141, 147
- ANFIS, *see* Adaptive network fuzzy inference system
- ANN, *see* Artificial neural network
- Anonymized call detail records (CDR), Sri Lanka, 299–300
  - Census and Statistics Department, 299, 300
  - cleaning and preprocessing data, 301–302
  - feature extraction, 301
    - behavioural features, 302
    - mobility features, 302–303
    - social network, 303
  - feature selection, 306
  - Hadoop, 301
  - high level process architecture, 300
  - home location detection, 303–304, 308
  - PCA, 309
  - RStudio, 301
  - socioeconomic index, 304–306
  - socioeconomic status groups, 306
  - SVM, 306–308
- Ant-based clustering
  - AFC algorithm (*see* Ant–fuzzy clustering algorithm)
  - EACA, 9, 10
  - K-means, 10
  - natural ant behavior, 9
  - PACE, 10
- Ant colony building (ACB), 10

- Ant colony optimization (ACO), 32
- Ant-fuzzy clustering (AFC) algorithm, 10
- bio-inspired metaheuristics algorithms, 11
  - for distributed database, 11–12
  - fuzzy C-means, 11
  - fuzzy partitioning, 11
  - Iris and Wine datasets
    - accuracy of clustering, 12, 14
    - error rates, 12, 13, 15
    - F*-measure, 12, 13
    - plotted predicted data, 12, 13
    - reality-based fuzzy algorithms, 11
- Anthropometric model, 271
- Anti Social Behaviors (ASBs), 290
- Ant odor identification model (AOIM), 10
- Arc welding, 347
- Arduino, 320
  - algorithm, 316–317
  - implementation, 317–319
  - results, 319–320
  - system design, 316
  - system model, 317–318
- Area Deviation Factor (ADF), 113
- ARMA, *see* Autoregressive-moving-average model
- Artificial intelligence (AI)
  - based analytics
    - data uploaded, 364–365
    - registered contacts, 365
    - for simulation, 365–366
    - vibration values, 365
  - neural network, 355
- Artificial neural network (ANN), 182, 208–210, 240, 343
- Assisted self-learning of yoga practice
  - angle calculation, 235
  - assistive technology act, 232
  - flow of work, 234
  - interventions, 231
  - Nuitrack, 233
  - risk of falling, 231
  - standing posture/sitting posture, 233–235
- Automatic Vehicle Identification (AVI), 362
- Autoregressive-moving-average model (ARMA), 213, 384–386, 388–389
- B**
- BAC, *see* Blood alcohol concentration level
- Bayes decision strategy, 156
- Beam Division Multiple Access (BDMA), 101
- Best margin classifier, 35
- Big data and blockchain (BD-BC) based decision support model
- blockchain-based crop management system, 81–83
- demand-based efficient decision support system
  - Aadhaar data, 80, 81
  - cloud-based framework, 79
  - Pahani records, 81
  - quantity of yield, 81
  - system architecture, crop selection model, 79, 80
- implementation
  - datasets, 83
  - demand supply gap, 83
  - expected reduced gap, 84
  - price inflation, 85
  - regulated demand supply, 85
  - Web server, crop selection and allocation, 83, 84
  - land ownership registration, 78
  - supply chain management, 78
- Bitcoin prediction
  - applications, 382
  - factors, 381–382
  - literature survey, 383
  - time series analysis
    - ARIMA, 384–386
    - characteristics, 383
    - co-variance, 384
    - deep learning based models, 384
    - dependent variable, 384
    - independent variable, 384
    - non-stationarity nature, 385
    - random component, 384
    - results, 387–390
    - semi-supervised learning, 384
    - serial correlation, 384
    - stationary time series, 385
  - transactional properties, 381
- Blockchain-based crop management system, 81–83
- Blood alcohol concentration level (BAC)
  - breathalyser, 394–396
  - brethometer, 394–395
  - drunkometer, 394
  - fuel cell sensor, 396
  - IID, 396–397
  - infrared testing, 395
  - non-dispersive infrared gas detection, 395
  - relay box, 395
  - skin sensor, 395
  - victim blows, 395
  - Wearable sensor devices, 395–396



## C

- Center of area (COA) method, 71
- Civil unrest detection system
  - architecture of, 290–291
  - clustering model, 292–293
  - extracted keywords, 294–295
  - frequent words, 294–295
  - keyword learning and filtering, 292
  - preprocessing, 291
  - protest-related keywords, 293–294
  - SSE, 294–296
  - Twitter API and R tool, 293
- Classifier process
  - CNN, 355
  - decision rules, 357
  - evaluation, 357–359
  - feature selection, 356
  - HE method, 357
  - IBL, 355–356
  - neural network, 355
  - normalized information, 356–357
  - and prediction, 353–354
  - rule based learning, 356
  - training data, 356
- Cluster head (CH), 90–92
- Cluster Head Election mechanism using Fuzzy logic (CHEF), 66, 74
- Clustering model, 259, 267
- CMF detection, *see* Copy-move forgery detection
- CMM, *see* Computer Aided Design mesh models
- CNN, *see* Convolutional neural network
- Coded Cooperation (CC) relaying protocol, 249
- Code Division Multiple Access (CDMA) multiplexing, 97–98, 101
- Cognitive radio (CR), 181, 185–188
- Cognitive radio network (CRN), 184
- Comma Separated Values (CSV), 365
- Compact Muon Solenoid (CMS), 341
- Compress and Forward (CF) relaying protocol, 249
- Computer Aided Design mesh models (CMM)
  - applications, 109
  - B-rep models, 109, 110
  - curvature, 109–110
  - mesh segmentation, 109, 110
  - methodology
    - hybrid mesh segmentation, 114–115
    - preprocessing, 112–113
    - volumetric feature recognition, 115–116
  - results, 116–117
  - rule-based reasoning, 110
- Confusion matrix, 358
- Context analysis, 123–124
- Contour finding algorithm, 226
- Control plane (CP), 88, 90–92
- Convolutional neural network (CNN), 182, 187, 355
- Cooperative network
  - destination node (D), 249
  - multiple intermediate relay nodes (R)
    - AF relaying protocol, 249
    - CC relaying protocol, 249
    - CF relaying protocol, 249
    - DF relaying protocol (*see* Decode and Forward (DF) relay model)
    - spatial diversity, 249
  - source node (S), 249
- Cooperative spectrum sensing (CSS), 182
- Copy-move forgery (CMF) detection
  - block-based detection methods, 17–18
  - feature-based detection methods, 18
  - simulation results and performance analysis
    - average contingency values, 27
    - vs.* FMM, SVM, NB, and KNN
      - performance, 27, 28
    - MIFCC\_600 dataset, 19, 20
    - precision *vs.* recall, 27
  - by Stockwell transform (S transform)
    - amplitude and phase representation, 20
    - Euclidean distance, 21
    - feature vector, 20–21
    - FMMNN-DT (*see* FMMNN-DT, forgery detection)
    - image blocks, 20–21
    - mean, standard deviation, and average residual, 20
    - proposed CMF detection system, 19
- Crack detection
  - image acquisition, 339
  - image enhancement, 340
  - image restoration, 340
  - preprocessing, 340
  - process of, 339–340
  - welding (*see* Welding)
- Crop yield
  - ARMA, 213
  - clustering algorithms, 214
  - defuzzification, 217, 218
  - fuzzy forecasting models, 214–216
  - fuzzy logical relationships, 215–216
  - performance, 217–218
  - rice production, 217
  - universe of discourse, 215–216
- Cryptocurrency, *see* Bitcoin prediction
- CSS, *see* Cooperative spectrum sensing

- Customer churn
  - classical algorithms
    - decision tree, 207
    - K-nearest neighbor, 207
    - logistic regression, 206
    - Naive Bayes, 206
    - Random forest, 206–207
  - data preprocessing
    - attribute types, 204
    - standardization, 205
    - tokenization, 205
  - dataset selection, 204
  - deep learning algorithm
    - ANN, 208–210
    - model representation, 209, 210
    - ROC curve, 210, 211
    - variables and features, 203
- Cyclostationary signal (PU), 185–186
- D**
- Data bandwidth, 97–98
- Dataloader.io, 365
- Data mining, 9, 137, 138
  - See also* Feature selection
- Data preprocessing, 204–205
- D-2-D communications, 104
- Decision tree (DT), 138, 139, 207, 356
- Decode and Forward (DF) relay model
  - end-to-end bit error probability, 250
  - evaluation and simulation result, 255–257
  - multihop DF relay model with single branch
    - SEP, 251–252
    - system model, 250–251
  - multihop multiple branch DF relaying model
    - SEP, 254–255
    - system model, 253–254
  - multihop single branch relaying without diversity, 249
- Deep belief network (DBN)
  - cost sensitive ordinal hyper planes, 270
  - DeepConvNets, 270
  - DeePID, 270
  - evaluation metrics, 274–275
  - experimental results, 275–276
  - feature extraction
    - AAM, 271–272
    - anthropometric model, 271
    - local and global features, 271
    - scattering features, 272
  - IMDB face database, 274
  - LUPI framework, 270
  - MAE, 270–271
  - preprocessing, 271
  - RBM
    - features, 272–273
    - gradient problem, 273
    - interlayer communication, 272
    - Layers, 272–273
    - training, 273–274
  - SVM, 270
- Deep convolution neural network (DeepConvNets), 270
- Deep learning algorithm
  - ANN, 208–210
  - model representation, 209, 210
  - ROC curve, 210, 211
- Deep neural network (DNN), 184
- Defuzzification, 64, 71, 217, 218
- Degree of differencing, 386
- Demand-based efficient decision support system, 79–81
- Deviance residual, 146
- DF relay model. *see* Decode and Forward relay model
- DHP. *see* Diagonal Hexadecimal Pattern
- Diabetes Risk Assessment Tools, 324–325
- Diagonal Hexadecimal Pattern (DHP), 223
- Dice coefficient, 132
- Dickey-Fuller test, 385
- Digital image forgery detection, *see* Copy-move forgery detection
- Disaster management
  - ANN, 182
  - CNN, 184
  - DNN, 184
  - flying cell towers, 183
  - signal-to-noise ratio vs. detection probability, 188
  - SpecCNN model
    - cyclostationary signal feature extraction, 185–186
    - training algorithm, 187–188
  - UAVs, 181, 182
- Discrete artificial bee colony (DABC) algorithm, 32
- Distributed database, 9
- Distributed energy-efficient clustering algorithm (DEEC), 66
- Distributed fuzzy logic control (DFLC) approach, 66
- Distributed Unequal Clustering using Fuzzy logic (DUCF) scheme, 66, 74
- Divisional secretariat divisions (DSDs), *see* Anonymized call detail records (CDR), Sri Lanka

- Document similarity assessment
  - grammatical linkages, 132
  - graph-based centrality, 132
  - graphical unit-based methods, 132
  - semantic graphs, 132
  - semantic similarity, 131, 132
  - system architecture
    - Data preprocessing module, 133
    - tokenization, 133
    - verbal intent modelling, 134
  - text-based techniques, 132
  - word distribution statistics, 131–132
- DOT-NET software, 330
- Dynamic Source Routing (DSR), 170
- DynamoDB table, 167
- E**
- EEDUC technique. *see* Energy-Efficient Distributed Unequal Clustering technique
- Electrooculogram, 152
- ElGamal method, 331, 335
- Elliptic curve cryptography (ECC)
  - additive abelian group, 331
  - cryptoprocessor, 330
  - decryption method, 332–333
  - definition, 329
  - encryption architecture, 331–332
  - entropy analysis, 336–337
  - isomorphism, 331
  - literature survey, 330
  - motivation, 330
  - performance analysis, 335–336
  - simulation setup and results, 333–334
- Emoticons, 192–194
- Energy-Aware multicast cluster (EAMC)-based routing, 54
- Energy-Efficient Distributed Unequal Clustering (EEDUC) technique
  - CH election, 70–71
  - defuzzification, 64, 71
  - fuzzy logic inference system, 64, 65
  - fuzzy parameters, 67–68
  - if then rules, 65
  - Mamdani fuzzy inference system, 67
  - network assumptions, 68
  - simulation setup and performance metrics
    - network lifetime, 71–73
    - per round energy consumption, 74
    - simulation parameters, 71, 72
    - sink location, 71
  - states, 68–69
  - tentative cluster head, 69
  - unequal clustering, 68, 69
- Energy-efficient multiple distance-aware clustering (EEMDC) protocols, 54
- Energy-efficient unequal clustering (EEUC) protocol, 54
- Enhanced ant clustering algorithm (EACA), 9, 10
- Entity Resolution (ER)
  - cluster, 41, 43
  - de-duplication, 41, 42
  - entities, 41, 42
  - flowchart, 44, 45
  - OYSTER, 41
    - handling ambiguous data, 48
    - identity capture, 43–48
    - match rules, 43
- Entropy ( $H$ ), 207
- European Agency for Health and Consumers (EAHC), 324–325
- F**
- Facial based human age estimation. *see* Deep belief network (DBN)
- FAIR tool analysis, 378–379
- Feature selection, 31
  - literature survey, 32
  - motivation, 31–32
  - NF-ABC (*see* Neuro-fuzzy ant bee colony based feature selection)
- Feed forward neural network (FFNN), 5, 7
- Fisher's scoring algorithm, 146
- 5G wireless network
  - deployment challenges
    - bandwidth utilization, 105
    - communication, navigation, and sensing, 106
  - efficient medium access control, 106
  - legislation of cyber law, 106
  - radiation hazards, 105
  - RAN Research, 105
  - security and privacy, 106
  - traffic management, 106
- features, 98
- implementation
  - D-2-D communications, 104
  - massive MIMO, 103
  - MMC, 104–105
  - moving network, 103
  - UDN, 103
  - URC, 104
- OFDM technique, 98
- packet switched wireless system, 101

5G wireless network (*cont.*)  
 performance parameters  
 evaluation, 102  
 network performance, 101–102  
 QoS performance, 102  
 related work  
 Alcatel-Lucent, 100  
 Ericsson, 99  
 Federal Communications Commission (FCC), 100–101  
 Fujitsu and DOCOMO, 100  
 iJOIN EU project, 99  
 METIS, 99  
 NEC, 99  
 Nokia, 99  
 Samsung, 100  
 Tokyo Institute of Technology, 100  
 Flat Rayleigh fading channel, 257  
 FMMNN-DT, forgery detection  
 algorithm, 23–24  
 hyperbox creation, 21–22  
 hyper container fuzzy set, 21  
 $k$ th pattern class, 21  
 neural network structure, 22  
 structure and membership function, 25, 26  
 unpredictable data, 23  
 upper distance, 23  
 4G network, 97  
 Fuel cell sensor, 396  
 Fully connected (FC), 187  
 Fuzzy ARTMAP, 342  
 Fuzzy-based bio-inspired approach, 9  
 Fuzzy forecasting models, 214–216  
 Fuzzy-neural system, *see* Crop yield

## G

Gaussian curvature, 112  
 Gaussian kernel function, 156  
 Generalized Cross-Correlation (GCC)  
 approach, 222  
 General linear model (GLM), 144–148  
 General self-organization tree-based  
 energy-balance (GSTEb) routing  
 protocol, 51  
 block diagram, 55–56  
 delay plot, 57, 59  
 node packet delivery plot, 57, 59  
 simulation window  
 existing system, 57, 58  
 proposed system, 57, 58  
 S-LZW compression algorithm, 57  
 tree-based data aggregation scheme, 56

Geographic energy-aware routing (GEAR-CC)  
 protocol, 54  
 Global Database of Events, Location, and Tone (GDELT), 290  
 Global Positioning System (GPS), 222  
 ‘GlobFit’ method, 111  
 Graph-based centrality, 132  
 Gray level co-occurrence matrix (GLCM), 343  
 Group of Pictures (GOP), 226  
 Group prototypes, 174–176  
 GSTEB routing protocol, *see* General self-organization tree-based energy-balance routing protocol

## H

Hierarchical Fitting Primitives (HFP)  
 technique, 110  
 Histogram of Orientated Gradients (HOG), 18  
 Homomorphic encryption (HE) method, 357  
 Human Machine Interaction, 320  
 algorithm, 316–317  
 implementation, 317–319  
 results, 319–320  
 system design, 316  
 system model, 317–318  
 Human organs  
 bioinformatics, 377  
 DNA sequences, 378  
 experiment analysis, 378–380  
 repeats of sequences, 377–378  
 single-amino-acid tandem repeats, 378  
 Hybrid energy-efficient distributed (HEED)  
 clustering algorithm, 66  
 Hybrid methodology, 192

## I

IBM Telco Customer Churn dataset, 204  
 Ignition interlock device (IID), 396–397  
 iJOIN EU project, 99  
 Image compression  
 flowchart, 5, 6  
 literature survey, 4–5  
 lossless compression technique, 3  
 lossy compression technique, 3–4  
 methodology, 5, 6  
 parameters, 7  
 value of  $K$ , 6–7  
 IMAGE Group 2006309-IMAGE, 324–325  
 IMDB face database, 274–276  
 Improved linear regression model, 259,  
 263–265  
 Improved UFHLSNN (IUFHLSNN)  
 dataset, 244

- fuzzy neural network, 239–240
- HLS, 241
- IBM's POWER8 and Intel's Xeon E5-2620, 241, 245, 246
- OpenMP, 240
- pattern classification, 239
- recognition algorithm, 242, 243
- serial and parallel experimentation, 244, 245
- speed up and percentage gain, 244
- test bench specification, 244
- training algorithm, 242, 243
- Indian Diabetes Risk Score (IDRS), 324
- Input image, 5
- Input pre-processing, 5, 6
- Instance based learning (IBL), 355–356
- Intelligent Transportation Systems (ITS), 361–362
- Intel Xeon E5-2620, 245, 246
- Internet of Things (IoT), 163, 165, 397, 398
- IoT Cloud Platform, 324–325
- IRHO database, 377, 379–380
  
- K**
- Keyword-based approach, *see* Social media
- K-nearest neighbor (KNN) classifier, 10
  - customer churn, 207
  - DBN, 275
  - IBL, 355–356, 358–359
  - NF-ABC based feature selection, 35
  - See also* Modified Fuzzy KNN Classifier (MFKNN)
  
- L**
- Lambda rules, 167
- Laser welding, 344
- LEACH, *see* Low-Energy Adaptive Clustering Hierarchy
- Learner analysis, 122, 123
- Learning styles, 122, 126, 128
- Learning using privileged information (LUPI) framework, 270
- Legislation of cyber law, 106
- Lempel-Ziv-Welch (LZW) algorithm, 52
- Lexicon-based methodology, 192
- Lex rank mechanism, 132
- Linear Discriminant Analysis (LDA), 139
- Linear Regression, 384–386, 388–389
- Line of sight (LoS), 188
- Liquid (dye) penetrant method, 348
- Local binary pattern (LBP), 18
- Local binary pattern variance (LBPV), 18
  
- Logistic regression, 206, 362
- Loop/radar Detector (LD) data, 362
- Low-Energy Adaptive Clustering Hierarchy (LEACH), 65–66, 74, 89
  
- M**
- Machine learning, *see* Customer churn
- Machine-to-Machine Intelligence (M2Mi) Corp, 98
- Mamdani fuzzy inference system, 67
- MANETs, 54
- Map-Reduce technology, 281–282
- Massive machine communications (MMC), 104–105
- Massive MIMO, 103
- MATrixLABoratory (MATLAB), 227
- MaxRel algorithm, 306
- Mean absolute error (MAE)
  - age estimation process, 270–271
  - classifier, 274–276
- Mean imputation method, *see* Missing data
- Medical record system
  - CD-ROM, 42
  - electronic medical record systems, 42
  - ER (*see* Entity Resolution)
  - paper-based medical record system, 42
- Message Queue Telemetry Transport (MQTT), 167
- Metal inert gas (MIG) welding, 347
- METIS, 99
- Metric F-score, 223
- MFKNN, *see* Modified Fuzzy KNN Classifier
- Minimum Feature Dimension (MFD), 112
- Missing at Random (MAR), 140
- Missing completely at Random (MCAR), 140
- Missing data
  - ANOVA, 147
  - data mining techniques, 137, 138
  - decision tree method, 138, 139
  - Fisher's scoring algorithm, 146
  - GLM, 144–148
  - MAR, 140
  - margin plot of dataset, 142
  - MCAR, 140
  - missing pattern, 142
  - MNAR, 140–141
  - order method type, 143–144
  - principal component analysis, 139
  - revenue and product cost, 142, 143
  - SVM algorithms, 138
  - training dataset, 139
  - variables, 141

- Missing not a Random (MNAR), 140–141
- m-Learning, 121–122
- factors affecting
    - context analysis, 123–124
    - learner analysis, 122, 123
    - learning styles, 122
    - meta-cognition, 122
  - Java Programming, case study
    - content formats, 125, 127
    - context characteristics and possible values, 124, 125
    - context scenario, 124, 126, 127
    - learning styles, 126, 128
    - number of students participated, 124
    - 240 students' interests in, 124
- Mobile learning, *see* m-Learning
- Mobile wireless sensor networks (MWSN)
- advantages, 88
  - applications, 88
  - components, 88
  - definition, 88
  - SDN (*see* Software defined networking)
  - vs.* static WSN, 88
- Modified Fuzzy KNN Classifier (MFKNN)
- actual test data *vs.* model prediction, 178, 179
  - fuzzy logic and fuzzy sets, 175–176
  - generalized representation, 174
  - group prototypes, 174–176
  - performance of, 177, 178
  - plot of distance  $d$  *vs.* number of prototypes, 178, 179
  - plot of  $K$  *vs.* percentage recognition rate, 177, 178
  - prototype creation, 177
  - skin segmentation, 173
  - testing phase, 176
  - training and test sets, 177
- MoodLens, 193
- Moving network (MN), 103
- Moving vehicles and speed estimation
- accuracy, 227–229
  - classification method, 223
  - contour finding algorithm, 226
  - equipment, 221–222
  - GCC, 222
  - MATLAB, 227
  - PPCA, 224–225
  - SBF, 226
  - STVF, 226
  - threshold algorithm, 223
  - VSM, 223
  - VTSM, 223
- MPSK modulation, 255, 256
- mRMR algorithm, 306–309
- Multilayer perceptron artificial neural network (MLP-ANN), 343
- Multilayer perceptron neural network (MLP NN), 341
- Multiple Object Tracking Accuracy (MOTA), 223
- Multistage vector quantization (MSVQ), 4
- Multi-step radiographic image enhancement algorithm (MSRE), 343–344
- MWSN, *see* Mobile wireless sensor networks
- N**
- Naïve Bayes (NB) algorithm, 195, 357–359
- Naïve Bayes classifier, 197–198, 206
- Nakagami channel fading, 249–250
- Neural network (NN), 275, 355
- Neuro-fuzzy ant bee colony (NF-ABC) based feature selection
- ABC, 32, 33
  - ANFIS, 32, 33
    - degrees of membership, 34–35
    - membership function and gradient descent, 34
    - sigmoidal functions, 34
    - supervised learning, 34
  - ant bee algorithm, 33
  - ant colony optimization, 32
  - classification methods
    - k-NN algorithm, 35
    - SVM algorithm, 35–36
  - vs.* kNN and SVM
    - accuracy, 36, 39
    - data collection and clustering, 36
    - feature subset selection, 36, 37
    - selected features, 36–38
  - objective function, 32–33
  - system architecture, 33
- n-gram, 132
- No Line of Sight (NLoS), 188
- Non-linear predictive model, 355
- Nonparametric methods, 152, 153
- Non-standardized index (NSI), 305–306
- Northbound interface (NI), 164
- Nuitrack, 233
- O**
- On-demand multipath routing ad hoc protocol, 89
- 1G network, 97
- Online sites, 9
- Open Multi-Processing (OpenMP), 240

Open system entity resolution (OYSTER), 41  
 handling ambiguous data, 48  
 identity capture, 43  
   attributes, 44, 46  
   cluster as output, 46, 47  
   false-negatives, 46  
   false-positives, 45–46  
   match rules, 43, 44, 46  
   number of clusters, 46, 47  
   OYSTER ID (OID), 48  
   true-negatives, 45  
   true-positives, 44–45  
   XML code, 44, 46

Order method type, 143, 144

**P**

Parametric method, 153

Part-of-Speech (POS) tagging, 133

Periodogram method  
   left-hander subjects, 156–157  
   right-hander subjects, 157–159

Plotting rolling statistics, 385

Polar harmonic transform (PHT), 18

Polar sine transform (PST), 18

Polynomial Regression, 390

POWER8, 241, 245–246

PPCA, *see* Probabilistic Principal Component Analysis

Principal component analysis (PCA)  
   anonymized CDR, 304–305, 309  
   DBN, 272  
   PPCA, 224–225

Probabilistic ant-based clustering (PACE)  
   algorithms, 10

Probabilistic neural network (PNN), 156

Probabilistic Principal Component Analysis (PPCA), 224–225

Property prices  
   factors affecting, 259, 261–262  
   prediction models  
     clustering model, 259, 267  
     generalized model, 260–261  
     improved linear regression model, 259, 263–265  
     related work, 260  
     result analysis, 267  
     surveying regional factors, 262  
     universal prediction algorithm, 262–263  
     weight points calculation, 265–266

Protein sequences, 377–378

PSNR, 4, 8

**Q**

Quality of Experience (QoE) constraints, 103  
 Quality of Service (QoS), 101

**R**

Radial basis function (RBF) neural network, 4

Radiation hazards, 105

Radio Access Network (RAN) Research, 105

Random forest, 206–207

RANSAC (RANdom Sample Consensus)-based framework, 111

Recognition algorithm, 242, 243

Rectified linear unit (ReLU), 187

Recurrent neural network (RNN), 184

Reduced Instruction Set Computing (RISC), 241

Region of interest (ROI) selection, 343

Restricted Boltzmann machine (RBM)  
   features, 272–273  
   gradient problem, 273  
   interlayer communication, 272  
   Layers, 272–273  
   training, 273–274

RFID tracking system, 372, 373

Road Side Unit (RSU), 165

**S**

Sales order method, 143, 144

Scan-derived mesh, 109

ScatNet, 272

Scattering transform, 272

SDN, *see* Software defined network

Secondary users (SUs), 185

Sentiment analysis (SA)  
   accuracy, 197  
   components, 194  
   consumer-generated content, 191  
   data preparation, 195  
   emoticons and E-type values, 196, 197  
   with emoticons and without emoticons, 198, 199  
   experimental setups, 196, 197  
   F-score, 197  
   machine learning tool(s), 195  
   methodologies, 192  
   Naïve Bayes classifier, 197–198  
   precision, 197  
   recall, 197  
   system architecture, 195–196  
   text v/s emoticons, 192–193  
   Twitter, 195  
   word2vec and K-means algorithm, 193

- SEP, 251–252, 254–255, 257
- Sequential covering algorithm (SCA), 356
- Simple Boundary Follower (SBF) method, 226
- Simple linear iterative clustering (SLIC) algorithm, 18
- S-LZW compression algorithm, 57
- “Small cell” technology, 99
- Smart active helmet
  - breathe alcohol testing equipment
    - breathalyser, 394–396
    - brethometer, 394–395
    - drunkometer, 394
    - fuel cell sensor, 396
    - IID, 396–397
    - infrared testing, 395
    - non-dispersive infrared gas detection, 395
    - relay box, 395
    - skin sensor, 395
    - victim blows, 395
  - Wearable sensor devices, 395–396
- IoT, 397, 398
- liquor-free driving, 398
- procedures to, 397–398
- in urban areas, 399
- Wi-Fi communication protocols, 397
- Smart trash can system, 167–168
- Social media
  - automatic detection, 288
  - challenge, 288
  - civil unrest detection (*see* Civil unrest detection system)
  - criminal gangs and terrorist organizations, 288
  - event detection and forecasting, 289
  - spatiotemporal mining, 289
  - Suspicious Behaviors, early detection of, 289–290
  - volume of uncensored data, 288
- Socioeconomic status (SES), *see* Anonymized call detail records (CDR), Sri Lanka
- Software defined network (SDN)
  - architecture, 166–167
  - cluster head, 90–91
  - control plane, 88, 90–91
  - database structure, 170
  - data forwarding phase, 92–93
  - initialization phase, 91–92
  - Internet of Things, 163, 165
  - northbound interface, 164
  - packet reception ratio, 94
  - performance metric reliability, 93–94
  - priority and control information generation phase, 92
  - related work, 89–90
  - RSU, 165
  - salient features, 170, 171
  - SDON, 165
  - shortest path, 169–170
  - smart trash can system, 167–168
  - southbound interface, 164
  - system model, 90–91
  - trash can level, 169
  - user interface, 168, 169
- Software-Defined Optical Network (SDON), 165
- Somewhat Homomorphic Encryption (SHE), 357
- Southbound interface (SI), 164
- Spatio-temporal Varying Filter (STVF), 222, 226
- Spawn 8 threads (SMT8), 246
- SpecCNN model
  - cyclostationary signal, 185–186
  - training algorithm, 187–188
- Spectral correlation function (SCF), 186
- Speed Estimation and Detection (SED) method, 227–228
- Standardized index (SI), 306
- Standard triangulated language (STL), 110
- Stationary time series (TS), 385
- Stationary wavelet transform (SWT)-based features, 18
- Statistical analysis of classification algorithm, *see* Missing data
- Sum of Squared Error (SSE), 294–296
- Supervised learning, 34
- Support vector machine (SVM), 35–36, 306
  - confusion matrix, 358–359
  - DBN, 270, 275
  - MaxRel features, 307–308
  - mRMR-MID, 307–308
  - mRMR-MIQ features, 307, 309
  - RBF, 307
- Sustainable agriculture system
  - BD-BC-based crop management system (*see* Big data and blockchain based decision support model)
  - demand-supply management service system, 77
  - in India, 77
  - supply chain management, 78
- SVM, *see* Support vector machine
- Symbol error probability, 257



**T**

- Task identification system
  - classification method, 156
  - experimental protocol, 153, 154
  - feature extraction, 153, 155
  - graphical user interface, 152
  - HCI, 152
  - instrumental amplifier AD620 and operational amplifier LM741, 153
  - left-hander subjects, 156–157
  - right-hander subjects, 157–159
  - signal-processing algorithm, 152
  - types, 151
  - voltage threshold algorithm, 152
- Telecommunication customer retention
  - potential churners, 280
  - Randomized Method
    - ABM, 282–284
    - input dataset, 280–281
    - Map-Reduce technology, 281–282
    - results, 284
    - unstructured user dataset, 281
  - real-world entities, 280
  - statistical methods, 280
- Temporal filtering method, 226
- Term frequency-inverse document frequency (TF-IDF), 292–294
- Text segment, 192–193
- 3G network, 97
- Time Division Multiple Access (TDMA)
  - multiplexing, 97
- Tokenization, 205
- Traffic management, 106
- Tree-based data aggregation scheme, 52, 56
- True positive (TP) rate, 358
- Tungsten inert gas (TIG) welding, 347
- Twitter, 195
- 2-dimensional discrete wavelet transform (2D-DWT), 4
- 2G network, 97
- Type 1 diabetes (T1DM), 323–324
- Type 2 diabetes (T2DM), 323–324

**U**

- Ultradense networks (UDN), 103
- Ultrareliable communication (URC), 104
- Ultrasonic sensor, 319–320
- Unmanned aerial vehicles (UAVs), 181, 182, 372–373

**V**

- Vector quantization (VQ), 5, 7
- Vector Space prototype, 134
- Vehicle Speed Measurement (VSM) method, 223, 227–229
- Vehicle Tracking and Speed Measurement (VTSM), 223, 227–229
- Virtual Private Network (VPN), 101
- Visibility Related (VR) crash occurrence, 362
- Volumetric feature recognition
  - in CAD mesh models (*see* Computer Aided Design mesh models)
  - literature review, 110–112
  - scan-derived mesh model, 109
  - STL, 110

**W**

- Warehouse inventory management
  - architecture of, 374
  - benefits, 374–375
  - in Flipkart, 371–372
  - hardware setups, 375
  - implementation procedure, 373
  - MSM solutions, 372
  - system reliability, 374
  - UAV, 372–373
- Weight points, property prices, 265–266
- Welding
  - acoustic emission measurements, 342
  - ANN classifier, 343
  - arc welding, 347
  - automated weld defect recognition, 343
  - data-driven, 342
  - electromagnetic acoustic transducer, 341
  - features, 344–346
  - flux-cored arc welding, 347
  - fusion, 340
  - fuzzy ARTMAP, 342
  - fuzzy enhancement algorithm, 343–344
  - GLCM, 343
  - image registration techniques, 342
  - imperfections, 342–343
  - laser welding, 344
  - literature survey, 344
  - MIG welding, 347
  - MLP-ANN, 343
  - MLP neural networks, 341
  - MSRE, 343–344
  - non-destructive methods, 341

- Welding (*cont.*)
- acoustic method, 349–350
  - characteristics defects, 350
  - eddy current testing, 349
  - liquid (dye) penetrant method, 348
  - magnetic particles, 348–349
  - radiography, 348
  - ultrasonic inspection, 350
  - visual inspection, 348
- parameters, 343
- pattern recognition method, 341
- TIG welding, 347
- ultrasonic phased-array technology, 341
- vision inspection system, 342
- Wireless sensor networks (WSNs)
- additive abelian group, 331
  - AODV protocol, 63
  - base station, 87
  - clustering
    - CHEF, 66
    - Cluster Heads, 63, 64
    - Cluster Members, 63, 64
    - DEEC, 66
    - DFLC, 66
    - distributed clustering technique, 64
    - DUCF scheme, 66
    - EEDUC technique (*see* Energy-Efficient Distributed Unequal Clustering technique)
    - HEED, 66
    - LEACH, 65–66
  - cryptoprocessor, 330
  - DAIC, 53
  - data compression, 51
  - decryption method, 332–333
  - definition, 329
  - dynamic tree-based routing protocol, 51
  - ELDC protocol, 54
  - encryption architecture, 331–332
  - entropy analysis, 336–337
  - evolutionary algorithm, 53
  - fitness function, 53
  - GEAR-CC protocol, 54
  - GSTEB protocol (*see* General self-organization tree-based energy-balance routing protocol)
  - isomorphism, 331
  - LEACH, 52
  - literature survey, 330
  - load balanced clustering algorithm, 52
  - LZW compression algorithm, 52
  - motivation, 330
  - multi-hop communication, 53–54
  - MWSN (*see* Mobile wireless sensor networks)
  - node deployments, 64
  - performance analysis, 335–336
  - sensor nodes, 63, 87, 88
  - simulation setup and results, 333–334
  - TDMA scheduling, 52
  - total energy consumption, 51
- Word Net, 131
- word2vec, 193
- WSNs, *see* Wireless sensor networks