SPRINGER BRIEFS IN APPLIED SCIENCES AND TECHNOLOGY · FORENSIC AND MEDICAL BIOINFORMATICS

Sasikumar Gurumoorthy
Bangole Narendra Kumar Rao
Xiao-Zhi Gao

Cognitive Science and Artificial Intelligence

Advances and Applications

Springer

SpringerBriefs in Applied Sciences and Technology

Forensic and Medical Bioinformatics

More information about this series at http://www.springer.com/series/11910

Sasikumar Gurumoorthy
Bangole Narendra Kumar Rao
Xiao-Zhi Gao

Cognitive Science and Artificial Intelligence

Advances and Applications

Springer

Sasikumar Gurumoorthy
Department of Computer Science
and Systems Engineering
Sree Vidyanikethan Engineering College
Tirupati, Andhra Pradesh
India

Bangole Narendra Kumar Rao
Department of Computer Science
and Systems Engineering
Sree Vidyanikethan Engineering College
Tirupati, Andhra Pradesh
India

Xiao-Zhi Gao
Machine Vision and Pattern Recognition
Laboratory
Lappeenranta University of Technology
Lappeenranta
Finland

ISSN 2191-530X ISSN 2191-5318 (electronic)
SpringerBriefs in Applied Sciences and Technology
ISSN 2196-8845 ISSN 2196-8853 (electronic)
SpringerBriefs in Forensic and Medical Bioinformatics
ISBN 978-981-10-6697-9 ISBN 978-981-10-6698-6 (eBook)
https://doi.org/10.1007/978-981-10-6698-6

Library of Congress Control Number: 2017962048

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Measurement of Disease Severity of Rice Crop Using Machine Learning and Computational Intelligence 1
Prabira Kumar Sethy, Baishalee Negi, Nalini Kanta Barpanda, Santi Kumari Behera and Amiya Kumar Rath

1 Introduction 1
2 Literature Review 2
3 Proposed Methodology 4
4 Experimental Results 7
5 Conclusion and Future Scope 8
References 10

Flue-Cured Tobacco Leaves Classification: A Generalized Approach Using Deep Convolutional Neural Networks 13
Siva Krishna Dasari, Koteswara Rao Chintada and Muralidhar Patruni

1 Introduction 13
2 Convolutional Neural Networks 14
2.1 Convolutional Layer (CL) 14
2.2 Pooling Layer 14
2.3 Back Propagation on Conv Layer 15
3 Tobacco Leaves Classification System 16
3.1 Convnet Architecture and Model 16
3.2 Details of Learning 17
4 Resultant Outputs 18
5 Compartive Study 19
6 Conclusion 20
7 Future Scope 20
References 20

The Adaptive Strategies Improving Web Personalization Using the Tree Seed Algorithm (TSA) 23
P. Srinivasa Rao, D. Vasumathi and K. Suresh
1 Introduction 23
2 Motivation 25
3 Problem Statement 26
4 Proposed Work 26
5 Conclusion 28
References 28

An Experimental Evaluation of Integrated Dematal and Fuzzy Cognitive Maps for Cotton Yield Prediction 31
N. Manoharan and Arunkumar Thangavelu
1 Introduction 31
2 Materials and Methods 32
2.1 The DEMATEL Approach 32
2.2 The Model of Fuzzy Cognitive Map 34
3 Results and Discussion 38
3.1 Classification Results 38
3.2 Simulation of FCM Tool 41
4 Conclusion 42
References 43

A Refined K-Means Technique to Find the Frequent Item Sets 45
A. Sarvani, B. Venugopal and Nagaraju Devarakonda
1 Introduction 45
2 Calculation of Minkowski Distance Measure 46
3 Clustering Methods 47
3.1 Hierarchical Methods 47
3.2 Partitioning Methods 47
3.3 Model-Based Clustering Methods 48
3.4 Grid-Based Methods 48
4 K-Means Clustering 48
5 GSP Algorithm 49
6 Methodology 50
6.1 Acquiring the Data 50
6.2 Data Pre-processing 50
6.3 Applying the Refined K Means 50
6.4 Applying the GSP Algorithm 51
7 Experiment 51
8 Conclusion 53
References 53

Finger Vein Detection Using Gabor Filter and Region of Interest 55
Saritha Reddy Venna, Suresh Thommandru and Ramesh Babu Inampudi
1 Introduction 55
2 Related Work 58
3 Proposed Work 58
4 Acquiring the Image 59
5 Image Enhancement 60
6 Feature Extraction 60
6.1 Region Growing 60
6.2 Image Histogram 61
6.3 Gabor Filter and Region of Interest (ROI) 62
7 Experimental Results 63
8 Conclusion 64
References 64

Design of Rheumatoid Arthritis Predictor Model Using Machine Learning Algorithms 67
S. Shanmugam and J. Preethi
1 Introduction 67
2 Related Works 70
2.1 Survey Based on Clinical Research 70
2.2 Survey Based on Machine Learning Algorithms 72
3 Objectives 73
4 Proposed Architecture 73
5 Conclusion and Future Scope 76
References 76

Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic 79
B. Vamshi Krishna, Ajeet Kumar Pandey and A. P. Siva Kumar
1 Introduction 79
2 Related Works 81
3 Research Background 82
4 Proposed Work 85
5 Results and Discussion 86
6 Conclusion 88
References 88

Hexagonal Image Processing and Transformations: A Practical Approach Using R 91
E. Ramalakshmi and Neeharika Kompala
1 Why Is R Used in Processing and Transformations 91
2 Hexagonal Grid: Compendium 92
3 Structure and Addressing 93
3.1 Tessellations 93

4 Operations On The Grid 94
4.1 Construction of Hexagonal Pixels 94
5 Simulations 96
6 Results 97
7 Conclusion 98
References 98

EEG Based Emotion Recognition Using Wavelets and Neural Networks Classifier 101
S. Thejaswini, K. M. Ravi Kumar, Shyam Rupali and Vijayendra Abijith
1 Introduction 101
2 Related Work 102
3 Methodology 103
3.1 SEED Database 104
3.2 Feature Extraction and Selection 105
3.3 Classification 107
3.4 Confusion Matrix 108
4 Results 109
5 Conclusion 111
References 111

Measurement of Disease Severity of Rice Crop Using Machine Learning and Computational Intelligence

Prabira Kumar Sethy, Baishalee Negi, Nalini Kanta Barpanda, Santi Kumari Behera and Amiya Kumar Rath

Abstract This study was conducted to develop a prototype which computes the severity of diseases appears in the rice crop using machine learning and computational intelligence. The symptoms of rice crop diseases imply the seriousness of the disease and suggest choosing the best approach to dealing with the disease. Most of the diseases in rice crop appear as a spot on the leaves. It is also needful to diagnose the disease properly and on-time to avoid the great harm of the rice crop. The treatment of rice crop diseases by applying disproportionate pesticides increases the cost and environmental pollution. So the use of pesticides must be minimized. This can be actualizing by targeting the diseased area, with the appropriate quantity and concentration of pesticide by estimating disease severity. This paper introduces Fuzzy Logic with K-Means segmentation technique to compute the degree of disease severity of leaves in rice crop. The proposed method estimated to give up to about 86.35% of accuracy.

Keywords Disease grading · K-means clustering · Fuzzy logic
Percentage of infection · SVM

1 Introduction

Agriculture has become a main source of life to feed ever growing population. So, plants are playing a key role in our society and a fundamental piece of the puzzle to solve each and every issue. There are many diseases that hamper the growth and productivity plants which lead to great ecological and economical losses. For this reason, it is better to diagnose diseases accurately and measure the severity of the disease to avoid such losses timely. The detection of plant disease can be done in many ways including manual and computer based systems. Manual measurements are visually made by plant pathologists. So, they must be used their experience to carefully manner. Such kind of process can be costly, lengthy and tiresome and can also leads to errors due to fatigue. So, new techniques are like image processing, computer visions are used to get more accurate with consuming less time and at an

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_1

affordable cost [1]. In this paper, introduce an approach to measure the severity of disease by the automatically grading system. For automatic grading purpose, Fuzzy Logic is implemented [2]. In the field of disease management, grade of the disease is determined to provide a precision and accurate treatment advisory. The results are getting from the automatically grading system using Fuzzy Logic which proved to be accurate and satisfactory in contrast with manual grading. K-means based color image segmentation technique is used for segmentation and from that segmented image features will be extracted as well as infected area and disease grade will be evaluated. SVM is used as a classifier in this paper to identify the risk of that disease.

2 Literature Review

The disease appear on leaves can be measured by use of several quantification methods.

i. Sannakki et al. [2] proposed a technique to measure disease based on Fuzzy logic. They demonstrate by use of pomegranate leaves. At first they convert the samples image to L*a*b* color space. Then by application of K-Means clustering pixel are grouped in certain classes. They also suggest a disease grading system based on Fuzzy Logic.

ii. Rahul S. Phadatare and Sanjay S. Pawar [3] proposed a technique to determine as well as quantifying the leaf disease based on image processing, ANN and Fuzzy logic. Here ANN and Fuzzy logic are used for classification and Grading purposes respectively.

iii. Huang et al. [4], apply Canopy Spectral Data Analysis for measuring disease severity. They use ASD field spec pro spectrometer fitted with 25° field for spectral measurement. The disease index is calculated using Eq. (1) given below,

$$DI(\%) = \frac{\sum(x * f)}{n * \sum f} * 100 \tag{1}$$

where ‘f’ is the total number of leaves of each degree of disease severity, ‘x’ is incidence level and ‘n’ is the light incidence level.

iv. Rashedul Islam and Md. Rafiqul Islam [5], they proposed a method to measure the severity of disease found on paddy leaf. First, the input leaf image is segmented using K-Means segmentation, then the cluster contains healthy and diseased portion will be converted to a binary image. From that binary image, white pixel contains by the both cluster has been calculated because that white pixel will help to calculate a number of pixels affected by the leaf disease. The percentage of the infected pixel can be calculated by Eq. (2).

$$P_A(\%) = \frac{wp_a}{P_l} * 100 \tag{2}$$

$$\text{where}\, P_l = wp_a + wp_u$$

v. Powbunthorn et al. [6] experimented on brown spot leaf disease in cassava crop. As per their method the RGB of sample image is transformed to HIS color space, then the infected pixels are extracted by differentiating the Hue. The disease severity is calculate using Eq. (3).

$$PI = \frac{A_d}{A_t} * 100 \tag{3}$$

where A_d = Area affected by disease, A_d = Total Area

vi. Bharambe et al. [7], proposed a method to measure the severity of leaf disease of ground nut. Here they apply Geometric moment to calculate degree og disease severity.

vii. C. H. Bock, G. H. Poole [8], use imaging spectroscopy to determine the disease severity. A large amount of data order of hundreds Mb per image is pre-processed and data is proceeded to take many forms as per its similarity. The popular classification technique is supervised classification which is based on statistical similarity.

viii. Sanjay B. Patil et al. [9], used Triangle method of the thresholding to measure the severity of brown spot leaf disease found in sugar cane crop. The severity of the diseased plant leaves is measured by the leaf area and lesion area ratio. The disease severity can be expressed in below Eq. (4).

$$S = \frac{A_d}{A_l} \tag{4}$$

where A_d = Diseased Leaf Area, A_l = Total Leaf Area

ix. Jayme Garcia Arnal Barbedo [10], presented a method to detect and measure the severity of leaf disease using digital image processing. The proposed image analysis method is based on widely used morphological mathematical operations. Also by the use of the ‘a’ channel of the L*a*b* color representation, which made it possible to derive general rules that hold true even when the leaves and symptoms have different shade and hue characteristics. They suggest that their method is simple to implement and is not computationally complex.

x. P. Saranya et al. [11], presented a method to measure the severity of fungi caused disease in leaf using image processing technique. Triangle thresholding and Simple threshold methods are used to segment the leaf area and disease region area correspondingly. They suggest that image processing tools to measure plant disease severity are suitable and correct because this eliminates the subjectivity of usual methods and person induced errors. The disease severity extent can be expressed in Eq. (5).

$$\mathrm{DSE} = \mathrm{LA}/\mathrm{AL} \tag{5}$$

where DSE = Disease Severity Extent, LA = Lesion Area, AL = Total Leaf Area

3 Proposed Methodology

This paper presents an efficient approach of fuzzy logic and SVM to quantify the disease severity accurately. The flow graph of this system is represented in Fig. 1.

The step below explained the proposed disease grading and risk recognition system:

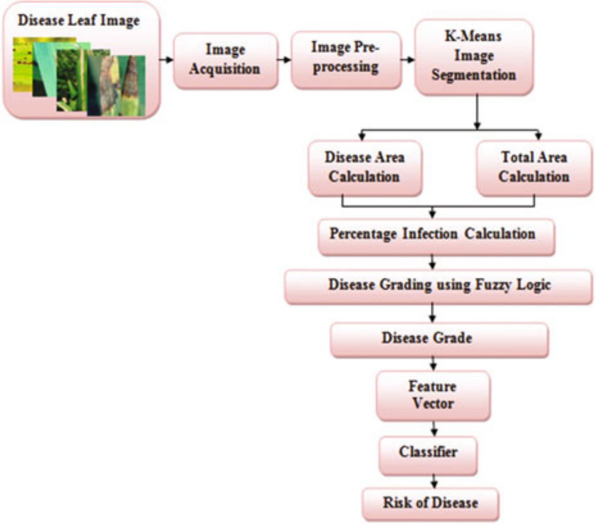


Fig. 1 Proposed disease grading and risk recognition system

(1) Image Acquisition considers as the first step for the proposed methodology. First, take the input leaf image of the rice crop has captured by the camera. The input image is in RGB (Red, Green, and Blue) form. Then, the RGB leaf image is converted into suitable color space as per the requirements.

(2) In Preprocessing phase, the query image converted to suitable color space i.e. L*a*b color space on which the algorithm can be worked. The required information from the image extracted more efficiently by image resizing and contrast enhancement.

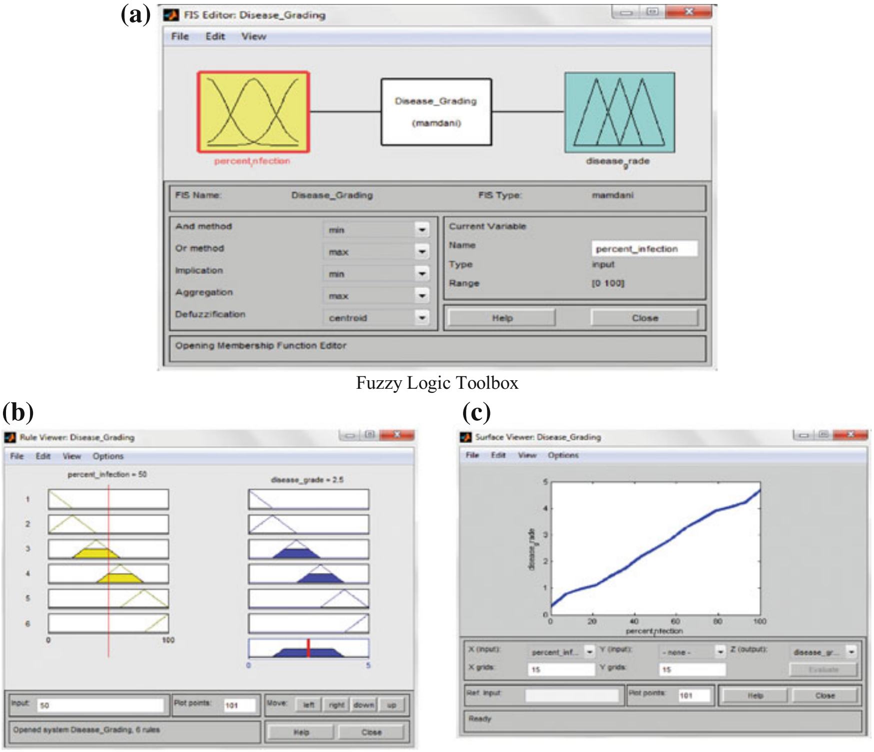
(3) This step includes the segmentation of an image using K-Means algorithm. It is quite a helpful method for detection of the object which is based on a set of features into K number of classes [12]. By minimizing the sum of the squares of the distance between the corresponding cluster and the object it can able to detect the interesting part of the input image. In K-Means clustering techniques, the clusters are determined by the grouping of pixels having the same value present in an image. Practically, the computational speed of this new image processing technique is very fast as well as gives more accurate output. The input dataset is partitioned into K number of clusters and each cluster is considered by a cluster center which is adaptive by nature. Initially considered values are known as seed-points and inputs are also known as data points. Estimation of the distances between the centers, inputs, and allocate inputs to the nearest center is only possible by using K-Means clustering technique.

(4) After the successful execution of K-Means algorithm, we can calculate the total leaf area (A_t) as well as diseased area (A_d) of the leaf; the percentage of infection (P) is calculated by the following equation.

$$P = \frac{A_d}{A_t} * 100 \tag{6}$$

(5) Fuzzy logic builds on and start with set of user supplied human language rules. Later the user supplied rules are converted into mathematical equivalents. So as a user input, percentage of infection has given to the design Fuzzy system. Again the Fuzzy system consists of three parts i.e. Fuzzier, Inference System and Defuzzifier. In Fuzzifier the user input or percentage of infection is converted to a set of fuzzy input values which consist of some membership function and this fuzzy input values are provided to Inference System. In Inference System, rules are designed according to the user need and also decision has been taken here. Then finally the decision is provided to Defuzzifier, where Defuzzifier is converted the set of fuzzy output values into a single user output value i.e. disease grade.

Fuzzy system is used here because of its flexible nature and conceptually easy to understand.



FIS Name: Disease_Grading
FIS Type: mamdani

And method: min
Or method: max
Implication: min
Aggregation: max
Defuzzification: centroid

Current Variable
Name: percent_infection
Type: input
Range: [0 100]

Opening Membership Function Editor

Fuzzy Logic Toolbox

percent_infection = 50
disease_grade = 2.5

Input: 50
Plot points: 101
Move:

Opened system Disease_Grading, 6 rules

X (input): percent_inf...
Y (input): - none -
Z (output): disease_gr...
X grids: 15
Y grids: 15
Ref. Input:
Plot points: 101

Ready

Rules for Disease Grading

Surface View of Disease Grading

Fig. 2 Fuzzy logic toolbox for disease grading

From the research paper [13], it conclude that in precision agriculture, the fuzzy logic permeate in several application of agriculture sector such as texture analyser, grading system and herbicide sprayer. Here we use fuzzy logic for grading of leaf disease in rice crop which shown in Fig. 2.

(6) Feature extraction plays a vital role in image classification. It allows representing the content of the image as perfect as possible. In feature extraction phase, GLCM (Gray Level Co-occurrences Matrix) is used to extract features from the segmented image [14].

(7) Support Vector Machine is used in this paper to identify the risk of disease. SVM is a machine learning technique which is basically used for classification. It is a kernel based classifier; initially, it was developed for linear separation which has able to classify data into two classes only, now it can be used as multi-class SVM. SVM has been used for different realistic problems [15]. The extracted features, the percentage of infection and disease grade is given as input to the SVM. On the basis of comparison between the training data and testing data, the result of risk will declare.

4 Experimental Results

For experimental purpose different disease infected leaf samples has been taken. After the successful computation of the algorithm in MATLAB software, the percent of affected area and disease grade will be observed by using Fuzzy Logic Toolbox which is illustrated in Table 1.

Grade stage of the disease is decided by using Table 2 [16]. According to their percentage of infection, the risk will be decided.

After the measurement of the percentage of infection and grading of disease using the fuzzy system, stage of grade or risk can be predicted by using SVM. SVM gives the percentage of infection and disease grade, by the features extracted from segmented image as input. The result will be decided on the basis of comparison between the test samples with respect to the train data. The result of input pre-processed image, segmented image and classifier for risk management will present in Figs. 3 and 4.

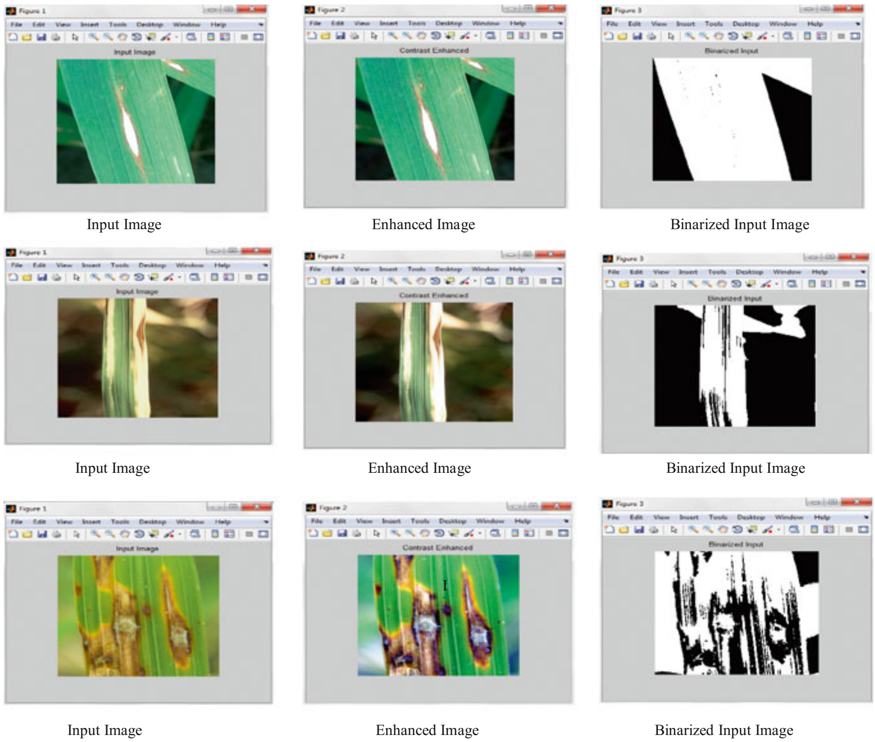
The accuracy of the proposed methodology for severity measurement of individual disease will be illustrated in Table 3. In our demonstration we have taken 4 number of sample image of four different type of disease that is in a whole 16 number of samples are taken. And the accuracy for brown spots is 85.71%, Bacterial Blight is 86.02%, Leaf Scald is 86.56%, Leaf Blast is 87.12% and the average accuracy is 86.35% (Fig. 5).

Table 1 Percentage of infection and grade of disease result per disease for proposed method

Sl no.	Dimension of sample image	Type of disease	Percent of affected area (%)	Disease grade
1	363 × 319 × 3	Brown spots	15.017	0.964
	354 × 360 × 3		15.0015	0.94
	494 × 563 × 3		16.0016	0.976
	300 × 349 × 3		19.1881	0.998
2	498 × 387 × 3	Bacterial blight	15.0016	0.964
	355 × 256 × 3		21.001	1.07
	640 × 428 × 3		10.9182	0.891
	640 × 428 × 3		16.1408	0.977
3	310 × 199 × 3	Leaf scald	15.001	0.964
	509 × 541 × 3		31.0021	1.54
	937 × 640 × 3		12.0278	0.915
	296 × 239 × 3		18.003	0.993
4	1595 × 1344 × 3	Leaf blast	21.4305	1.1
	491 × 462 × 3		15.7133	0.973
	300 × 324 × 3		7.1156	0.778
	395 × 432 × 3		5.004	0.688

Table 2 Disease scoring scale for leaves

Class	Risk	Percentage of infection (%)
1	Very low risk	Between 1–10
2	Low risk	Between 10–20
3	Medium risk	Between 20–30
4	High risk	Between 30–50
5	Very high risk	Between 50–100



Input Image Enhanced Image Binarized Input Image

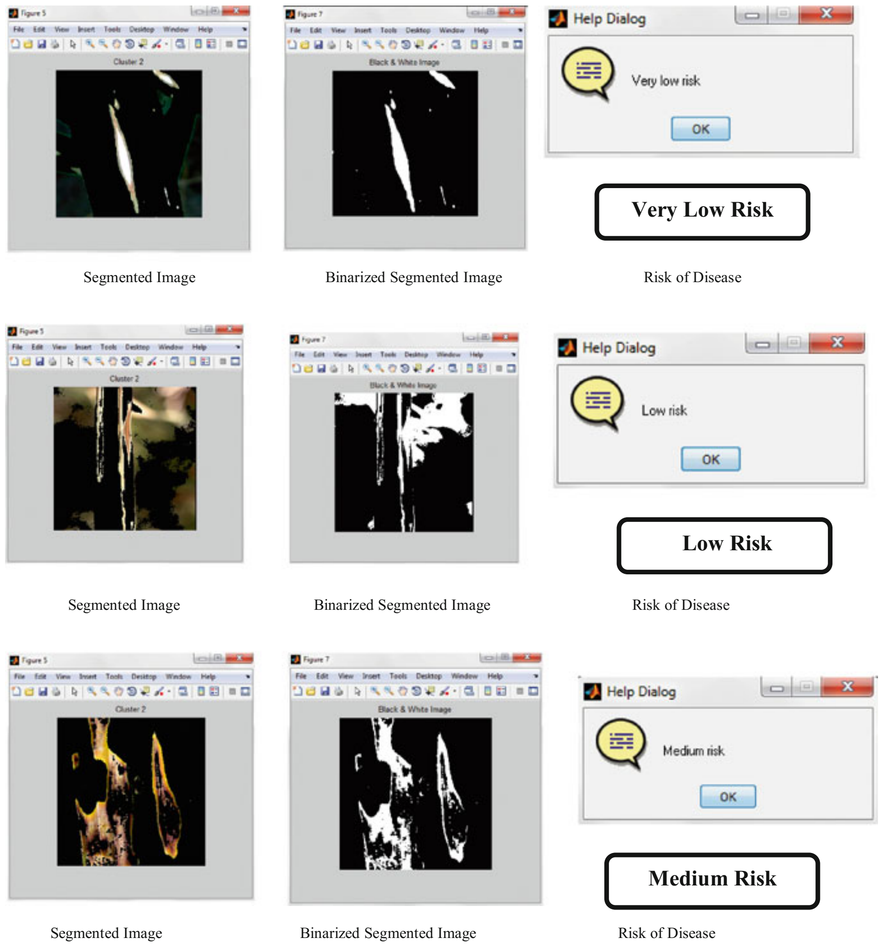
Input Image Enhanced Image Binarized Input Image

Input Image Enhanced Image Binarized Input Image

Fig. 3 Disease input and pre-processed images

5 Conclusion and Future Scope

This paper represents a prototype for leaf disease severity measurement and grades the leaf disease using Fuzzy Logic. The proposed grading system comprises of fuzzy logic and machine vision tool for estimating the severity of leaf disease in rice crop and implies 86.35% accuracy. The method is not computationally complex as



Very low risk

Very Low Risk

Segmented Image

Binarized Segmented Image

Risk of Disease

Low risk

Low Risk

Segmented Image

Binarized Segmented Image

Risk of Disease

Medium risk

Medium Risk

Segmented Image

Binarized Segmented Image

Risk of Disease

Fig. 4 Segmented and classified image as per disease severity/risk

Table 3 Accuracy result per disease for proposed algorithm

Sl no.	Name of disease	Percentage of accuracy
1	Brown spots	85.71
2	Bacterial blight	86.02
3	Leaf scald	86.56
4	Leaf blast	87.12
Average accuracy		86.35

well as is easy to implement. Also, this system gives a fast and accurate grading of disease severity and predicts the risk of disease into different stage as compared to manual method. The study may be extended by considering more types of disease and with large data set.

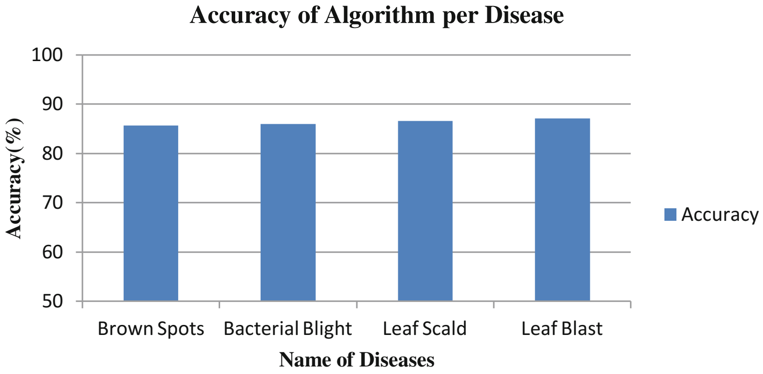


Fig. 5 Accuracy result per disease for proposed algorithm

References

1. Barbedo, Jayme Garcia Arnal (2013). Digital Image Processing Techniques for Detecting, Quantifying and Classifying Plant Diseases, SpringerPlus.
2. Sannakki, Sanjeev S. et al. (2011). Leaf Disease by Machine Vision and Fuzzy Logic. *International Journal of Computer Applications* 2 no. 5: 1709–1716, ISSN: 2229-6093.
3. Phadatare, Rahul S., and Sanjay S. Pawar. (2016). Leaf Disease Detection and Grading using Image Processing. *International Journal for scientific research & Development* 4 no. 9, ISSN: 2321-0613.
4. Huang, Wenjiang, Qingsong, Guan et al. (2014). New Optimized Spectral Indices for Identifying and Monitoring Winter Wheat Diseases. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 no. 6.
5. Islam, Rashedul, and Md. Rafiqul, Islam. 2015. An Image Processing Technique to Calculate Percentage of Disease Affected Pixels of Paddy Leaf. *International Journal of Computer Applications* (0975–8887) 123, no. 12.
6. Powbunthorn, Kittipong, Wanrat, Abudullakasim, and Jintana Unartngam. (2012). Assessment of the Severity of Brown Leaf Spot Disease in Cassava using Image Analysis. *The International conference of the Thai Society of Agricultural Engineering*.
7. Bharambe, Chandan J., Vidya N. More, Sumeet S. Nisale. (2011). Detection and Analysis of Deficiencies in Groundnut Plant using Geometric Moments. *World Academy of Science, Engineering and Technology International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering* 5, no. 10.
8. Bock, C.H., and G.H. Poole. (2010). Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis. *Reviews in Plant Sciences* 29, no. 2: 59–107. https://doi.org/10.1080/07352681003617285.
9. Patil, Sanjay B. et al. 2012. Leaf Disease Severity Measurement using Image Processing. *International Journal of Engineering and Technology* 3 no. 5: 297–301.
10. Barbedo, Jayme Garcia Arnal. 2014. An Automatic Method to Detect and Measure Leaf Disease Symptoms using Digital Image Processing. *APS Journal* 98, no. 12. http://dx.doi.org/10.1094/PDIS-03-14-0290-RE.
11. Saranya, P., S. Karthick, and C. Thulasiyammal. 2014. Image Processing Method to Measure the Severity of Fungi Caused Disease in Leaf. *International Journal of Advance Research* 2, no. 2: 95–100.

12. Sethy, Prabira, Baishalee Negi, and Nilamani Bhoi. 2017. Detection of Healthy & Defected Diseased Leaf of Rice Crop using K-Means Clustering Technique. *International Journal of Computer Applications* 157, no. 1: 0975–8887.
13. Kavdir, Ismail, and Daniel E. Guyer. (2003). Apple Grading Using Fuzzy Logic. *Turkish Journal of Agriculture and Forestry,* 375–382 © T. BÜTAK.
14. Gebejes, A., and R. Huertas. 2013. Texture Characterization based on Grey-Level Co-occurrence Matrix, ICTIC.
15. Byun, Hyeran, and Seong-Whan Lee. 2002. Applications of Support Vector Machines for Pattern Recognition—A Survey. Springer, SVM 2002, LNCS 2388, pp. 213–236.
16. Standard Evaluation System for Rice. 2015. *International Rice Research Institute (IRRI)*, 5th ed.

Flue-Cured Tobacco Leaves Classification: A Generalized Approach Using Deep Convolutional Neural Networks

Siva Krishna Dasari, Koteswara Rao Chintada and Muralidhar Patruni

Abstract In this paper, a solution is defined based on convolutional neural networks (CNN) for the grading of flue-cured tobacco leaves. A performance analysis of CNN on 120 samples of cured tobacco leaves is reduced from 1450 × 1680 Red-Green-Blue (RGB) to 256 × 256, consisting 16, 32 and 64 feature kernels for hidden layers respectively. The neural network comprised of four hidden layers where the performance of convolution and pooling on first three hidden layers and fourth layer a fully connected as in regular neural networks. Max pooling technique (MPT) is used in the proposed model to reduce the size. Classification is done on three major classes' namely class-1, class-2 and class-3 for obtaining global efficiency of 85.10% on the test set consisting about fifteen images of each cluster. A comparative study is performed on the results from the proposed model with existing models, state of the art models on tobacco leaf classification.

Keywords Convolutional neural network · Flue-cured tobacco
Max pooling · Multi-layer perceptron · Tobacco leaves classification system

1 Introduction

Tobacco is one of the most successful commercial crops cultivated on this planet. China, India, Brazil and USA are the major producers of tobacco worldwide and these four nations alone contribute around 86% of the global production. India is the top two contributors to the global tobacco production and it's estimated that around 750 M kgs of tobacco is being produced in the area around 0.45 M hectares [1]. Harvested tobacco will undergo a process called curing for further stages of refinement.

There are few methods existing on grading of FCT leaves [2], proposed a method based on nearest neighbors which uses the mean of hue and chroma as color features and factors such as the ratio of leaf length with respect to the width and shape as features [3]. A fuzzy classification system which uses color, texture and

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_2

shape as features to design the model and grade the tobacco leaves using Support Vector Machine (SVM) to classify FCT leaves. There is a model proposed by [4] which uses symbolic data and uses Munsell coloring system to extract average and standard deviation of hue as color features.

In this work, a solution is proposed based on CNN for the grading of tobacco. As CNN have much fewer connections and parameters compared to standard feed forward neural networks with similar sized layers, it's easy to train them. They can handle high dimensional inputs cleverly. The feature selection is done unsurprisingly without explicitly stating the features with the help of filter kernels. They have many advantages as they adjust the filter weights to recognize features such as color, shape and size.

2 Convolutional Neural Networks

2.1 Convolutional Layer (CL)

In CNN each neuron doesn't sense the total image as in MLP which is fully connected and process all the pixels of the image. Every neuron gets connections from a portion of the image (local to particular neuron) which makes CNN faster to train (Fig. 1). The connectivity along the depth axis is same as the depth of the input volume (number of color channels).

2.2 Pooling Layer

Pooling Layer reduces the size of the input volume there by reducing the number of parameters and computation. It controls over-fitting of the data. While resizing, it operates on each depth slice independently. There are different methods of doing pooling, the proposed model uses MAX Pooling (MP) [5]. If a p × p filter is used for MP, then it take one maximum value out of all p2 values on the input volume. In pooling depth dimension remains same (Fig. 2).

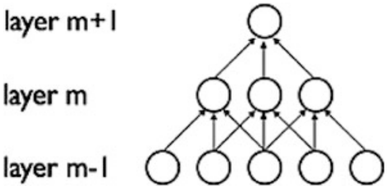
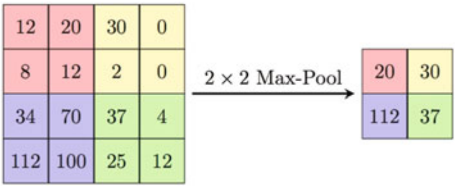


Fig. 1 Sparse connectivity of CNN



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Fig. 2 MAX pooling on a depth slice

2.3 *Back Propagation on Conv Layer*

For an error function E and having the current error value of the layer m, calculate the error values of the previous layer and gradients for each weight. The error for previous layer is partial of E with respect to neuron output $\frac{\partial E}{\partial y_{ij}^{m}}$. Here is the formula for calculating the gradient for each weight by using chain rule [6].

$$\frac{\partial E}{\partial \omega_{ab}} = \sum_{i=0}^{s-t} \sum_{j=0}^{s-t} \frac{\partial y}{\partial x_{ij}^{m}} \frac{\partial x_{ij}^{m}}{\partial \omega_{ab}}$$

$$= \sum_{i=0}^{s-t} \sum_{j=0}^{s-t} \frac{\partial E}{\partial x_{ij}^{m}} y_{(i+a)(j+b)}^{m-1}$$

$$\frac{\partial E}{\partial x_{ij}^{m}} = \frac{\partial E}{\partial y_{ij}^{m}} \cdot \frac{\partial y_{ij}^{m}}{\partial x_{ij}^{m}}$$

$$= \frac{\partial E}{\partial y_{ij}^{m}} \cdot \frac{\partial}{\partial x_{ij}^{m}} \left(\sigma \left(x_{ij}^{m} \right) \right)$$

$$= \frac{\partial E}{\partial y_{ij}^{m}} \cdot \ \sigma^{1} \left(x_{ij}^{m} \right)$$

The error at current layer $\frac{\partial E}{\partial y_{ij}^{m}}$ it's straight forward to compute $\frac{\partial E}{\partial x_{ij}^{m}}$ at present layer by knowing the first derivative of the activation function, i.e. $\sigma^{1}\left(x_{ij}^{m} \right)$.

$$\frac{\partial E}{\partial \omega_{ab}} = \sum_{i=0}^{s-t} \sum_{j=0}^{s-t} \frac{\partial y}{\partial x_{ij}^{m}} \frac{\partial x_{ij}^{m}}{\partial \omega_{ab}}$$

$$= \sum_{i=0}^{s-t} \sum_{j=0}^{s-t} \frac{\partial E}{\partial x_{ij}^{m}} y_{(i+a)(j+b)}^{m-1}$$

Therefore, the error at the current layer computes the gradient with respect to weights on the layer, using chain rule to back propagate.

3 Tobacco Leaves Classification System

3.1 Convnet Architecture and Model

Tobacco Leaves classification system (TLCS) is based on deep CNN. Torch 7 Frame work in deep learning for building the system [7]. The network consists a total of nine layers. The full architecture of this model is depicted as shown in the Fig. 3 with the exclusion of Input and Output layers. There are total seven layers, out of which three convolutional, three pooling and one fully connected layer.

The design depth of the network is as follows:

Input Layer which is followed by a CL Conv1, Pooling Layer Pool1, CL Conv2, Pooling Layer Pool2, CL Conv3, Pooling Layer Pool3, a Rectified Linear Unit (ReLU) unit sandwiched between the Convolutional and Pooling Layers as a fully CL and the Output Layer.

In this model the following hyper parameters filter size or receptive field of 5, Pooling filter of size 2, with a stride of 1 and non Zero-padding applied. The CL Conv1 which filters the 256 × 256 × 3 input image with 16 Convolutional Kernels each of size 5 × 5. Total number of neurons in this layer is 252 × 252 × 16.

The Pooling Layer Pool1 which filters the input volume from Conv1 16 Pooling Kernels each of size 2 × 2.

Total number of neurons in this Layer is 126 × 126 × 16.

There is ReLU unit which works like an activation function applied on the output volume of the CL [8]. This outputs the same dimension as of Input volume with activation function applied on each pixel of the Input volume.

The output volume of Pool1 is fed into Conv2 Layer which has 32 filters, each filter operates on some portion of 126 × 126 × 16 input volume.

This Layer has a total of 122 × 122 × 32 neurons. Pool_2 Layer which filters the output of ReLU with 32 filters has a total of 61 × 61 × 32 neurons.

Conv3 Layer which filters the output of Pool2 functions with 64 filters has a total of 57 × 57 × 67 neurons. Pool3 Layer which filters the output of ReLU functions with 64 filters has a total of 28 × 28 × 64 neurons. The output volume of Pool3 is

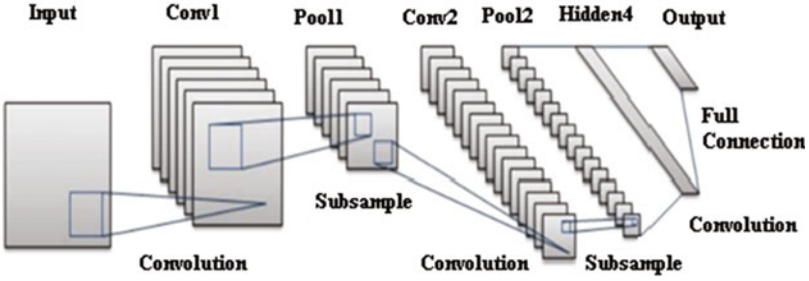


Fig. 3 Convolutional neural network architecture of tobacco leaves classifier

being fed into a fully connected layer having 128 neurons. All the outputs of fully connected layers are piped into the output layer which is a three way classifier with three neurons.

3.2 Details of Learning

The model is trained using gradient descent with a batch size of 1, weight decay of 0.0001 and with a momentum of 0.5.

Initialized learning rate is 0.01.

The generated formula for weights becomes

$$\vartheta_{i+1} = 0.5V_i - 0.0001 \cdot \in \cdot \, \omega_i - \in \cdot < \frac{\partial \mathrm{L}}{\partial \omega} |\omega_{\mathrm{i}} > {}_{\mathrm{D_i}}$$

$$\omega_{i+1} = \omega_i + \vartheta_{i+1}$$

where

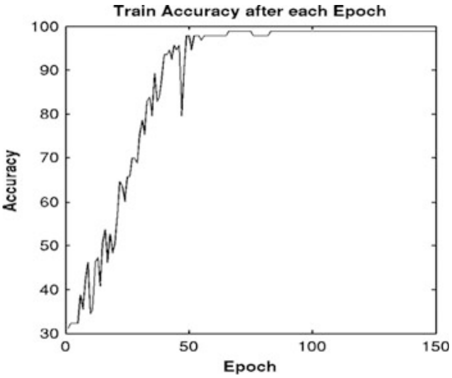
i Iteration index
$\in$ Learning rate
ϑ Momentum variable

$< \frac{\partial L}{\partial \omega} |\omega_i > {}_{D_i}$: is the average over ith batch $\mathrm{D_i}$ of the derivative of the objective with respective ω valuated at ω_i [9].

The network is trained on 120 images and It took approximately 4 h for 300 epochs on a machine with 2.3 GHz Intel Quad Core Processor and 3 GB of Main Memory.

Figure 4 illustrates the training accuracy over the number of epochs.

Fig. 4 Training accuracy over the number of epochs



4 Resultant Outputs

The proposed model on the test data set consisting of 47 image samples belonging to three different classes i.e. (class-1: 15 samples, class-2: 17 samples class-3: 15 samples). It took on average about 68.46 ms to test or classify each test sample.

A global correctness of 85.10% on 47 samples. Here is the Confusion matrix for the three classes.

$$\begin{array}{c} \begin{array}{ccc} \text{class1} & \text{class2} & \text{class3} \end{array} \\ \begin{pmatrix} 13 & 0 & 2 \\ 0 & 14 & 3 \\ 0 & 2 & 13 \end{pmatrix} \end{array}$$

Out of 15 samples of class-1, two samples are incorrectly classified into class-3. Out of 17 samples of class-2, 3 samples are incorrectly classified into class-3 from the 15 samples of class-3 two are incorrectly classified into class-2. Here is the bar chart depicting in Fig. 5.

For the test sample plotted a graph for the percentage accuracy with respect to number of epochs. At around 100 epochs the accuracy get saturated and have minimal fluctuations after that. This is represented in the Fig. 6.

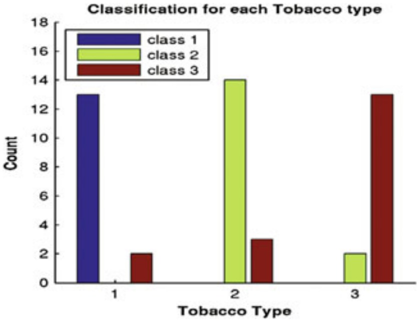


Fig. 5 Test sample correctness

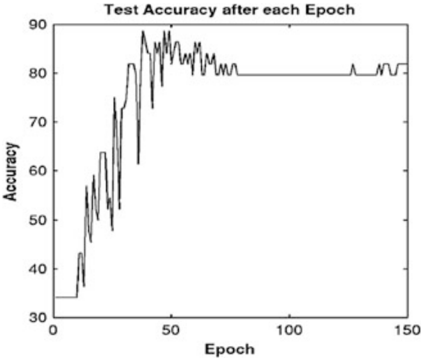


Fig. 6 Accuracy plot for a number of epochs

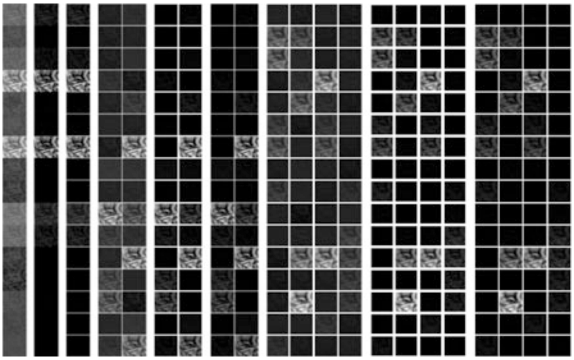


Fig. 7 A sample input volume as it passes through the layers

Figure 7 shows the outputs from different layers as the test sample passes through each layer (Conv Layer, ReLU Unit, Pooling Layer).

Columns-1, 2, 3 are convolutional, ReLU and Pooling Layers respectively with 16 filters. Columns-4, 5, 6 having 32 filters each and Columns-7, 8, 9 having 64 filters each.

5 Compartive Study

In the proposed system, the model takes 120 images for training sample. The existing models trained on more number of samples than considered in the proposed model and did more quality samples taken with better equipment.

Description: Table 1 shows that the proposed system produced 85.10% accuracy measure, though it is less in measure the concept of image enhancing in this work is scheduled with 120 as a size which surprisingly gave good result by considering back propagation.

Table 1 Comparison of different models with the proposed model

Title	Sample size	Classification technique	Accuracy (%)
A trainable grading system for tobacco leaves [2]	110	NNC	64
Proposed model (tobacco leave classification system using max pool technique)	120	DCNN	85.10
Recognition of the part of growth of flue cured tobacco leaves on support vector machine [10]	1712	SVM	86.62
Min-max representation of features for grading cured Tobacco leaves [4]	887	SC	87.18

6 Conclusion

Proposed TLCS performance even on limited training samples (image data set) compared to existing models. The proposed model achieved a global correctness of 85.10%. On application, the state of art image classification technique CCN for the model. This can be applied on samples in which it's hard to describe the features. As disclosed in Table 1, though the performance measure of the proposed model is next to [10]. It is clearly understood that the sampling in the above references handled huge clusters of image data whereas in the proposed model, it is considered strongly for 120 images. Here, the max pooling technique had shown its effect in abundant reduction of image size by reducing the unwanted kernel features of the respective hidden layers. The CL itself produced more betterment results compared with the three individual hidden layers. The major difference between the other models and the proposed model is that, the secure data of hidden layer could be accessed at any levels of all the algorithms whereas in this study it is clearly informed that there is no passage of data transfer between CL and hidden layers. This study gives us the parametric results as 85.10% which is accurate and appropriate both technically and theoretically as per the samples of tobacco data collected.

7 Future Scope

This model is further extended for application among larger training sample set to achieve better results. The result obtained through deep learning algorithms could be enhanced by using some optimization algorithms to the images which are considered. Consistent images i.e., with huge information and with less noise are considered to enrich the outputs at all levels.

References

1. CTRI-Rajahmundry. Tobacco in Indian Economy. http://www.ctri.org.in/fortobaccoEconomy.php 2015.
2. Zhang, J., et al. 1997. A Trainable Grading System for Tobacco Leaves. *Computers and Electronics in Agriculture* 16 (3): 231–244.
3. LeCun, Yann, et al. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86 (11): 2278–2324.
4. Guru, D.S., et al. 2011. Min-max Representation of Features for Grading Cured Tobacco Leaves. *Statistics and Applications* 9 (1&2): 15–29.
5. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

6. Boureau, Y-Lan, Jean Ponce, and Yann LeCun. 2010. A Theoretical Analysis of Feature Pooling in Visual Recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), 111–118.
7. Prasad, V., T.S. Rao, and M. Babu. 2016. Thyroid Disease Diagnosis Via Hybrid Architecture Composing Rough Data Sets Theory and Machine Learning Algorithms. *Soft Computing* 20 (3): 1179–1189.
8. Prasad, V., Rao, T. S., and Reddy, P. P..2016. Improvised Prophecy Using Regularization Method of Machine Learning Algorithms on Medical Data. *Personalized Medicine Universe* 5, 32–40.
9. Lawrence, Steve, et al. 1997. Face Recognition: A Convolutional Neural-Network Approach. *Neural Networks, IEEE Transactions on* 8 (1): 98–113.
10. LeCun, Yann, et al. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1 (4): 541–551.

The Adaptive Strategies Improving Web Personalization Using the Tree Seed Algorithm (TSA)

P. Srinivasa Rao, D. Vasumathi and K. Suresh

Abstract Web personalization is a method of modifying a web site to the requirements of exact users gaining benefit of information attained for the study of the user's directional conduct in association with other material composed in the web framework. In this education primarily we give request to distinctive search engine what's more, the top n list from each web search tool is picked for all the ore dealing with our technique to fulfill data location period. Finding Large Itemsets: create all blends of things that have a bolster an incentive over a client characterized least support. The support for an itemset is the quantity of exchanges that contain the itemset. These things are called substantial itemsets. We then union the top n list in view of one of a kind connections and we do some parameter computations, for example, title based figuring, piece based estimation, content based count, address based computation, interface based estimation, URL based count and co-event based count. We give the arrangements of the counts with the client given the positioning of connections to the fluffy bat to prepare the framework. The framework then positions and unions the connections we get from various web indexes for the question we give. In this paper the reaction time in view of main fifty connections and hundred connections of our system is better contrasted with the current fluffy strategy and the exactness of our method is high contrasted with the current procedure for the distinctive questions we gave. The computed esteems with the client positioned rundown are given to the fluffy bat to rank the rundown. The calculated values with the user ranked list are given to the fuzzy-bat to rank the list.

Keywords Web mining · Recommendation system · Fuzzy-bat
Precision and response time

1 Introduction

The hyper fast development of the World-Wide Web postures extraordinary mounting challenges for inquiry engines. In the advanced time of great volume data era, internet searcher ends up being a crucial technology of information mining and

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_3

data recovery. Universally useful web crawlers [1] have accomplished a lot of achievement in giving significant data to the client. They used to be a viable instrument for retrieving data from the tremendous data archive. Case in point, Google, which is one of the prominent web crawlers, not just gives fitting indexed lists to the client on the planet by pack up of more than 20 hundred millions website pages, furthermore a perfect chance to chase is not for the most part recent second [2]. The inescapability of the Internet and Web has provoked the ascent of a couple web records with evolving capacities. These web crawlers list Web districts, pictures, Usenet news clusters, content-based indexes, and news sources with the goal of conveying inquiry things that are most correlated to customer request. Regardless, only somewhat number of web customers truly know how to utilize the bona fide force of web indexes. Keeping in mind the end goal to address this issue, web indexes have begun giving access to their administrations by means of different interfaces [3].

Web crawler as an apparatus to research the Web must acquire the fancied results for any given inquiry. Accomplishment of a web index is straightforwardly subject to the fulfillment level of the client. Clients seek the information to be exhibited to them inside a brief span interim. They additionally expect that the most significant and late data to be exhibited [4]. A large portion of the web crawlers can't totally fulfill client's requirements and the query.

ACO, Kernel SVM, Recommendation System, Precision, Response Time items are frequently extremely mistaken and unimportant [3, 5]. There are as of now numerous analysts who have provided details regarding about different parts of web crawlers in [3, 6, 7]. A meta-web searcher is the kind of web hunt device to outfit customers with information organizations and it doesn't have its own particular database of website pages. It sends look terms to the databases kept up by other web seek apparatuses and gives customers the results that begin from all the web files addressed [3]. The deficiency of a particular structure and broad assortment of data disseminated on the web makes it exceedingly striving for the customer to find the data with no external offer assistance. It is a general dependability [8] that a lone all around helpful web list for all web information is impossible since its preparing power can't scale up to the quick expanding and boundless measure of web information. A device that quickly picks up endorsement among clients is Meta web search tools [9]. The Meta internet searchers can run client question over various segment web search tools simultaneously, recover the produced results and amassed them. The advantages of Meta web search tools against the web crawlers are striking [10].

The Meta web index upgrades the hunt scope of the web giving higher review. The cover among the essential web crawlers is by and large little [5] furthermore, it can be little as three rates of the total outcomes recuperated. The Meta web searcher fathoms the flexibility issue of looking the web and supports the usage of different web lists engaging consistency checking [11]. The Meta internet searcher improves the recovery viability giving higher exactness due to 'theme impact' [6]. Web Meta looking for in uniqueness to rank aggregate is an issue addressing its own specific remarkable challenges. The outcomes that a Meta look for structure aggregates from its section engines dislike votes or whatever other single dimensional substances: Apart from the individual situating it is consigned by a section engine, a Web out-come in like manner joins a title, a little bit of substance which addresses its criticalness to the submitted request [7] (abstract piece) and a uniform asset locator (URL). Apparently, the conventional rank aggregation systems are lacking for giving a hearty positioning component fitting for Meta internet searchers, since they disregard the semantics going with every Web comes about.

2 Motivation

Web Usage Mining turns into an essential viewpoint in today's period on the grounds that the amount of information is continuously expanding. Web utilization excavating is the use of general mining methods to discovery use strategies from Web information, with a specific end goal to comprehend and better serve the necessities of Web-based applications. The information gathered in web mining from the customer side, server-side, associations database, intermediary servers. Web log mining is one of the late ranges of exploration in Data mining. Web mining comprehensively isolated into three classes: Content mining: Extract data of substance of web mining. Utilization mining [12]: To break down the connections between pages through the web structure to surmise the learning. Structure mining: Extracting the data from web log record which is gotten to by clients. Web utilization mining comprises of three stages: Preprocessing, Pattern revelation and Pattern investigation. Preprocessing comprises of changing over the utilization, substance, and structure data contained in the different accessible information sources into the information reflections essential for example revelation. In the information pre-preparing, it takes web log information as information and afterward handle the web log information and gives the dependable information. To accomplish its objective Data preprocessing is separated into Data Cleaning, client distinguishing proof, and Session Identification. Once the preprocessing stage is all around performed, we can apply information mining systems like grouping, association, characterization, and so forth for utilizations of web use mining, for example, business knowledge, e-trade, e-learning, personalization, and so on. Design disclosure draws upon strategies and calculations created from a few fields, for example, insights, information mining, machine learning and example acknowledgment. Bunching web information is finding the gatherings which offer normal interests and conduct by dissecting the information gathered in the web servers. Association guideline mining (ARM) is an outstanding combinatorial issue and a standout amongst the most dynamic exploration fields in information preparing. Essentially, it distinguishes down to earth and intriguing conditions between things in an exchange al database to help for basic leadership. Affiliation rules [13] have turned out to be extremely helpful instruments in an endeavor as it endeavors to enhance its aggressiveness and productivity. The inspiration driving example investigation is to sift through uninteresting standards or examples from the set found in the example revelation stage.

3 Problem Statement

The most important methods working with the purpose of the web page recommendation and personalization. The major issue concerned with our existing study is discussed as follows: (1) The infringement of protection is additionally a surely understand issue in Personalized Web Search approach. It creates moral and security issues. Another restriction in this technique is that clients' needs are not static, it changes consistently. And in addition there are a few events, clients do scanning for others needs too. So there is issue to web indexes to recognize these scenarios. (2) A primary downside of personalization on the customer side is that the personalization calculation can't utilize some learning that is just accessible on the server side. Moreover, because of the points of confinement of system data transmission, the customer can normally just process constrained top outcomes. (3) In our work we introduce kernel SVM algorithm [14] will increase a performance and good accuracy result. The above mentioned issues are overcome to carrying out this investigation on the personalized web search and recommendation using the Tree seed Algorithm.

4 Proposed Work

1. Finding Large Itemsets: produce all blends of things that have a bolster esteem over a client characterized least backing. The backing for an itemset is the amount of exchanges that contain the itemset. These things are called broad itemsets [15].
2. Making Association Rules: connection fundamentals are delivered from the discovered broad thing sets. To make association runs all the nonempty subsets for every vast item set are generated.

Consideration the leading equations for the web personalization [16] flow issue and under Tree seed optimization approximation are given by

$$\frac{\partial u^*}{\partial t^*} = \nu \frac{\partial^2 u^*}{\partial y^{*2}} - \frac{\sigma B_0^2}{\rho(1+m^2)}\left(u^* + mw^*\right) + g\beta\left(T^* - T_\infty^*\right) + g\beta^*\left(C^* - C_\infty^*\right) - \frac{\upsilon}{k^*}u^* \tag{1}$$

$$\frac{\partial w^*}{\partial t^*} = \nu \frac{\partial^2 w^*}{\partial y^{*2}} + \frac{\sigma B_0^2}{\rho(1+m^2)}\left(mu^* - w^*\right) - \frac{\upsilon}{k^*}w^* \tag{2}$$

$$\frac{\partial T^*}{\partial t^*} = \frac{k}{\rho c_p}\frac{\partial^2 T^*}{\partial y^{*2}} - \frac{Q_0}{\rho c_p}\left(T^* - T^*_\infty\right) - \frac{1}{\rho c_p}\frac{\partial q_r}{\partial y^*} + \frac{Q_1^*}{\rho c_p}\left(C^* - C^*_\infty\right) \tag{3}$$

$$\frac{\partial C^*}{\partial t^*} = D\frac{\partial^2 C^*}{\partial y^{*2}} - K_r^*\left(C^* - C^*_\infty\right) \tag{4}$$

Initial and boundary conditions to be satisfied are as follows

$$t^* \leq 0\text{: } u^* = 0, \quad w^* = 0, \quad T^* = T^*_\infty, \quad C^* = C^*_\infty, \quad \text{for all y}^* \geq 0 \tag{5}$$

$$t^* > 0 : u^* = \mathrm{e}^{a^* t^*}, w^* = 0,\ T^* = \begin{cases} T^*_\infty + \left(T^*_w - T^*_\infty\right)\frac{t^*}{t_0} & \text{at y}^* = 0 \text{ when } 0\ < \mathrm{t}^* \leq t_0, \\ \mathrm{T}^*_w & \text{at y}^* = 0 \text{ when } \mathrm{t}^* > t_0. \end{cases}$$

$$when \quad t^* > 0, \quad C^* = C^*_\infty + \left(C^*_w - C^*_\infty\right)At^* \quad at\,\mathrm{y}^* = 0 \tag{6}$$

$$t^* > 0\text{: } u^* \to 0, \quad w^* \to 0, \quad T^* \to T^*_\infty, \quad C^* \to C^*_\infty, \quad \text{as} \quad \mathrm{y}^* \to \infty \tag{7}$$

For an optically low, the web pages q_r is approximated by Roseland approximation which is given

$$q_r = -\frac{4\sigma^*}{3k^*}\frac{\partial T^{*^4}}{\partial y^*} \tag{8}$$

The adequacy of the recovery assessment of our system is contrasted and the current fluffy method. The assessment we done in view of fifty clients i.e. we gave the questions we utilized for our assessment to fifty clients and they took the main ten records from each web crawlers we utilized and blended it in view of the exceptional connections and positioned the connections in light of the association [17] with the inquiry and their acumen. In the long run, the positioned arrangements [18] of the fifty clients are changed over to a solitary positioned rundown to play out the assessment of our system. The top rundown and important archives of the query when our technique is applied (Table 1).

Table 1 Relevant document in the top ten lists for the query "data mining techniques"

Engine	1	2	3	4	5	6	7	8	9	10	R
Our technique	R	R	–	R	R	–	R	R	R	R	8
Old technique	R	R	R	R	–	–	R	R	–	R	7
Google	R	R	–	R	R	–	R	–	R	–	6
Bing	R	R	R	–	R	–	–	R	R	–	6
Yahoo	R	–	R	R	–	R	–	R	–	R	6

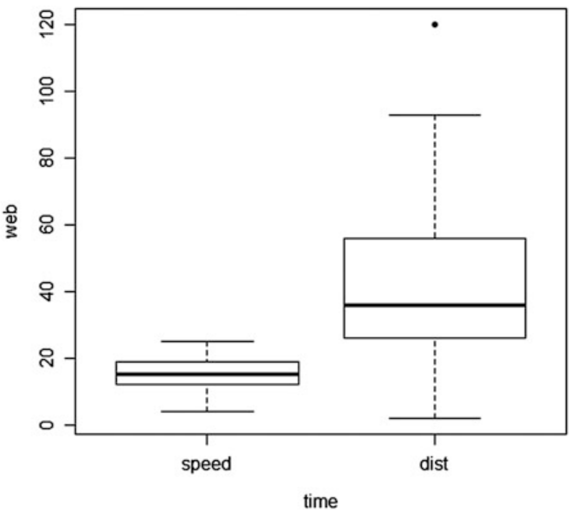


Fig. 1 Precision assessment for the user given query "Data mining" when top most links are taken

The accuracy using user given queries is explained in this section. Figure 1 shows the web sites comparison for the user given time taken for links are taken from each search.

5 Conclusion

The issue of mining affiliation governs in an arrangement of exchanges D can be characterized as the issue of producing all the affiliation decides that have a bolster esteem more noteworthy than a client characterized least support and a certainty esteem more noteworthy than a client characterized least certainty. Much of the time the reaction time in light of main fifty connections and hundred connections of our strategy is better contrasted with the current fluffy procedure and the exactness of our system is high contrasted with the current method for the diverse questions we gave. The computed esteems with the client positioned rundown are given to the fluffy bat to rank the rundown. We compared our technique with the existing fuzzy technique in terms of precision and response time.

References

1. Araus, et al. 2001. Searching the Web. *ACM Transactions on Internet Technology* 1: 243.
2. Tang, Juan, Ya-Jun Du, and Ke-Liang Wang. 2007. Design and implement of personalize meta-search engine based on FCA. In *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 19–22 August 2007.

3. Abawajy, J.H., and M.J. Hu. 2005. A new internet meta-search engine and implementation. In *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*.
4. Satya Sai Prakash, K., and S.V. Raghavan. 2001. *DLAPANGSE: Distributed Intelligent Agent based Parallel Architecture for Next Generation Search Engines*. Department of Computer Science & Engineering, Indian Institute of Technology Madras, India.
5. Aslam, J.A., and M.H. Montague. 2001. Metasearch consistency. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 386–387.
6. Dwork, C., R. Kumar, and M. Naor, D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the ACM International Conference on World Wide Web (WWW)*, 613–622.
7. Lamberti, Fabrizio, Andrea Sanna, and Claudio Demartini. 2009. A relation-based page rank algorithm for semantic web search engines. IEEE Transactions on Knowledge and Data Engineering 21(1).
8. Telang, Aditya, Chengkai Li, and Sharma Chakravarthy. 2012. One size does not fit all: to-ward user- and query-dependent ranking for web databases. *IEEE Transactions on Knowledge and Data Engineering* 24 (9): 1671–1685.
9. Meng, W., C. Yu, and K.-L. Liu. 2002. Building efficient and effective metasearch engines. *ACM Computing Surveys* 34 (1): 48–89.
10. Spink, A., B.J. Jansen, C. Blakely, and S. Koshman. 2006. Overlap among major Web search engines. In *Proceedings of the IEEE International Conference on Information Technology: New Generations (ITNG)*, pp. 370–374.
11. Vogt, C.C. 1999. *Adaptive Combination of Evidence for Information Retrieval*. Ph.D. thesis. University of California at San Diego.
12. Akritidis, Leonidas, Dimitrios Katsaros, and Panayiotis Bozanis. 2011. Effective rank aggregation for metasearching. *The Journal of Systems and Software* 84: 130–143.
13. Ishii, Hideaki, Roberto Tempo, and Er-Wei Bai. 2012. A web aggregation approach for distributed randomized pagerank algorithms. *IEEE Transactions ON Automatic Control* 57 (11): 2703–2717.
14. Durao, Frederico, and Peter Dolog. 2012. *A Personalized Tag-Based Recommendation in Social Web Systems*.
15. Zhou, Zhurong, and Dengwu Yang. 2014. Personalized recommendation of preferred paths based on web log. *Journal of Software* 9 (3): 684–688.
16. Rizvi, N.T.S.H., and Ranjit R. Keole. 2015, January. A preliminary review of web-page recommendation in information retrieval using domain knowledge and web usage mining. *International Journal of Advance Research in Computer Science and Management Studies* 3(1).
17. Yang, Linjun, and Alan Hanjalic. 2012. Prototype-based image search reranking. *IEEE Transactions On Multimedia* 14(3): 871–872.
18. Abrishami, Soheila, Mahmoud Naghibzadeh, and Mehrdad Jalali. 2012. *Web page recommendation based on semantic web usage mining*, vol. 7710 of the series Lecture Notes in Computer Science, pp. 393–405.

An Experimental Evaluation of Integrated Dematal and Fuzzy Cognitive Maps for Cotton Yield Prediction

N. Manoharan and Arunkumar Thangavelu

Abstract The main objective of this paper is to build a novel integrated fuzzy approach to find the influential factors as well as ranking of the factors related to the cotton yield with a fuzzy Decision Making Trial And Evaluation Laboratory **(DEMATEL)**. Fuzzy Cognitive Maps **(FCM)** and its important elements has been used for assessment of cotton yield prediction. This paper proposed the interdependence of every factors on each other and how it directly or indirectly affects the cotton yield using hybrid DEMATAL and FCM. No previous studies have integrated FCM and DEMANTAL for the prediction of cotton yield. Furthermore, evaluation results of this study reveal that the integration of DEMATAL-FCM could be effective and as well accurate compared to existing approaches for evaluating cotton yield prediction.

Keywords Fuzzy cognitive maps · Fuzzy decision making trial and evaluation laboratory · Cotton yield prediction · Benchmark algorithms

1 Introduction

The economic growth of any country in a global platform is largely depends on the agriculture sector. Due to the incredible focus given for improving growth of a agriculture and textile sector, a wide range of practices such as learning and adaptation approaches for improving and predicting the cotton yield have been adapted. Early prediction of agricultural production has been significant impact for agricultural strategy and textile trade. The prediction of the yield depends on based on the combination of the factors affecting it, is greatly reliable for farm management. Moreover, cotton is a very vital yield particularly in India. Despite the important need for estimating the cotton yield in well in advance.

The earlier studies have utilized the traditional approaches, models, algorithms and statistical techniques (correlation and multiple regression models) via empirical methods for the evaluation of crop prediction. Traditional methods have certain limitations to deal with human perception and subjective vagueness in the decision making process. Moreover, the factors influencing and predicting the cotton yield

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_4

Table 1 Preprocessed dataset attributes

Concept	Description
C1: shallow EC	Soil shallow electric conductivity
C2: Mg	Magnesium
C3: Ca	Measured calcium in the soil
C4: Na	Measured sodium in the soil
C5: K	Measured potassium in the soil
C6: P	Measured phosphorous in the soil
C7: N	Measured NO_3 in the soil
C8: OM	The % of organic matter in the soil
C9: Ph	PH of the soil
C10: sand	The % of organic matter in the soil
C11: clay	The % of organic matter in the soil
C12: yield	Seed cotton yield
C1: shallow EC	Soil shallow electric conductivity

via statistical approaches might not provide the precise results [1]. To overcome such inadequacy, the recent studies have utilized various fuzzy multi-criteria decision making (MCDM) approaches for especially handling various attributes in the decision making process to the wide variety of applications. The earlier studies have been used FCMs to predict yield in cotton crop. Based on this research gap, this study have proposes a novel fuzzy approach that integrates the DEMATEL-FCM methodology for predicting the cotton yield. This research has investigated the uses of DEMATEL-FCMs extensively and shows how the proposed methodology would be applied for efficient decision making, analysis and prediction purposes on a given data. The data is normalized and modified based on the type of problem being solved.

In this paper, fuzzy DEMATEL method is used to determine the weights of various factors of cotton yield, identify their significance, the DEMATEL weights has been applied in FCM for different evaluation criteria and map cause-effect relationships and interactions among them. Transforming a linguistic evaluation into fuzzy rating is quite effective to handle this ambiguity from the decision maker's side (Table 1).

2 Materials and Methods

2.1 The DEMATEL Approach

In this study, the main intuition of using fuzzy **DEMATEL** method to evaluate the seventeen **SC** risk attributes was to establish the inter relationship among the four **SC** criteria especially with respect to uncertainty and subjective vagueness within the decision making process. Thus, for computing the weights and ranking of **SC** risk factors, the fuzzy **DEMATEL** method is adopted and its computational procedure involves 8 steps which are summarized as follows:

Step 1: Obtain pair-wise comparison of risks against each other. Based on the survey conducted, we obtain a pair wise comparison of each criterion to form a linguistic variable matrix X.

Step 2: Obtain the triangular fuzzy value and substitute then with a set of linguistic variables. For each linguistic variable ranging from NI (No Influence) to VHI (Very High Influence), its corresponding triangular fuzzy value is substituted using Table 2.

Step 3: Construct initial direct relation matrix D. The initial direct-relation matrix D is obtained by converting the linguistic scale into TFN values. Each element $d_{ij}$is a positive integer that denotes the impact of criteria i on factor j. It is important to note that the diagonal elements d_{ij} are equal to 0.

Step 4: Normalize initial direct relation matrix D. Normalized matrix N = [d_{ij}] is obtained using Eqs. (1) and (2) as shown:

$$m = \max_{1 \le i \le n} \sum_{j=1}^{n} d_{ij} \tag{1}$$

$$N = \frac{1}{m} D \tag{2}$$

Step 5: Compute the total relation matrix R. In order to calculate the total relation matrix R, an identity matrix of the same size as the normalized matrix N has to be constructed. Total relation matrix R can be calculated using the formula given in Eq. (3)

$$\mathrm{R} = \mathrm{N}(\mathrm{I} - \mathrm{N})^{-1} \tag{3}$$

Table 2 TFN substitution for DEMATEL

No.	Influence
0	No influence
0.5	Neg. low/very low
1	Low
1.5	Neg. medium
2	Medium
2.5	Neg. High
3	High
3.5	Neg. very high
4	Very high

Step 6: Defuzzification is done by converting the fuzzy linguistic variable into crisp scores (CFCS). Consider a set of fuzzy numbers $a_k = (l_k, m_k, u_k)$ where k = 1, 2, 3… n. The corresponding crisp values a_{def} are obtained from Eq. (4).

$$a_{def} = L + \Delta \frac{(m-l)(\Delta+u-m)^2(R-l) + (u-L)^2(\Delta+m-l)^2}{(\Delta+m-l)(\Delta+u-m)^2(R-l) + (u-L)(\Delta+m-l)^2(\Delta+u-m)} \tag{4}$$

where $L = \min(l_k)$, $R = \max(u_k)$, $\Delta = R - L$.

Step 7: Compute row sum (r_i) and column sum (c_j) for each risk. In order to calculate the priority weights, the row sum and column sum are computed based on the following equations:

$$r_i = \sum_{1 \le j \le n} R_{ij} \tag{5}$$

$$c_i = \sum_{1 \le i \le n} R_{ij} \tag{6}$$

For each criterion, the value of (r_i + c_i) represents the overall effect of that corresponding criterion on the other criteria and also effect of other criteria on the current one. Hence, priority weight of the criteria is calculated using (r_i + c_i) since it depicts the overall importance of that criterion.

Step 8: Finally, the weight of the influencing criteria is determined. The priority weight of each influencing criterion is calculated using the Eq. (7) as shown below:

$$W_j = \sum_{j=1}^{n} \left(r_i + c_j\right) \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n} \left(r_i + c_j\right) \tag{7}$$

2.2 *The Model of Fuzzy Cognitive Map*

2.2.1 Fuzzy Cognitive Map Aspects

Fuzzy Cognitive Map has been constructed with the integration of fuzzy logic and cognitive map. It was presented by Kosko. It has a graphical structure illustrating the concepts and the weights.

Figure 1 shows a simple fuzzy cognitive map. The FCM has five concepts namely C1, C2, C3, C4 and C5.

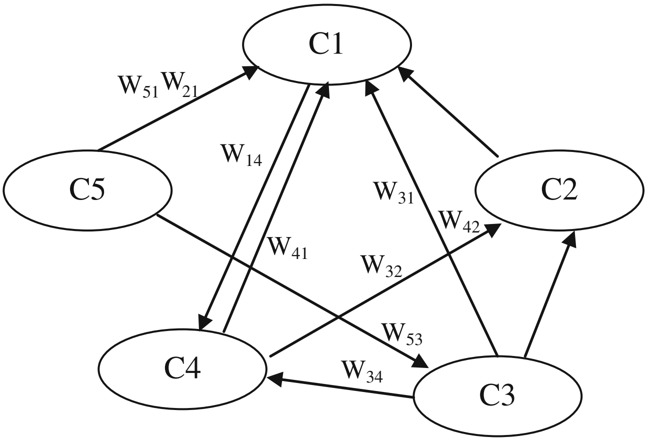


Fig. 1 A simple fuzzy cognitive map

The obtained linguistic variables from the various experts were aggregated using the weighted SUM method and then using the Centre of Gravity method to defuzzify the values obtain a crisp number [2]. This procedure is followed for every single interconnection to obtain the appropriate weight. In most cases, FCM are constructed manually due to lack of experts and hence it is not possible to apply such maps for various real world applications. In such instances, if data is available, data can be exploited to find the appropriate weight using DEMATEL approach.

2.2.2 Fuzzy Cognitive Map Inference Process

The inference algorithm is the most important part of an FCM mechanism. It shows how the concepts interact with each other and in turn affect the system as a whole. The initial vector called as the state vector needs to be set before the start of the inference process. The state vectors is depicted by $A^{(t)}$ at any given point of time t, where $A_i^{(t)}$ denotes the concept value of ci at time *t*. $A_i^{(t)} \in [0, 1]$ for all i = 1, 2, 3 … *N*.

2.2.3 FCM Inference Algorithm

The following algorithm is used in the FCM inference:

```
Begin
Step 1: Initialize the state vector $A^o$
Step 2: Initialize the weight matrix – W
Step 3: Calculate the state vector at time t using the rescaled activation function
        (4) and the sigmoid transformation function (3)
```

```
Step 4: The new state vector $A^{(t+1)}$ is obtained from $A^{(t)}$
Step 5: Repeat the step 3 and 4 until $A^{(t+1)} - A^{(t)} < 0.001$
End
```

The decision concepts of the final vector $A^{(f)}$ at the steady state are assessed and clarify the final decision of the specific decision making system.

2.2.4 Development of DEMATEL-FCM Model for Cotton Yield Prediction

The initial development of the FCM as done with the help of experts' knowledge. The experts are well known with how the various factors affect each other in daily life and to what extent. Hence it is possible for experts to define the degree to which a particular concept affects another concept. Ecologists were approached and asked for guidance in carefully building the FCM. For this, a survey was conducted which consisted of a questionnaire comprising of questions such as how a small change in a particular factor affects the rate of change in another factor and also its effect on the decision concept. The ecologists were asked to choose the degree of causality from a fuzzy linguistic scale shown in Fig. 4 to further crosscheck the model, was referred for verifying the relation between the components of the cotton yields indices. In this way knowledge can be extracted from the experts and used in the further development of the FCM demonstrated further. The results not only give a high prediction rate of 98% but also proved to be better than the existing benchmark machine learning algorithms.

The first most important thing in the development of the FCM model was to establish the important concepts to be chosen for further analysis. Ecologists were asked to state the degree of causality of a particular factor on the decision concept and also a correlation analysis was conducted on the data to find which factors were the most influential factors in decision making. Earlier studies were also taken into consideration in selecting the most important concepts, such as [3] stated that the best classification results were obtained using support vector machines (SVM) with the four weather data. Temperature proved to be the most sensitive attribute followed by rainfall, wind and relative humidity. Adding spatial and temporal attributes had no effect on the overall accuracy, in fact models tested without spatial and temporal attributed tend to give better performance. Hence, due to low correlation of the spatial and temporal attributes with the area burned and its low performance, they were ruled out. A total of 8 concepts were selected in the development of the FCM model and 1 decision concept. The description of the concepts of the FCM model is provided in Table 3.

A rule based approach was carried out to establish the relation between the various concepts. Experts used their knowledge to determine how a slight change in

Table 3 Concepts of the FCM model

Concepts	Description	Range of measured values
C1: FFMC	FFMC index from the FWI system	18.7–96.20
C2: DMC	DMC index from the FWI system	1.1–291.3
C3: DC	DC index from the FWI system	7.9–860.6
C4: ISI	ISI index from the FWI System	0.0–56.10
C5: temp	Temperature in Celsius degrees	2.2–33.30
C6: RH	Relative humidity in %	15.0–100
C7: wind	Wind speed in Km/h	0.40–9.40
C8: rain	Outside rain in mm/m^2	0.0–6.4
C9: area burned	The burned area of the forest (in ha)	0.0–1090.84

one concept could affect the other concept. On carrying out the following procedure, temperature, rain and wind speed were identified to be independent variables and directly affected the area burned. The various other cotton yield indexes were affected directly by the weather data. Rain, temperature, relative humidity and wind directly affect the FFMC index. The DCM index is affected by rain, relative humidity and temperature. Rain and temperature affects DC index. The ISI is affected by the FFMC and wind. The FCM model constructed in shown in Fig. 3.

Usually experts play a very important role in determining the weight matrix. Experts determine how a change in a particular concept affect the other concept. It is done by defining the relationship between concepts using rules. These rules are then de-fuzzified into crisp weight using linguistic concepts. Three experts were asked to separately describe the relationship between concepts using If-Then rules. The following shows the methodologies involved:

First Experts:
If a small change occurs in temperature (C5), then a very high change occurs in the area burned (C9).
It means: The influence from C5 to C9 is positively very strong.

Second Experts:
If a small change occurs in temperature (C5), then a high change occurs in the area burned (C9).
It means: The influence from C5 to C9 is positively strong.

Third Experts:
If a small change occurs in temperature (C5), then a very high change occurs in the area burned (C9).
It means: The influence from C5 to C9 is positively very strong.

After generating all the rules related to the influence of every factor with the other, the rules were then converted to numerical weights using the SUM method and the Centre of Gravity method shown in Fig. 2. The relationships obtained are shown in Tables 4 and 5.

Fig. 2 Weight calculation using DEMATEL Method

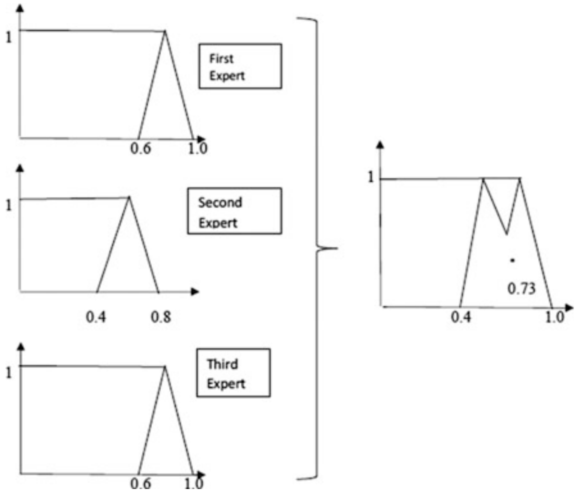


Table 4 Indirect relation matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	0	0	0	0	0	0	0	0	0	0	0	0.1245
C2	0	0	0	0	0	0	0	0	0.2491	0	0	0.2179
C3	0	0	0	0	0	0	0	0	0.3114	0	0	0.3736
C4	0	0	0	0	0	0	0	0	0	0	0	0.5916
C5	0	0	0	0	0	0	0	0	0.4982	0	0	0.4982
C6	0	0	0	0	0	0	0	0	0.1557	0	0	0.3736
C7	0	0	0	0	0	0	0	0	0.3114	0	0	0.2491
C8	0.3114	0	0	0	0	0	0	0	0	0	0	0.3114
C9	0	0	0	0	0	0	0	0	0	0	0	0.1245
C10	0.0934	0	0	0	0	0.1868	0	0	0	0	0	0.5293
C11	0.3114	0	0	0	0	0	0	0	0	0	0	0.4359
C12	0	0	0	0	0	0	0	0	0	0	0	0

3 Results and Discussion

3.1 Classification Results

The FCM model shows in Fig. 3 is used to predict the type of cotton yield in terms of the area growth. For prediction purposes, the data was first normalized. The dataset represent a majority of the small cotton yields. On the analyzing the data and as suggested by the experts after evaluation and careful consideration, values below 0.65 were considered as low cotton yield. The values between 0.65 and less than 0.8 were considered as medium cotton yield and anything above the value of

Table 5 Total relation matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Ri	Cj	ri + cj	Ri − cj
C1	0	0	0	0	0	0	0	0	0	0	0	0.125	0.125	0.715	0.84	−0.59
C2	0	0	0	0	0	0.048	0	0	0.257	0	0	0.268	0.573	0	0.573	0.573
C3	0	0	0	0	0	0.06	0	0	0.321	0	0	0.436	0.817	0	0.817	0.817
C4	0	0	0	0	0	0	0	0	0	0	0	0.592	0.592	0	0.592	0.592
C5	0	0	0	0	0	0.096	0	0	0.513	0	0	0.598	1.207	0	1.207	1.207
C6	0	0	0	0	0	0.03	0	0	0.16	0	0	0.405	0.595	0.486	1.081	0.109
C7	0	0	0	0	0	0.06	0	0	0.321	0	0	0.311	0.692	0	0.692	0.692
C8	0.311	0	0	0	0	0	0	0	0	0	0	0.350	0.661	0	0.661	0.661
C9	0	0	0	0	0	0.192	0	0	0.03	0	0	0.2	0.422	1.602	2.024	−1.18
C10	0.093	0	0	0	0	0	0	0	0	0	0	0.541	0.634	0	0.634	0.634
C11	0.311	0	0	0	0	0	0	0	0	0	0	0.475	0.786	0	0.786	0.786
C12	0	0	0	0	0	0	0	0	0	0	0	0	0	4.301	4.301	−4.301

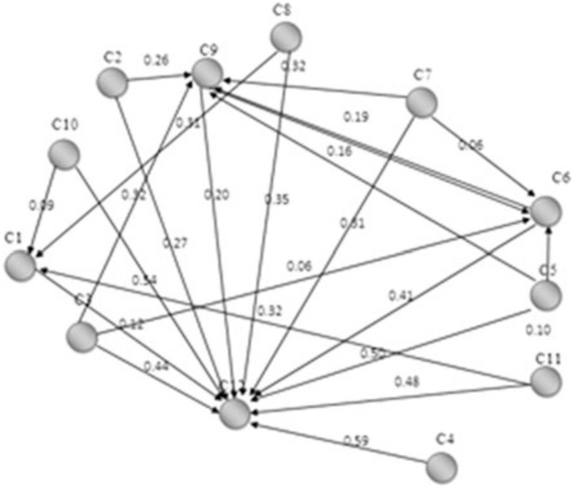


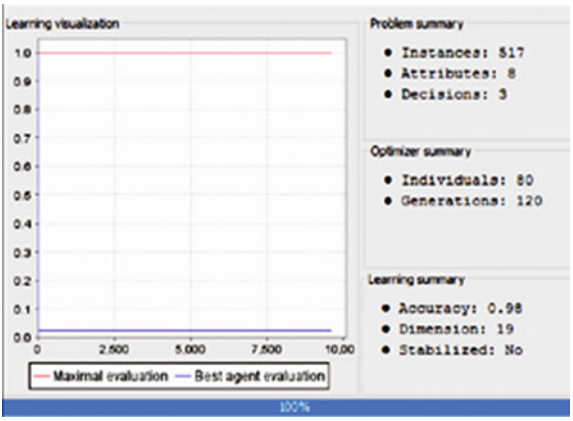
Fig. 3 The FCM model

Table 6 Decision parameters

Class	Min	Max
Low	0.0	0.65
Medium	0.65	0.80
High	0.80	1.0

0.8 was considered high cotton yield. All the experiments have been addressed in this study via FCM wizard tool. The setting of the decision parameters is shown in Table 6.

Further, five learning algorithms namely Data Driven Non-linear Hebbian Learning, Differential Evolution, variable mesh Optimization, Real-Coded Genetic algorithm were tested on the FCM model using the FCM wizard tool to select the most optimized learning algorithm. After a large number of conducted experiments with FCM wizard tool, DEMATEL was proved to perform better than the rest by giving an accuracy of 98% and hence was decided to be used for training the FCM model with Least Squares minimization function chosen as the optimization function to generate the best fit weights from the data provided [4]. The Fig. 4 shows the accuracy obtained along with the classification. The green and the blue points in Fig. 5 represent misclassification in predicting the small cotton yields. As shown in the decision parameter setting in Table 6, all instance of small cotton yields classified below the 0.65 mark depict the right classification of small cotton yields. The same data set tested with the benchmark algorithms discussed in Sect. 4. The FCM model gave better classification results and proved to be better classification model.



Problem summary

- Instances: 517
- Attributes: 8
- Decisions: 3

Optimizer summary

- Individuals: 80
- Generations: 120

Learning summary

- Accuracy: 0.98
- Dimension: 19
- Stabilized: No

Fig. 4 Accuracy of the prediction

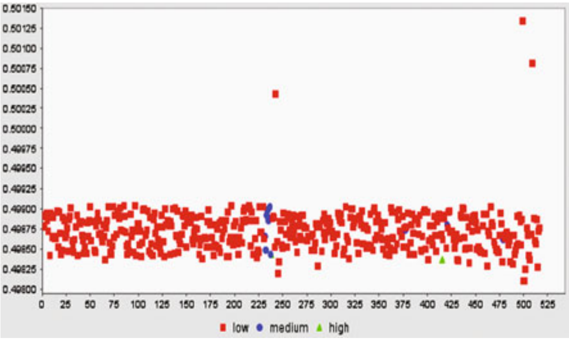


Fig. 5 Classification of forest data

3.2 *Simulation of FCM Tool*

The simulation of the FCM model shows how the system reacts when various concepts are activated. It depicts how each concepts influences other concepts and in turn the final decision concept. The concepts are updated using (4). To show the simulation of the model, all concepts were activated to 1 expect the decision concept. Hence the initial state vector is denoted as A(0) = [0.1 0.8 0.6 0.3 0.6 0.6 0.1 0.85 0.2 0.6 0.8 0.5]. The inference depicts how all the concepts will interact with each other and affect the decision concept. This is clearly shown through the graphical representation of the simulation denoted in Fig. 6. The inference table depicted in Table 7 shows how the concepts values change through the iteration. The system converges at the fourthstep. As shown in the inference, initially the area burned is less. But due to high temperature, the area burned increases. Further, a decrease takes place in the degree of the cotton yields due to the decrease of the temperature and the presence of rainfall and relative humidity leading to relatively

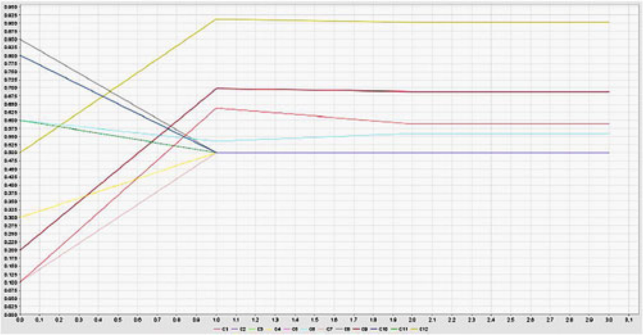


Fig. 6 Inference of the FCM model

Table 7 Inference table

Steps	C1	C2	C3	C4	C5	C6	C7	C8	C9
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
1	0.4182	0.5075	0.5498	0.81	0.5	0.4975	0.5	0.5	0.9227
2	0.4591	0.5039	0.525	0.6619	0.5	0.4988	0.5	0.5	0.824
3	0.459	0.5038	0.525	0.6678	0.5	0.4988	0.5	0.5	0.8019
4	0.459	0.5038	0.525	0.6678	0.5	0.4988	0.5	0.5	0.8027

stable system. The final value of the decision concept is 0.8027 which denotes an 80.27% serverity, hence showing high prediction of cotton yields. Such inferences help in making decisions and analyzing environmental situations.

4 Conclusion

The methodology applied in this work shows the successful prediction of cotton yields using FCMs. The learning of the model was enhanced with the use of DEMATEL approach. The evolutionary algorithm used on the data obtained from the UCI Machine Learning Repository helped to optimize the weight matrix further and give highly accurate results. The main objective of this work was to present a new method based on the FCM learning technique, the efficient use of the FCM wizard tool on making predictions and to develop an intelligent decision support system to predict cotton yields. The designed and learned model using FCm wizard tool provides accurate prediction on the small cotton yields obtaining 98% accuracy. The further scope of this work could be to try different hybrid approaches to further enhance the classification of the FCM and help enable better prediction. The methodology proposed in this study could also be used to analyze various other environmental problems and hence could be an efficient way to make intelligent decisions.

References

1. Mues, V., R. Fischer, R. Becker, V. Calatayud, N. Dise, G.H.M. Krause, et al. 2012. Forest condition in Europe, 165. Edited by M. Lorenz, & G. Becher. Thunen institute, Bundesforschungs institute wald and fischerei.
2. Colaco, M., F. Castro Rego, P. Meirs, and T. Santos. 2005. What are the opinions of foresters in portugal regarding fire? In *Proceedings of the 2nd International Conference on Prevention Strategies of Fire in southern Europe*, 9–11. Barcelona (Spain).
3. Cortez, P., and A.D.J.R. Morais. 2007. A data mining approach to predict cotton yields using meteorological data.
4. Papageorgiou, E.I., K.D. Aggelopoulou, T.A. Gemtos, and G.D. Nanos. 2013. Yield prediction in apples using FUZZY cognitive Map learning approach. *Computers and Electronics in Agriculture* 91, 19–29.

A Refined K-Means Technique to Find the Frequent Item Sets

A. Sarvani, B. Venugopal and Nagaraju Devarakonda

Abstract In this paper we have shown the behaviour of the new k-means algorithm. In k-means clustering first we take the 'n' number of item sets, then we group those item sets into the k clusters by placing the item set in the cluster with nearest mean. The traditional k-means clustering is completely depend on initial clusters and can be used only on spherical-shape clusters. The traditional k-means clustering uses the euclidean distance but in our paper we have replaced it with minkowski distance and combined with the Generalized Sequential Pattern algorithm (GSP algorithm) to find the frequent item sets in the sequential data stream. The GSP algorithm based on the frequent item sets, it traces the databases iteratively. The modified k-means clustering have reduce the complexity and calculations and the GSP algorithm has given the better result than any other algorithm to find the frequent item sets. The results show that this approach has given the better performance when compared to the traditional k means clustering.

Keywords Generalized sequential pattern algorithm · K-means algorithm
Minkowski distance

1 Introduction

Here in our paper we concentrate on clustering of the similar objects where the objects in one cluster have similar behaviour than the objects in other clusters. In a cluster we can have number of observations or records or events. There are number of techniques for clustering some of them are partition based, hierarchical based, density based, grid based and model based clustering. From these many clustering technique we have studied on k-means clustering which comes under typical partition based clustering.

K-means is designed to solve many of the clustering problem and this k-means comes under unsupervised learning algorithms. The k-means clustering starts by randomly selecting k number of clusters. The k can be any number of clusters. Here we will have k-centres one for each cluster.

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_5

Even though the k-means is most used clustering technique in many applications it has many issues to be resolved. The traditional k-means clustering is completely depend on initial clusters and can be used only on spherical-shape clusters. This has also increased computational complexity exponentially.

In traditional K-means algorithm we use the Euclidean distance but in our paper we used minkowski distance measure. For the two points X = {x_1, x_2 … x_n} and Y = {y_1, y_2 … y_n} the minkowski distance is defined as:

$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \quad (1)$$

This gives raise to k clusters, then for each cluster we apply GSP algorithm to find the frequent item sets with their support and confidence percentage. The GSP algorithm can efficiently find the frequent item sets in the sequential data stream. Our experiment result has proven that the refined k-means combined with GSP algorithm has given the better performance than any other algorithms.

The Redmond et al. [1] given the starting methods of k-means, here randomly we choose the k number of clusters and then we calculate their centroids. Hartigan et al. [2] suggested the two different methods. In the first method we select the initial k points which becomes the initial centres and in the second method we select the k centres in random fashion. This leads to two disadvantages, the k-centres are completely depend on the order of sequence if we choose the first method and if we choose the second method we may selects the points which are closer to each other and there is also chance to leave the points which are far away. kanungo et al. [3] has given a new method where the dataset is divided into J number of groups and then has applied k-means on each group which gives raise to many centres. Next these centres are grouped into superset. Again the k-means is applied on this superset. Now the centres having the members with least SSE (sum of the squared distance between each member of the cluster and its centroid) is chosen as initial clusters.

Qi et al. [4] this uses the hierarchical approach to find the initial clusters. In this paper Redmond et al. at different locations the volume of data is estimated using the k d trees and at every location maximin method is applied to find the initial cluster. Bradley et al. [5] brought out three optimization principles along with the optimized *k*-means clustering method.

2 Calculation of Minkowski Distance Measure

Suppose we have I = (X1, X2) and J = (Y1, Y2) let order is p. Then to calculate the distance from point (X1, Y1) to the point (X2, Y2) we use the mathematical equation

$$\mathrm{D}\,(\mathrm{I},\mathrm{J}) = \sqrt[p]{|X_1 - Y_1|^P + |X_2 - Y_2|^P} \tag{2}$$

By substituting the values we can get the distance between the points. If D(I, J) > 0 it is taken as positive.

If D(I, J) = D(J, I) it is taken as symmetric and if D(I, J) $\leq$ D(I, K) + D(K, J) it is taken as triangle inequality.

If p = 1 then it is Manhattan distance then

$$\mathrm{D}(\mathrm{I},\mathrm{J}) = |X_1 - Y_1| + |X_2 - Y_2| \tag{3}$$

Using this distance measure we calculate the block distance, this is relatively easier to calculate but can be used only for the order 1.

If p = 2 then it is Euclidean distance then

$$\mathrm{D}(\mathrm{I},\mathrm{J}) = \sqrt[2]{|X_1 - Y_1|^2 + |X_2 - Y_2|^2} \tag{4}$$

Using this distance measure we calculate the diagonal distance, but here the order of p is fixed to ‘2’. Compared to minkowski distance measure the Euclidean distance is simple, but the minkowski distance measure can be adopted to any order (p) and can be adopted to any application.

3 Clustering Methods

3.1 Hierarchical Methods

It iteratively divides the item sets into number of clusters. These can be done by both a top-down or bottom-up style and again these can be divided into Agglomerative clustering and hierarchical clustering. Here first we consider each item as one separate cluster which are again combined to form a cluster hierarchical structure. This continues till we reach our desired structure.

The reverse of above method can also be done by taking entire item sets in one large clusters and dividing it into separate cluster till we reach the clusters having the similar behaviour. This method is called Divisive hierarchical clustering.

3.2 Partitioning Methods

In this method the user need to tell beforehand the number of clustered needed. By this method we can change the position of the item sets from one cluster to another cluster.

3.2.1 Error Minimization Algorithms

This method can be used efficiently with compact and remote clusters. This is the algorithm which is used by most of the people. This is used to reduce some of errors in formation of cluster structure. K-means clustering can also be considered under this algorithm.

3.2.2 Graph-Theoretic Clustering

This uses the graphs theory to form the clusters. Here the item sets are represented by nodes. The relationship between item sets is represented with the edges and this uses the Minimal Spanning Tree.

3.3 Model-Based Clustering Methods

This fits the given data on to some mathematical models. In traditional clustering we identify the cluster of item sets. Besides clustering model-based methods also detect each cluster behaviour. Here the each group or each cluster can be concept or class. In this we use neural networks and decision trees.

3.4 Grid-Based Methods

Here the grid structure is formed by dividing the entire space into fixed number of rows and columns. On each cell the clustering operations are done. This method can produce the output in less time.

4 K-Means Clustering

This comes under unsupervised learning and used to group the item sets. In this first we need to decide the number of cluster needed then the given item sets are grouped into clusters based on distance between centres of the cluster and item set. Here for k cluster there are k centres. We should select the centres which are far-away to each other to yield the better performance. After selecting the appropriate centres now we need to select all item sets one after another and link them to the centre which is close to it. When we have linked all the item sets to their centres we are then completed with first stage of k-means clustering. After this again we need to construct the new centre for the newly formed k clusters. Again we need to calculate the distance between the item sets and the new centroid to assign the item sets to the closest centroid. This process should be done repeatedly until there are no more changes to the centroids.

Algorithmic steps for k-means clustering

Let the item sets = {x_1, x_2, x_3, …, x_k} and C = {c_1, c_2, … c_k} be the set of centres.

1. First select K centres of cluster.
2. Measure the distance between centre of a cluster and item set.
3. Associate the item set to the centre of a cluster which is nearer to item set.
4. Again measure the centre of newly formed clusters.
5. Again associate the item set to the new cluster centre based on distance.
6. Repeat from step 3 until there are no new centres.

5 GSP Algorithm

This GSP is also called as Generalized Sequential Pattern algorithm. This algorithm is used detect the most frequent item sets in a cluster. This can efficiently detect the frequent item sets from the data which is in the form of sequential patterns. This algorithm repeatedly scans the data number of times. The first scan of the data reveals the support of items i.e., the occurrence of the item in data. After the first scan we can know the support of each item in the data and also which items have passed the minimum support. The support or threshold value is decided based on the application. In the first scan we can know only the support of individual item. Now by using the frequent items (candidate 1 sequence) from the first scan the algorithm measure the 2 level frequent items (candidate 2 sequence) this continues until no more frequent item sets are present. After completion of each scan the algorithm measure the support of each item set. Finally after successful completion of the algorithm we can know the most frequent item set with their support and confidence (Fig. 1).

Algorithm

```
F1 = candidate 1-sequence set
Continue until F (k-1)! = Null;
Construct the candidate sets $C_k$
For every input sequences s present in database D
do
Increment count of items in $C_k$ if same item is present in s Fk
= {Items whose frequency is above threshold comes under $C_k$}
k = k + 1;
Result (Fks) = Sum of all frequent sequences
end
```

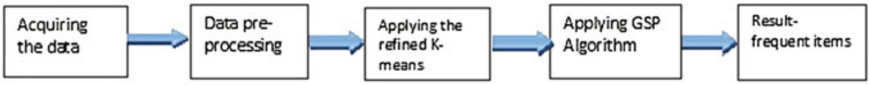


Fig. 1 Flowchart

6 Methodology

Here we assign different ids to each every item in the clusters. This helps use to uniquely identify the items in cluster. Each item has a separate location in cluster which depends on the environment. Now the user selects the k number of clusters, the k can be any number. Now from the item sets the k centres are calculated. Now each item is linked to the closest centre based the distance between item and cluster centre. Here we have used minkowski distance. This process continues till there are no more new centres are formed. After we have completed the k means clustering, we apply the GSP algorithm on each and every cluster to get the frequent items with their support and confidence.

6.1 Acquiring the Data

In this experiment we have taken the real world datasets one i.e., related to stock price and other dataset is transactions which describes main characteristics of the Stock Price and transactions respectively.

6.2 Data Pre-processing

In this set we have removed the inconsistent data, missing values, duplicate data. This is an important step to produce the accurate result.

6.3 Applying the Refined K Means

Here we have applied the modified k-means clustering. The traditional k-means clustering had used the euclidean distance but for k-means in this paper we have used minkowski distance. This refined k-means is applied on the dataset.

6.4 *Applying the GSP Algorithm*

After the application of k-means clustering. We get the k number of clusters. Now on each cluster we applied the GSP algorithm which gives the frequent item sets present in the dataset. For each frequent item we calculate the support and confidence.

7 Experiment

To check the performance of the refined k-means clustering and also GSP algorithm we have used the two real world datasets. This experiment in done on window 10 operating system with ram 2 GB, hard disk of 500 GB and on the processor Intel core i4. The dataset information is given in Table 1.

After we have done the pre-processing on the data then we applying the refined k-means algorithm using minkowski distance measure.

We have taken the dataset stock details and applied k-means clustering using both euclidean distance and minkowski distance measure which is shown in Tables 2, 3 and 4.

Table 1 Dataset information

Dataset name	Number of attributes	Number of instances	Missing values	Origin
Stock price	10	950	No	Real world
Transaction	3	120,427	No	Real world

Table 2 K-means using Euclidean distance when k = 2

Attribute	Full data (941.0)	Cluster 0 (493.0)	Cluster 1 (448.0)
Company 1	38.1145	46.8836	28.4645
Company 2	43.8961	34.8344	53.868
Company 3	18.6973	16.5773	21.0301
Company 4	45.3755	46.23	44.4353
Company 5	60.8601	50.1425	72.6543
Company 6	24.0937	20.2723	28.2988
Company 7	70.7109	71.2391	70.1297
Company 8	23.3478	25.396	21.0938
Company 9	44.1764	42.9092	45.5709
Company 10	47.1072	50.0545	43.8638

Table 3 K-means using Minkowski distance when k = 2

Attribute	Full data (941.0)	Cluster 0 (473.0)	Cluster 1 (468.0)
Company 1	39	46.875	27.781
Company 2	47	35.25	53.375
Company 3	19.375	16	20.875
Company 4	44	43.375	44.375
Company 5	61.75	49.625	72.0625
Company 6	25.625	18	28.3125
Company 7	68.75	68	69
Company 8	22.5	26.375	21.625
Company 9	44.75	42.25	45.25
Company 10	46.75	48.25	41.5

Table 4 Comparison between k-means using euclidean distance and minkowski distance

Attribute	Cluster 0 in Euclidean distance	Cluster 0 in Minkowski distance	Cluster1 in Euclidian distance	Cluster 1 in Minkowski distance
Company 1	46.8836	46.875	26.4645	24.781
Company 2	34.8344	35.25	53.868	53.375
Company 3	16.5773	16	21.0301	20.875
Company 4	46.23	43.375	44.4353	44.375
Company 5	50.1425	49.625	72.6543	72.0625
Company 6	20.2723	18	28.2988	28.3125
Company 7	71.2391	68	70.1297	69
Company 8	25.396	26.375	21.0938	21.625
Company 9	42.9092	42.25	45.5709	45.25
Company 10	50.0545	48.25	43.8638	41.5

From the above table it is clear that for the attribute “Company 1” cluster 0 using euclidean distance and minkowski distance are having 46.8836 and 46.875 item sets respectively and for the same attribute cluster 1 using euclidean distance and minkowski distance are having 26.4645 and 24.781 item sets respectively. This continues for rest of attribute from company 2 to company 10.

From the experiment it is clear that k-means using minkowski have given almost the same result of the traditional k-means clustering. But k-means using minkowski had required less iterations and has less squared error when compared with traditional k-means clustering.

After the completion of k-means clustering we have applied GSP algorithm on both cluster 0 and cluster 1 of all the attributes from company 1 to company 10 which have given their respective support and confidence. The support and confidence values are shown in Table 5.

Table 5 Support and confidence of frequent item

Data sequence	Support	Confidence (%)
Cluster 0 {company 1, company 3, company 7}	70	70
Cluster 1 {company 2, company 5, company 9}	62	86

From the above table the sequence “company 1, company 3, company 7” is having fraction of 70 of total data sequence and the sequence “company 2, company 5, company 9” is having fraction of 62 of total data sequence The most frequent item set in the cluster 0 is {company 1, company 3, company 7} having the confidence of 70%. The most frequent item sets in the cluster 1 is {company 2, company 5, company 9} having the confidence of 86%.

8 Conclusion

This paper focused on finding the frequent item sets from the sequential data by first applying the refined k-means algorithm on the data and then applying the GSP algorithm on each and every cluster. We have shown an experiment that the k-means with minkowski distance is better than traditional k-means clustering in the aspect of iterations and shared error and also we have calculated support and confidence of the frequent set. The future scope is to reduce the time duration and complexity of the algorithm.

References

1. Redmond, S.J., and C. Heneghan. 2007. A method for initialising the K-means clustering algorithm using kd-trees. *Pattern Recognition Letters* 28 (8): 965–973.
2. Hartigan, J.A., and M.A. Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1): 100–108.
3. Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (7): 881–892.
4. Qi, J., Y. Yu, L. Wang, and J. Liu. 2016, October. K*-means: An effective and efficient k-means clustering algorithm. In *Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, 2016 IEEE International Conferences, 242–249. IEEE.
5. Bradley, P.S., and U.M. Fayyad. 1998, July. Refining Initial Points for K-Means Clustering. *ICML* 98: 91–99.

Finger Vein Detection Using Gabor Filter and Region of Interest

Saritha Reddy Venna, Suresh Thommandru
and Ramesh Babu Inampudi

Abstract One of the most biometric recognition systems is Finger vein recognition. It affords authentication using veins, which are within the finger of individuals. This technology has the dominance of being defiant to falsification because the vein pattern is hidden inside a finger. In this paper, we present the key problems and feature extraction methods in order to acquaint finger vein recognition research domain. In our proposed method, vein image is subjected to Gabor filter in conjunction with Region of interest during the preprocessing and to efficiently enhance the invisible vein pattern of given image. Observations are present that our proposed could obtain patterns staunchly when vein dimensions and illumination oscillate, and also the experimental errors for personal identification are very low, in comparison with that of traditional methods.

Keywords Biometric finger vein · Gabor filter · Region of interest
Feature extraction

1 Introduction

Biometrics is in wide spread use at present, as security is of major concern in this modern and networked world. This has led the extensive use of biometrics i.e., the physiological or behavioral characteristics of human beings for authentication purpose. Authentication plays a vital role to protect sensitive data from unauthorized access. The notable features of biometrics such as Universality, Uniqueness, Collectability, Performance, Permanence, Acceptability and Circumvention have made them suitable for authenticating individuals based on their unique traits. This paved the way to biometric authentication and its applicability in areas such as access control, e-commerce, automated teller machines, remote access, financial transactions and network security. Biometric Authentication is primarily used for user identity verification using either physiological or behavioral traits such as finger print, iris, finger vein, palm vein, hand geometry, signature, voice, gait etc.

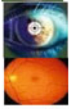
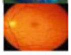
S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_6

Biometric authentication can be accomplished using any kind of biometric trait but each has its own set of merits and demerits [1–3].

Facial recognition is non-intrusive and a cheaper technology but requires camera for identification [4]. Aging, facial expressions, changes in lighting may decrease the recognition accuracy and also incurs high cost to build. Fingerprint is a widely used and accepted technology for personal authentication but it suffers from low genuine recognition rate and is relatively easy to forge. The finger image quality is highly dependent on the contact between the finger and the sensor and is also susceptible to various conditions such as sweat or wetness of fingers, age etc. [5, 6]. Hand shape recognition has attracted many people as the images are acquired in a user-friendly and non-intrusive manner using low cost sensors [7, 8], can use low quality images along with providing high recognition rate. The demerits would be the size of the hand that makes it applicable to a few applications and hand injury may affect the user acceptance rate. Keystroke is economical and can be easily integrated into the security systems. But the key stroke may differ each time leading to low accuracy in the feature extraction process. Another personal identification system is palm vein recognition using individual veins of palm [1]. But in comparison to the finger vein technology, it is relatively expensive; the size of template is large, and may not be applicable to a human subject who doesn't possess both the palms. Iris pattern possess many advantages such as distinctiveness, life time stability, difficulty to forge etc. that makes it an acceptable trait for biometric authentication. Iris has high recognition accuracy but during image capture it poses few demerits such as necessity of proper amount of light to capture the image, can be obscured by objects such as eyelids and eyelashes and individuals who are blind and having cataract eventually results in the difficulty to capture the iris image. Retinal recognition is difficult to replicate and has high accuracy but it is intrusive and very expensive. Signature is a behavioral trait that can be used for identity check and is also easy to implement. But the signatures may vary with time, leading to low genuine recognition rate and can easily be forged. Voice recognition is non-intrusive with high social acceptability, affordable technology and has less processing time. But it poses few demerits such as the voice of an individual can easily be recorded and used for authentication and the voice may become difficult to recognize during illness which in turn leads to low recognition accuracy [9].

To overcome these challenges, we have opted finger vein as a reliable biometric trait to authenticate persons that assures good recognition accuracy and high security. Our proposed system relies on the finger vein patterns which are more reliable for personal identification [2] needs. The contact less capture of finger vein images by passing infrared light signals through the finger would be an added benefit. Finger vein patterns reside beneath the skin surface and are distinct for each human being. High levels of recognition accuracy, security, long-term stability and sufficiency of less memory to store the vein template are the notable advantages of finger veins. The main aim would be to emphasize the integrity and resilience of finger veins that are user friendly and tamper resistant for authentication purpose (Table 1).

Table 1 Comparison of various biometric traits

Biometric		Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face		☼	☻	☺	☼	☻	☼	☻
Finger Print		☺	☼	☼	☺	☼	☺	☼
Hand Geometry		☺	☺	☺	☼	☺	☺	☺
Key Strokes		☻	☻	☻	☺	☻	☺	☺
Palm vein		☺	☺	☺	☺	☺	☺	☼
Iris		☼	☼	☼	☺	☼	☻	☼
Retianl		☼	☼	☺	☻	☼	☻	☼
Signature		☻	☻	☻	☼	☻	☼	☻
Voice		☺	☻	☻	☺	☻	☼	☻

☼ – good, ☺- normal, ☻ –insufficient

Finger images can be easily captured by passing infrared rays from the back of the hand without causing any harm to the human subject. But most of the captured images have the finger veins with capricious obscure and noise. Hence, there is a necessity to pre-process the image in contemplation of restoring the quality of the image by reducing the noise, irregular shading etc., so that the vein pattern is clearly visible which helps in better recognition rate. Finger vein pattern as it resides internal to a finger, it can be used to provide confidentiality that makes it a reliable security solution. The below mentioned features or advantages makes finger vein pattern the best option for constructing a secure and a safe uni-modal recognition system.

1. Veins reside within the finger and are hence difficult to forge.
2. Vein pattern is distinct to each individual.
3. Image acquisition is resilient to conditions such as oil, sweat and dirt.
4. Vein pattern has long-term stability, as it doesn't deteriorate with time.
5. Images can be captured in a contactless manner.
6. Vein pattern matching is very fast as it completes within the span of an eye blink.
7. Veins are more reliable to ensure both privacy and security.
8. Vein authentication devices are compact and can be easily used as embedded devices in various applications.

The related work reveals the existing methods for finger vein detection. In proposed work reveals proposed method with limitations of existing work. In experiment results, discussed the database considering for experiments and gave results of proposed system. In last section reveals the conclusion.

2 Related Work

Naoto et al. [3] used digitization of the vein images and transform of distance for degradation process. Disadvantage is that more amount of time is being consumed for pre-processing due to the iterative nature of repeated line tracking algorithm. In [6], the features were extracted using bifurcation points and termination points that led to single isolation points around the main detected pattern. To overcome this Gaussian filters are used. In [9], Tanushri Chakravorty, used the linear transformation and histogram equalization technique to enhance the image features but could not produce accurate results with feature extraction. In [10], J. Yang proposed detection of finger vein by combining the Gabor wavelet transformation and Gabor filter. It could produce good result using Gabor filter but could not highlight the vein regions accurately. In [11], X. Chen used the Maximum Margin Locality Preserving Projection feature extraction by processing the input image. Region of Interest localization is a fundamental activity for a finger vein recognition system to capture required regions from the input image [12]. It gives best caliber to the recognition system. It is generally composed of finger vein segmentation of required regions, rotation correction, and detection of interested regions.

The demerits of the above mentioned works such as consumption of more storage space and processing time is due to the fact that they have considered the entire image. As Gabor filter performs best at image enhancement and feature extraction, We are proposed a method which both Gabor Filter along with Region of Interest for reducing the time of processing and only required portion in the input image. The features obtained from the vein image can be stored as a single feature vector [13].

3 Proposed Work

Biometric authentication system basically aims at recognizing individuals based on their unique physical or behavioural traits. It generally involves many phases such as Image Acquisition, Pre-processing the image to enhance the image quality which helps in better recognition rate, Next step to extract the features from the preprocessed image and finally a matching algorithm is used to match the stored template with the acquired image features to identify a human subject either as genuine or imposter. In our work, the input image is taken from the SDMULA-HMT database and we have pre-processed this image to enhance the quality of the vein pattern. In

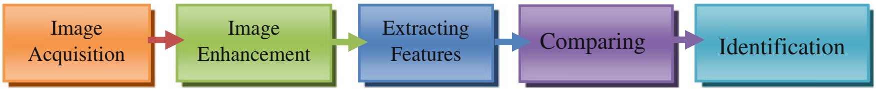


Fig. 1 Basic flow of biometric authentication system

the pre-processing phase, initially we have extracted the region of interest from the input vein image and then applied proposed method to predominantly improve the quality of the required pattern (vein). Extracting the feature points are depends on the quality of the image. This would eventually increase the genuine acceptance rate of the vein biometric system. The below figure specifies the basic flow of the general biometric authentication system in Fig. 1.

4 Acquiring the Image

One of the advantage of the Finger veins are not visible to the human eye, only when the near infrared rays (NIR) of wavelength between 700 and 1000 nm are passed through the finger [14]. There are two techniques namely light reflection method and light transmission method to acquire the required patterns as shown in Figs. 2 and 3.

When the light source (NIR) is passed through the finger, the light will be reflected in haemoglobin. From this reflected light, a charge coupled device camera captures the vein pattern image. From this captured image, the vein pattern is

Fig. 2 Reflection method

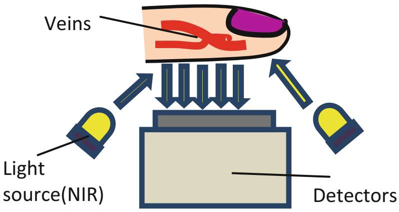
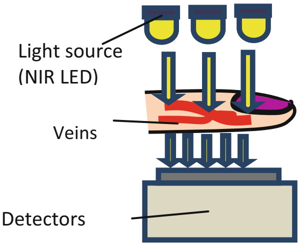


Fig. 3 Transmission method



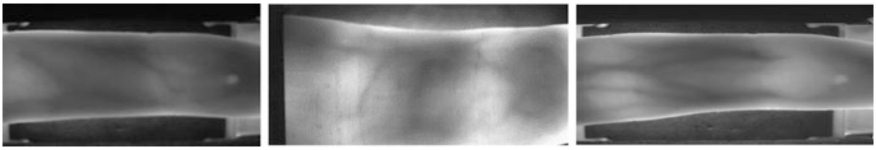


Fig. 4 Sample of finger vein images

constructed using image processing techniques. This pattern is then compressed, converted into digital form and hence registered as a template which is then put in comparison to the stored template of the user. Then process of comparing between images to determine a match. The sample of finger vein images as shown below is taken randomly from the databases of different universities (Fig. 4).

5 Image Enhancement

Input is enhanced based on contrast and noise reduction using spatial filters. Median filter is used for noise reduction in the input image. If we are not gone through this phase, we can get low level features to process in the feature extraction. In Fig. 6g represent the image after median filter on given image.

6 Feature Extraction

In this paper, different existing techniques are put in comparison to augment the need of the proposed method. One of the preprocessing techniques is region growing method; another one is on histogram equalization, median filter and segmentation. Finally, the proposed method is compared with these methods that show the better performance of proposed method.

6.1 Region Growing

This method is used to segment the image based on regions. It is one of the simplest segmentation methods to detect discontinues and regions from the image [15]. It follows edge detection and region identification. The performance of these methods differs with individual segmentation and both will not provide us with the same result. The main criterion for this method is homogeneity of regions like gray level, color, shape, texture etc. Initially, we have to segment the image into small regions and varies aspect ratios of pixels. Region description is compared with neighbour

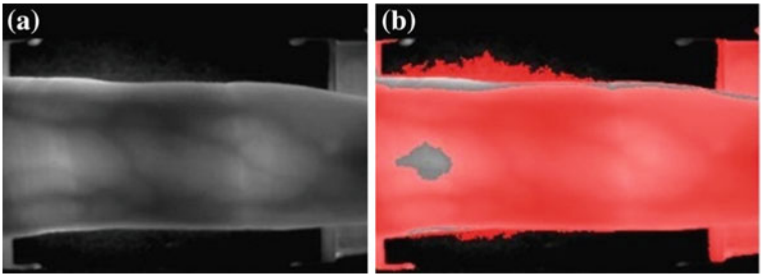


Fig. 5 Image segmentation using region growing **a** input image **b** region growing segmented image

regions. If they match, then those regions will be combined to make a large region until required segmentation of the image (Fig. 5).

Algorithm: Region Growing

Step 1: *Find the Rmin (smallest regions)*
Step 2: *Find the regions which are similar to Rmin, based on homogeneity.*
Step 3: *Merge those regions.*
Step 4: *Repeat the steps 1, 2, 3 until the homogeneity criteria is satisfied.*

6.2 Image Histogram

Histogram is the process of representing the tonal dissemination of image in graphical manner. Characterize the pixels at every tonal value and represent density of pixels each tonal value. It is a simple scanning process that counts the density of pixels found at every change in gray levels. In one of the gray scale image, having 256 possible gray levels and hence the representation will depict the distribution of pixels amongst those 256 gray scale values. The notable advantage of image histogram is that we can decide upon a threshold value when converting a gray scale image to a binary one. Manually we can adjust the scale on the y-axis. If it is automatically, will give high peak values and lead to force a scale. It will negotiate the smaller features. Image histograms can be used by other operations such as contrast stretching and histogram equalization. These operations assume the full intensity range to show the maximum contrast i.e., pixels are distributed or spread evenly over the intensity range.

Histogram equalization is a global enhancement technique that enhances the image contrast by adjusting the intensities of pixels in an image. But this method does not provide us with required contrast to further enhance and segment the finger veins [16]. To overcome this problem, Contrast Limited Adaptive Histogram Equalization (CLAHE) is used to enrich the contrast of a gray scale image to a greater extent as it is a local region based enhancement technique. It divides the entire image into tiles known as regions and operates on these regions, instead on the entire image. It also produces less noise and CLAHE can be calculated as

$$S = Histogram(i) * \frac{255}{M * M} \tag{1}$$

where *Histogram* is the representing the tonal dissemination of regions of image that emulates number of pixels of the *i* region. *S* represent calculated occurrence of new pixel value. *M* represent regional filter or window size. Median filter is a non-linear smoothing filter which is good at removing noise from the image along with preserving the edges and uniformity. This will enhance the image quality. Figure 6g shows the smoothness and noise reduction of the Fig. 6a. The input image has low contrast and hence is difficult to detect and segment the venous regions. As the tip of the vein is not highlighted properly, we consider the segmentation with local and global threshold which will extract the veins with low contrast based on local or regional threshold. Figure 6h shows the segmented image, but some regions near the tip of the vein are difficult to segment. Regions with high visual quality can be observed clearly.

6.3 Gabor Filter and Region of Interest (ROI)

Our proposed technique encompasses the Gabor Filter with Region of Interest (ROI). Region of Interest extracts the required regions from the input image to improve the performance of proposed method due to it is taking less computation time. Existing methods have considered the whole vein image during the process of

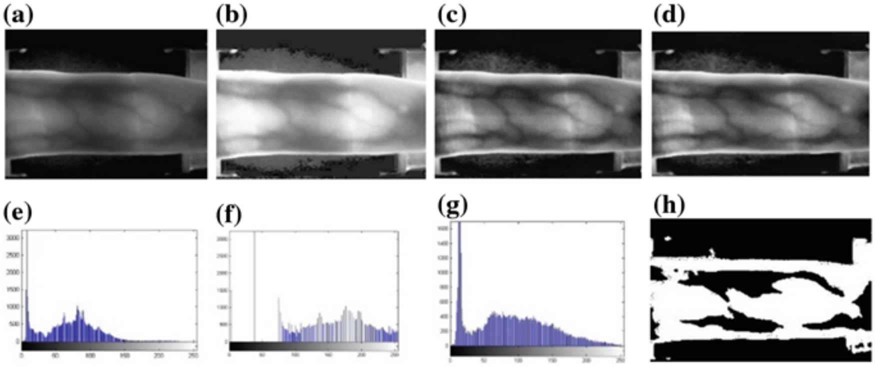


Fig. 6 Seconds method histogram, CLAHE, median filter and segmentation. **a** Input image, **b** image after histogram equalization, **c** CLAHE applied on input image, **d**–**f** histogram of **a**–**c**, **g** image after median filter on input image, **h** segmented image

filtering due to which the computation time increases. The integration of ROI with the Gabor Filter decreases the processing time and also solves the problem of low contrast of the vein images. The Gabor filter is a two-dimensional, adjustable band pass filter; it performs the vein filtering efficiently by considering the magnitude, phase, orientation and peak of the pixel values. As specified above, the finger veins are spread across the finger, the tip of the veins is of non-linear size and the dimension of the veins is also different [17]. It captures the local rotation and frequency occurrence of the venous network. The extraction of features from the image by the using below equation

$$G(x, y, f, \theta) = e^{\left\{\frac{-1}{2}\left[\frac{x'^2}{\delta_{x'}^2} + \frac{y'^2}{\delta_{y'}^2}\right]\right\}} \cos(2\pi f x') \tag{2}$$

where x' and y' are the oriented points after rotation in x and y direction and $\delta_{y'}^2$, $\delta_{x'}^2$ are Gaussian filter spatial constants about the new oriented points, x' and y'. f is the frequency of one of the wave along plane orientated. Obtained features represent the finger venous network.

7 Experimental Results

The results to confirm robustness of the finger vein detection using the Gabor filter with ROI for larger database SDUMLA-HMT [18]. The database helped to experiment proposed method and it is open to researchers with minor authentication.

An Acquisition system was designed by the intelligent computing of Shandong University. Images were acquired through light transmission method for 106 persons (subjects). 6 finger numbers per subject and each finger 6 images are captured and total number of images is 3816 with 320 × 240 pixels image size (Table 2).

Table 2 Performance metrics for finger vein extraction

Feature extraction	Number of images	Performance evaluation metrics	Execution time
Conventional method [19]	3816	EER = 2.36	9 s
Region growing method [19]	3816	Mean sensitivity = 0.711	13 s
Histogram	3816	EER = 0.89	10 s
CLAHE	3816	EER = 0.13	8 s
Proposed method	3816	EER = 0.0009	2.5 ms

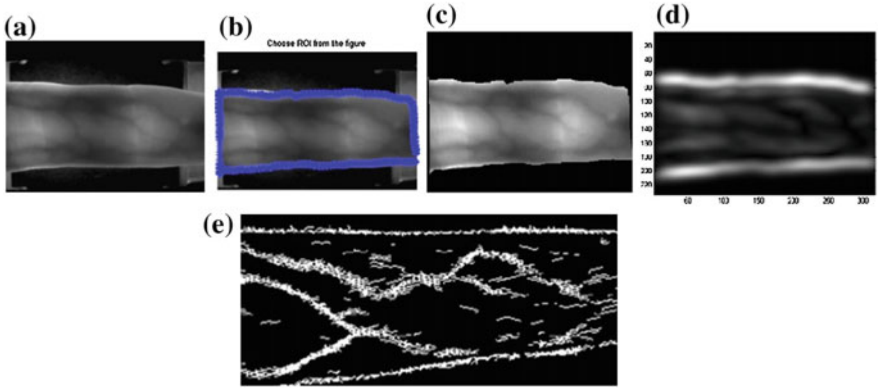


Fig. 7 Feature enhancement with gabor filter. **a** Input image **b** choosing ROI region **c** processing for ROI **d** output of gabor filter **e** detected finger veins

The above experiment gave good results and took less processing time to detect the finger vein regions based on Region of Interest in Fig. 7. But the existing methods consider the total input image that includes unwanted regions.

8 Conclusion

The performance of the finger vein detection by using the Gabor filter and ROI is improved by considering the existing systems. The finger vein pattern identification has improved and also the processing time for extracting the vein pattern is very low with our proposed method. The proposed method was tested and the above results presented were based on the finger vein images of the SDUMLA-HMT [18] database. Our proposed system combining Gabor and Region of Interest can be considered as the most appropriate method in comparison with the above mentioned existing methods for finger vein detection.

References

1. Wassila, B. 2007. Identification Biometrique des Individus par Leursempreintes Palmaires. Mémoire de Magister, Université des Scienceset de la Technologied' Oran USTO-MB.
2. Hitachi. 2006. Finger Vein Authentication-White Paper, Copyright.
3. Naoto, M., A. Nagasaka, and M. Takafumi. 2004. Feature Extraction Offinger-vein Patterns Based on Repeated Line Tracking and Its Application to Personal Identification. *Machine Vision and Applications* 15 (4): S194–S203.
4. Jain, A.K., S. Pankanti, S. Prabhakar, H. Lin, and A. Ross. 2004. Biometrics: A Grand Challenge. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR),* vol. 2, 935–942.

5. Lin, Shang-Hung. 2000. An Introduction to Face Recognition Technology. *Informing Science Special Issue on Multimedia Informing Technologies—Part 2*, 3 (1): 1–5.
6. Vehils, Duque, and Jose Miguel. 1978. *Final Thesis Design and Implementation of a Finger Vein Identification System*, 2011. Cambridge, MA: Institute of Technology.
7. Amayeh, G., G. Bebis, A. Erol, and M. Nicolescu, Peg-free Hand Shape Verification Using High Order Zernike Moments. In *Proceedings of the IEEE Workshop on Biometrics at CVPR06*, New York, USA.
8. Kumar, A., D.C.M. Wong, H.C. Shen, and A.K. Jain. 2006. Personal Authentication Using Hand Images. *Pattern Recognition Letters* 27: 1478–1486.
9. Chakravorty, Tanushri. 2011. *Low Cost Subcutaneous Vein Detection System Using ARM9 Single Board Computer*. Pune: Department of Instrumentation and Control.
10. Yang, J. 2009. Combination of Gabor Wavelets and Circular Gabor Filter for Finger-vein Extraction. *Lecture Notes in Computer Science* S346–S354.
11. Chen, X., X. Bai, and X. Tao. 2013. *Chaotic Random Projection for Cancelable Biometric Key Generation*. Berlin, Heidelberg: Springer.
12. Meng, X.J., G.P. Yang, Y.L. Yin, and R.Y. Xiao. 2012. Finger Vein Recognition Based on Local Directional Code. *Sensors* 12: 14937–14952.
13. Elmir, Y., Z. Elberrichi, and R. Adjoudj. Support Vector Machine Based Finger Print Identification. In *CTCI 2012 Conference*, Adrar University, Algeria.
14. Wang, K., and Z. Yuan. 2007. Finger Vein Recognition Based on Wavelet Moment Fused with PCA Transform. *Journal Pattern Recognition and Artificial Intelligence* 20 (5): S692–S697.
15. Huafeng, Q., Q. Lan, and Y. Chengbo. 2011. Region Growth-Based Feature Extraction Method for Finger Vein Recognition. *Optical Engineering* 50: 281–307.
16. Caixia, L. 2012. *The Research on Finger Vein Image Preprocessing Based on Mathematical Morphology*. College of Information Science and Engineering, Zaozhuang University, China, London: Springer.
17. Elmir. Y. 2007. “L’identification Biométrique par les Empreintes Digitales,” Mémoire de Magister, Université des Sciences et de la Technologie d’Oran USTO-MB, LAMOSI.
18. Yin, Y., L., Liu, L.L., and X.W. Sun. 2011. SDUMLA-HMT: A Multimodal Biometric Database. In *The 6th Chinese Conference on Biometric Recognition, LNCS 7098*, 260–268. Beijing, China.
19. Malik, Iram, and Sharma Rohini. 2013. Analysis of Different Techniques for Finger-Vein Feature Extraction. *International Journal of Computer Trends and Technology* 4 (5): 1301–1305.

Design of Rheumatoid Arthritis Predictor Model Using Machine Learning Algorithms

S. Shanmugam and J. Preethi

Abstract The main aim of this paper is to investigate various data mining and machine learning techniques employed for the analysis of rheumatoid arthritis prediction based on clinical and genetic factors. The clinical characters and gene factors are collected from various hospitals in Coimbatore region through laboratory investigations from the blood serum samples and general investigations. Patients with viral fever more than six weeks and later arthritis affected compared with those patients with viral fever and no rheumatoid arthritis developed. This study involves detailed analysis of machine learning algorithms employed for rheumatoid arthritis disease, and genetic factors involved in this disease. The relevant attributes taken from the literature and consultation of rheumatologists, a combination of clinical and genetic factors evolved in this disease. The proposed model works in a big data environment named Machine Learning based Ensemble Analytic Approach (MLEAA) consists of two phases, namely learning phase and prediction phase. In learning phase data's are processed by map reduce framework in hadoop and the featured attributes are working towards prediction phase. The proposed MLEAA approach prediction phase consists of three different algorithms, namely Ababoost, SVM, ANN and based on voting system final predictive value is calculated. From this study achieve better results and it will be very useful for predict rheumatoid arthritis earlier.

Keywords Machine learning · MLEAA approach · Hadoop · Voting technique

1 Introduction

Nowadays around 45% of people affected by arthritis. It mainly focuses towards women's compared to males. Arthritis means causing inflammation in the joints. Inflammation causes redness, warmth, swelling, and pain within the joint by reducing the synovial fluid. Initially the arthritis that still not yet identified are called undifferent arthritis [15]. Once it is developed called Rheumatoid arthritis affects the whole human body, especially both the knees, wrists, and spine. It may

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_7

also affect the visceral organs like heart, blood, skin, nerves, lungs and eyes. The major symptoms faced by the peoples suffered from rheumatoid arthritis are long fever, fatigue, joint pain, swelling between joints and morning stiffness at a time of wakeup. This disease affects the people through many ways like some people may affect quicker and for some people affects gradually with symptoms over several years. Rheumatoid Arthritis affects both the men and women, but it occurs three times more than men includes at any age, but more in middle age. The cause of disease is unknown yet, but it's believed that the major cause is a combination of environmental and genetic factors involved For Diagnosing the Rheumatoid Arthritis disease cannot be diagnosed by a single test, it is a combination of multiple laboratory tests like blood tests, skin biopsy, X-rays and regular checkups for inspecting the location of painful joints. There are lots of data's are generated and stored while investigating the series of tests [6]. It is very much required to develop a rapid database that stores all the clinical and genetic data's related to RA. By analyzing these data will helpful for design a suitable predictor model. For this data mining plays an important role for classification and prediction of many diseases by processing these large data sets.

Data mining is a technique for extracting hidden information from a large set of database and it can help researchers to gain both descriptive and deep insights of understanding the complex types of sources. The major goal of using data mining is to extract the useful hidden information from large sets and to design either descriptive or predictive model. Descriptive model used to extract the common properties from the database. Predictive model is to find the future based on present values [12]. In general view data mining techniques are applied to healthcare for classification and prediction of diseases. It will help the medical practitioners for effective assessing complex diseases accurately and to create a better suitable decision support system.

For creating a predictor model using normal RDBMS is impossible. Using with help of suitable machine learning algorithms the predictive model is easily created. For measuring the accuracy the datasets are split into training and testing datasets and values are recorded. Machine Learning algorithms are works based on the past experience of data's and from their experience the model itself learned by owning and predict the future. To process a large amount and different varieties there is a need of advanced analytical processing framework.

Big data states that combination of tools and techniques for processing a large volume and complex nature of datasets in a quick period of time that are inadequate to deal with traditional processing database systems. The Challenges faced by the traditional database system includes capturing, cleaning, storing, processing, visualizing and querying the complex data formats. Big Data mainly focuses towards storing all kinds of data's, processes to extract hidden information from complex datasets that is very useful for behavior analysis, Predictive analysis for different set of domains. According to, Gartner the definition of Big data has high volume, high velocity, and/or high variety of information's that require new forms of processing to better decision making, imminent discovery and process optimization. The major characteristics of big data are tabulated at Table 1.

Table 1 Big data characteristics

S. No.	Characteristics	Explanation
1.	Volume	It represents the quantity of data generated and stored. Based on the size and potential insight it's identified as big data or not
2.	Variety	It represents the different formats and types of data. This helps to people wants to analyze it to effectively use the resulting insight
3.	Velocity	It represents the speed of data that can stored in different storage places
4.	Variability	Inconsistency of the data set can slow down processes to handle and manage it
5.	Veracity	Represents the quality of data captured can vary greatly, for accurate analysis

Medical Big Data contains various different fields involved in the medical field like textual information, numerical formats, pictorial representation that can be properly mined to extract meaningful information which is used for physicians. By mining the various medical data's several patterns are obtained and its very useful for medical practitioners to detect diseases, prediction of survivability after the disease occurs in patients and disease nature.

Nowadays the data are generated through wearable devices and stored databases processed directly through streaming and effectively analyzed. The Fig. 1 shows the different ways data's get generated and stored, real time streaming agents used in big data environment, and the operations performed with the streaming datasets are represented as 3 layers.

The rest of the paper is organized as follows. In Sect. 2 explains the related works of rheumatoid arthritis by eminent professionals in the field of medicine and computational intelligence. In Sect. 3 formulate the objectives for early prediction of rheumatoid arthritis in Sect. 3. The design details of the proposed architecture of MLEAA and important attributes taken for this work are presented in Sect. 4. Section 5 present conclusion and future works.

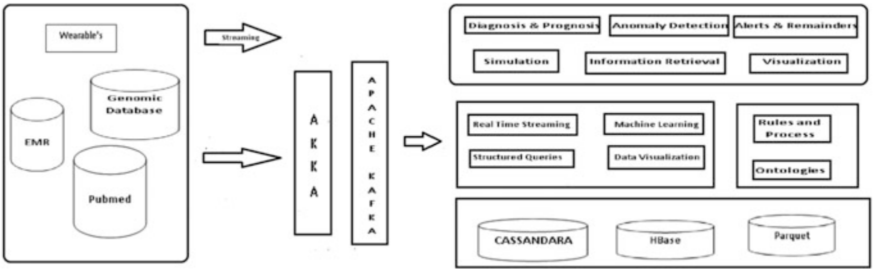


Fig. 1 General processing architecture of medical datasets using big data framework

2 Related Works

This related consists of two sections namely analysis based on clinical research and the machine learning algorithms employed for early prediction.

2.1 Survey Based on Clinical Research

Mohan et al. [1] presented review article for identifying the genetic susceptible biomarkers involved in Rheumatoid Arthritis and found that the following factors to be concentrated aggressively to immediate treatment with anti-CCP and genetic profile consisting of *HLA-DRB1*, *TNF-α*, *PADI4*, *PTPN22*, *STAT4*, *TRAF1-C5*, *CD244*, *CTLA4* and chromosome 6q 23.

Viatte et al. [2] investigated genetic risk prediction of RA using GWAS and related studies, new pathogenic pathways has been revealed, and disease nature will be clearer. Epigenetic Programming is very useful for identifying the correct gene functions between different stages and between tissue types and cells. The authors suggested in future epigenetic studies in RA to focus on the correct cell types will move towards targeted approaches in correct biological pathways.

Bridges et al. [3] has been explored the knowledge of genetic variations and influences of genetics for complex autoimmune and inflammatory diseases such as RA among different populations. The authors stated that class II MHC alleles are important contributors, there are likely to incorporate multiples of other genes that modulate the disease phenotype. With the help of genetic markers allows treatment response will determined, especially in light of the growing number of biologic agents undergoing clinical trials.

Soroka et al. [4] work towards identifying the adverse reactions of genes among genetic biomarkers for using sulphasalazine in rheumatoid arthritis patients and found that gastrointestinal Adverse Drug Reactions with sulphasalazine were registered in 20.7% (12/58) of patients, mucocutaneous ADR in 8.6%, urine analysis abnormalities in 8.6%, and haematological abnormalities in 5.2% of patients.

Oliver et al. [5] presented an review article to identify the success rate of anti-TNF drugs towards treatment of Rheumatoid Arthritis disease and studied the association between respondent and non respondent genetic markers using genomic wide association studies. The authors suggested to enhance the general studies with international collaboration and to discovery with CD84 association.

Scott et al. [6] work towards predicting the risk factors of rheumatoid arthritis based on age and disease onset through a modeling approach based on genetic variants with smoking. In this research work combination of ratios for 15 four-digit/10 two-digit HLA-DRB1 alleles, 31 single nucleotide polymorphisms (SNPs) and ever-smoking status in males to determine risk using computer simulation and confidence interval based risk categorization. From the study author concluded that the clinical information for RA prediction with Human Leukocyte Antigen and smoking status will very useful for predicting the risk of younger and older peoples.

Briggs et al. [7] focus towards finding the genetic interactions a phenomenon called epistatic by combining the laboratory findings with statistical methods. By this work author performed multistage analysis using supervised machine learning approach and association testing to investigate interactions with genetic factors (PTPN22 1858T) and complex disease Rheumatoid Arthritis (RA). The work consists of four stages are like predict chromosomal regions of 292 patients using random forest, candidate chromosomal regions for epistasis with PTPN22 1858T in 677 cases and 750 controls using logistic regression, to create Evidence of epistatic interactions in 947 cases and 1756 controls, Pooled analysis includes all 1624 RA cases and 2506 control subjects for estimate final effect size. Authors identified total of seven replicating epistatic interactions involved SNP variants within CDH13, MYO3A, CEP72 and near WFDC1 showed significant evidence for interaction with PTPN22, affecting susceptibility to RA.

Taniya et al. [8] classified novel gene structures for disease-susceptibility depending on the biological similarities to the known disease causing genes. Datasets are extracted from H-InvDB and RefSeq database and by analyzing the seven important features sequence similarity, InterPro annotation, EC number, three kinds of GO terms, and KEGG pathways. Rheumatoid Arthritis and Prostate Cancer are taken up for this study and the accuracy are evaluated based on highly scored genes obtained from TNFSF12 and OSM as candidate disease genes for RA and PC, respectively. Consequent classification of these genes based upon an all-embracing journals reinforced the validity of these highly scored genes as possible disease-susceptibility genes. By using, Prioritization Analysis of Disease Association (PANDA), is used to find an efficient and cost-effective method to short down a large set of genes into smaller subsequent that are most likely to be involved in the disease pathogenesis.

Paunic et al. [9] proposed a method by inferring HLA alleles from commonly available and reasonably SNP genotype data by taking into account the high linkage disequilibrium that exists in the region. For this study Human Leukocyte Antigen (HLA) gene system is very important to identify the variations in genes in the human body. By using HLA typing in human genomes to study the association between inflammation, infection, and with autoimmune diseases between normal and affected patients. Genomic Wide Association Studies are used for finding a strong association with HLA alleles are available through Single Nucleotide Polymorphism genome data. The authors mentioned that there is a great use of having complete HLA data's for clinicians and researchers, but it's highly time consuming and cost-prohibitive process.

2.2 Survey Based on Machine Learning Algorithms

Author	Year	Knowledge type	Knowledge resource	DM techniques/ applications
Briggs et al. [10]	2010	Rheumatoid arthritis	Identifying PTPN22 is interaction with SNP variants within CDH13, MYO3A, CEP72 and near WFDC1	Random forest, logistic regression
Yang et al. [11]	2014	Disease gene identification	EPU is able to achieve 84.8% in terms of F-measure, which is 3.5%, 15.3% and 16.1% better than MSVM, WNB and WKNN respectively	Ensembled approach namely nearest neighbor, naive Bayes and SVM
Shiezadeh et al. [12]	2015	Predictor model for rheumatoid arthritis	Compared to SVM, ANN, decision tree, AdaBoost with cuckoo search are combined named CS-boost gets higher accuracy	Adboost with decision stump as weak learner, cuckoo search
Feng et al. [13]	2015	Classifying RA based on clinical trial metadata	ACR 20, DAS 28, Adverse events and serious adverse events, Adverse events have accuracy higher 82.70% then other approaches	Random forest algorithm
Nair et al. [14]	2009	Classification of possibility of RA with gait using machine learning algorithms	Comparison and to classify RA affected patients and osteoarthritis. between Least Squares Kernel and Neural networks like SOM, LVQ, MP	Least squares kernal outperforms best classification
Garcia-Zapirain et al. [15]	2015	Classifying the patients affected by fibromyalgia and arthritis using machine learning	Accuracy level is achieved with medical-social features 89.9474% and the success rate for psychopathology features are 96.4035% and joining the two achieved rate is 95.8246%	AdaBoost algorithm

3 Objectives

(I) Rheumatoid Arthritis—Study based on computational techniques so far applied and genetic factors involved.
(II) To analyze the ways of using medical data mining for classification and prediction in rheumatoid arthritis and to identify an appropriate clinical and genetic involved in working with big data framework.
(III) To propose an effective framework for predicting rheumatoid arthritis earlier.

4 Proposed Architecture

The proposed architecture Fig. 2 consists of two phases. The first phase is learning phase and the second phase is prediction phase.

Phase 1: Learning phase—it consists of patient's raw information such as clinical and genetic factors that are collected are moved into hadoop framework. In hadoop framework the mapper takes charge of input data's, splitted into small chunks of files stored under different systems for parallel processing. The reducer gathers all the data's by combining the similar groups based on key-value pairs. used for predicting the new data from the knowledge obtained from learning phase. From the output of map reducer we get the important features of Rheumatoid Arthritis disease for early prediction.

Phase 2: Prediction phase: In this phase the extracted important features apply to help of rapid miner tool to different famous state-of-art machine learning algorithms like AdaBoost, SVM, ANN which produces better accuracy and sensitivity to get

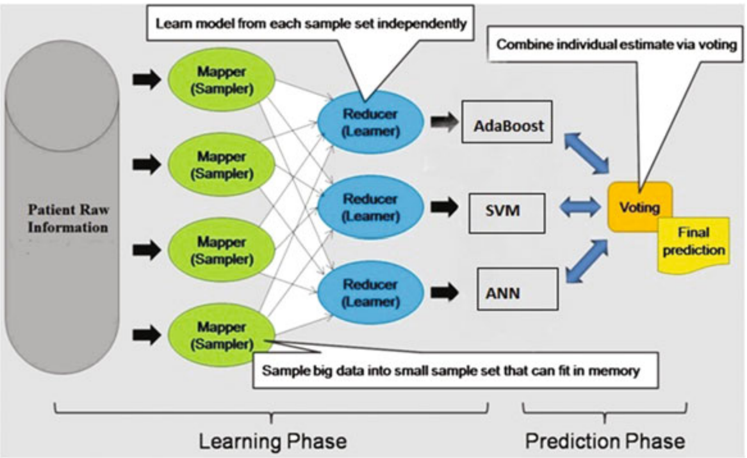


Fig. 2 Proposed MLEAA architecture

Table 2 Accuracy of classification algorithms

S. No.	Technique	Accuracy (%)
1.	AdaBoost	85
2.	SVM	75.4
3.	ANN	72.2
4.	Naive Bayes	71.1

better decisions. The data's are applied separately to the above mentioned algorithms and the results are combined for final prediction. Based on voting method accurate values are predicted between affected RA and unaffected RA patients. The results will help the doctors for better diagnosing the patients effectively and treat as earlier.

The various algorithms for this proposed work are chosen from various research works gives better accuracy for disease classification. The accuracy obtained from different classification algorithms are tabulated (Table 2).

From this we have selected top 3 classification algorithms which gives more accuracy for our proposed MLEAA approach using the rapid miner tool and the result is evaluated. The data sets are selected with the consultation of rheumatologist in Coimbatore Government Medical college the primary factors and attributes taken for early prediction using machine learning is mentioned in Table 3. Even though all these are primary factors consulting with three more physicians to rank this attributes and major features are selected as tabulated in Tables 3 and 4.

Table 3 Major Attributes involved for Rheumatoid Arthritis

S. No.	Attributes	Meaning
1.	Gender F, M	F—Female M—Male
2.	Age 0, 1, 2	0: <=20 1: >=21 and <=50 2: >51
3.	Sample type S, P, U	S: Serum, P: Plasma, U: Urine
4.	Disease activity F, M, S	F: Flare, M: Moderate, S: Severe
5.	ACR criteria 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	1: Malar rash 2: Discoid rash 3: Photosensitivity 4: Oral ulcers 5: Non erosive disorder 6: Pleuritis 7: Renal disorders 8: Neurologic disorders 9: Hematologic disorder 10: Immunologic disorder 11: Antinuclear antibody

(continued)

Table 3 (continued)

S. No.	Attributes	Meaning
6.	Joints involved 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	0: Spine 1: Proximal interphalangeal 2: Distal interphalangeal 3: Metacarpophalangeal 4: Wrist 5: Shoulder 6: Hip 7: Knee 8: Ankle 9: Metatarsophalangeal 10: Elbow 11: Foot proximal interphalangeal
7.	Extra auricular involved 0, 1, 2, 3, 4, 5	0: Eye 1: Heart 2: Lungs 3: Blood vessels 4: Muscle
8.	Blood investigations 0, 1, 2, 3, 4, 5, 6, 7, 8	0: Total count 1: Differential count 2: Erythrocyte sedimentation rate 3: Haemoglobin 4: Platlets 5: Renal function test 6: Rheumatoid factor 7: Anti-CCP 8: Antinuclear antibody
9.	HLA allele	Class A, B, DR

Table 4 Ranking of criteria by clinicians for prediction of rheumatoid arthritis disease

Clinician 1	Clinician 2	Clinician 3
Age Disease onset Anti nuclear antibody Rheumatoid factor Joints involved ACR criteria 2, 3, 4 HLA DRB1	Age Gender Nature of disease Anti nuclear antibody Rheumatoid factor ACR criteria 1, 3, 4 HLA DRB1	Age Disease onset Anti nuclear antibody Rheumatoid factor ACR criteria 2, 3, 4

5 Conclusion and Future Scope

The main purpose of this survey was to discover the most typical machine learning algorithms used for predicting rheumatoid arthritis disease. The ideas for future work includes the evaluation of choosing algorithms on the basis of real time rheumatoid arthritis disease with genetic variations. Other algorithms can be applied to the built-in dataset and the algorithm which gives best result will be applied on the test dataset. The experiments would be conducted on the selected medical records which design the analysis even more accurate. The good idea is taking also other algorithms to the experiments and compares their performance in medical field. This would evolve a new class and assist in scheming Medical Decision Support Systems by the selection of the most acceptable algorithms.

References

1. Mohan, Vasanth Konda, Ganesan, Nalini, and Rajasekhar, Gopalakrishnan. 2014. Association of Susceptible Genetic Markers and Autoantibodies in Rheumatoid Arthritis. *Journal of Genetics* 93(2): 597–605.
2. Viatte, Sebastien, Darren Plant, and Soumya Raychaudhuri. 2013. Genetics and Epigenetics of Rheumatoid Arthritis. *Nature Reviews Rheumatology* 9: 141–153. https://doi.org/10.1038/nrrheum.2012.237. published online 5 February 2013.
3. Bridges Jr., S. Louis, and Robert P. Kimberly. 2002. Genetic Influences on Treatment Response in Rheumatoid Arthritis. In *Modern Therapeutics in Rheumatic Diseases*, ed. G.C. Tsokos, et al. Totowa, NJ: Humana Press Inc.
4. Soroka, N., et al. 2012. Genetic Markers of Sulphasalazine Adverse Reactions in Rheumatoid Arthritis Patients. *Annals of the Rheumatic Diseases* 71 (Suppl3): 68.
5. Oliver, James, Darren Plant, Amy P. Webster, and Anne Barton. 2015. Genetic and Genomic Markers of Anti-TNF Treatment Response in Rheumatoid Arthritis. *Biomarkers in Medicine* 9 (6): 499–512.
6. Scott, Ian C., et al. 2013. Predicting the Risk of Rheumatoid Arthritis and Its Age of Onset through Modelling Genetic Risk Variants with Smoking. *PLOS Genetics* 9 (9): 1–14.
7. Briggs, F.B.S., and P.P. Ramsay. 2010. Supervised Machine Learning and Logistic Regression Identifies Novel Epistatic Risk Factors with PTPN22 for Rheumatoid Arthritis. *Genes and Immunity* 11: 199–208.
8. Takayuki, Taniya, et al. 2012. A Prioritization Analysis of Disease Association by Data-mining of Functional Annotation of Human Genes. *Genomics* 99: 1–9. https://doi.org/10.1016/j.ygeno.2011.10.002. ISSN:0888-7543.
9. Vanja, Paunic, Michael, Steinbach, Vipin, Kumar, Martin, Maiers. 2012. Prediction of HLA Genes from SNP Data and HLA Haplotype Frequencies. In *2012 IEEE 12th International Conference on Data Mining Workshops*.
10. Briggs, F.B.S., P.P. Ramsay, et al. 2010. Supervised Machine Learning and Logistic Regression Identifies Novel Epistatic Risk Factors with PTPN22 for Rheumatoid Arthritis. *Genes Immunity* 11 (3): 199–208. https://doi.org/10.1038/gene.2009.110.
11. Yang, Peng, Xiaoli, Li, Hon-Nian, Chua, Chee-Keong, Kwoh, See-Kiong, Ng. 2014. Ensemble Positive Unlabeled Learning for Disease Gene Identification. *PlusOne*. https://doi.org/10.1371/journal.pone.0097079.

12. Zahra, Shiezadeh, Hedieh, Sajedi and Elham Aflakie. 2015. Diagnosis of Rheumatoid Arthritis Using an Ensemble Learning Approach, 139–148. ICAITA, SAI, CDKP, Signal, NCO-2015.
13. Feng, Yuanyuan, Vandana P. Janeja, et al. 2015. Poster: Classifying Primary Outcomes in Rheumatoid Arthritis: Knowledge Discovery from Clinical Trial Metadata. *IEEE Transactions on Information Technology in Biomedicine* 10 (2): 254–263.
14. Nair, Sumitra S., Robert M. French, Davy Laroche, and Elizabeth Thomas. 2010. The Application of Machine Learning Algorithms to the Analysis of Electromyographic Patterns from Arthritic Patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 4: 1–10.
15. Garcia-Zapirain, Begoña, Garcia-Chimeno Yolanda, et al. 2015. Machine Learning Techniques for Automatic Classification of Patients with Fibromyalgia and Arthritis. *International Journal of Computer Trends and Technology (IJCTT)* 25 (3): 149–152.

Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic

B. Vamshi Krishna, Ajeet Kumar Pandey and A. P. Siva Kumar

Abstract This paper discusses a new model towards opinion mining and sentiment analysis of the text reviews posted in social media sites which are mostly in unstructured format. In recent years, web forums and social media has become an excellent platform to express or share opinions in the form of text about any product or any interested topic. These opinions are used for making decisions to choose a product or any entity. Opinion mining and sentiment analysis are related in a sense that opining mining deals with analyzing and summarizing expressed opinions whereas sentiment analysis classifies opinionated text into positive and negative. Feature extraction is a crucial problem in sentiment analysis. Model proposed in the paper utilizes machine learning techniques and fuzzy approach for opinion mining and classification of sentiment on textual reviews. The goal is to automate the process of mining attitudes, opinions and hidden emotions from text.

Keywords Opinion mining · Sentiment analysis · Machine learning
Natural language processing

1 Introduction

Nowadays, social media sites like Twitter, Facebook, blogs, LinkedIn etc. have become an excellent platform to share information or opinions mostly in the form of text. These social networking sites are used for exchanging views or opinions about a product, movies, and politics or about any user interested topics in the form of posting comments, pictures and get feedback from other users. This kind of user generated text on social web forums about any products, people, and any events is very useful in business, government and individual. Data from social networking sites are actively mined for trends and patterns of interests.

Opinion mining is a text mining problem. An opinion is defined as a quadruple, (g, s, h, t), where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed [1].

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_8

Opinion mining is a technique which analyzes and summarizes the opinions expressed in the form of huge text data. Sentiment analysis classifies opinions in text into different classes like positive, negative or neutral. Opinion mining deals with retrieving, classifying, analyzing and assessing the opinions in various online platforms which is in the form of text. Sentiment analysis aims to identify sentiment, affect, subjectivity and other emotional states hidden in the text. This paper proposes a feature based model for opining mining and sentiment analysis of unstructured textual reviews using data mining techniques [2] and fuzzy logic [3].

Opinion mining can be performed in various levels. Few of them are document level, sentence level or even word level and feature level. Document level is used to identify the overall sentiment expressed on an entity or object. Sentence level is used to identify overall sentiment or opinions about an entity but does not associate opinions with specific aspects or features of an object. To gain fine grained opinion analysis, need to deal with opinion feature or in short feature level of the product. Feature based opinion mining also known as aspect based opinion mining and is an effective technique which aims to identify opinion holding features and performs sentiment analysis.

Fine grained opinion may be used for purchasing a product or help in decision making. Feature identification and extraction is a sub problem in the opinion mining process. Every object and entity consists of implicit and explicit features. Explicit features are the features which are directly mentioned in reviews typically noun or noun phrase whereas implicit features are the features which are not mentioned in review but it is implied in the sentence of a review.

Many techniques have been identified to extract opinion features in the process of opinion mining. Supervised learning approaches which are trained perform well in the given domain, but the model shall undergo training to apply for another domain. Unsupervised learning approaches extract opinion features with the help of natural language processing techniques by defining syntactic rules for capturing the feature terms.

Topic modeling approaches aims for mining fine grained topics or aspects. Topic or aspect extraction aims to extract aspects of entities for which opinions are mentioned in the text. This aspect extraction is done in two stages. The first stage task is to extract aspect terms from an opinion source and in the second stage task is to group synonymous terms into an aspect category [4].

Rest of the paper is organized as follows: the following section presents literature survey related to the problem. Section 3 gives the brief idea about the classification techniques and fuzzy set theory. Section 4 describes the proposed model. Section 5 contains results and 6 conclusions and future work.

2 Related Works

A. *Opinion Mining*

Opinion mining and sentiment analysis is the new research domain in which various efforts have been made for sentiment classification by using machine learning techniques. On reviewing literature, it is found that various supervised machine learning algorithms like Naïve Bayes [5, 6] Support Vector Machines [7, 8] and Neural Networks [9] have been used for opinion mining of text to classify positive and negative sentiment. Supervised learning approaches performs well for only trained domain and shall be retrained for opining mining of another domain.

Topic modeling is a method for identifying "topics" in the collection of opinionated topics. Unsupervised topic models aims to identify topics which are represented as probabilistic distributions on words from a source of text. Latent Dirichlet Allocation (LDA) [10] is most popularly used topic modeling approach for aspect extraction which is a probabilistic graphical model.

Sequence models are popularly used for retrieving information and aspect extraction that can be viewed as sequence labeling tasks as product entities, aspects and text expressions can be considered as interdependent and occur in a sequence in a textual review. Hidden Markov Model (HMM) [11] is a directed sequence model which works for a multiple range of state series data. Conditional Random Fields (CRFs) is also a sequence model [12]. Bayesian models are used to classify text into multiple emotional categories and resulting into multi-dimensional sentiment classification [13].

B. *Opinion Feature Extraction*

Opinion feature extraction is a crucial problem in the process of opinion mining. Fine grained feature extraction is required for effective sentiment analysis. Each product has its own set of features and product reviews are about product features are good indicators in classifying the sentiment of product reviews.

Feature extraction step needs to identify opinionated features and sentiment holding words and in the given set of reviews. Natural language processing methods POS-Taggers shall be used to identify features and opinion holding words. Lexicon based approaches takes help of external lexicons which has predefined positive and negative scores and increases the performance of a sentiment classifier. Few standard lexicons are WordNet, ConceptNet, SenticNet and SentiWordNet [14]. Aspect based models are used for extracting aspects which are unsupervised and based on topic models Few models use seed words and are semi-supervised which jointly model both aspects and aspect specific sentiments [4].

From the literature it is found supervised machine learning methods like Naïve Bayes, Maximum Entropy, Support vector machines (SVM) and Neural Networks are most popularly used to classify sentiment. Among all these techniques, it is found that Support vector machines (SVM) is an efficient classification technique.

But the user generated text on social networking sites are more unstructured and fuzziness in nature. From the literature, it has been found that fuzzy lexicon [15–18] and fuzzy sets [19] are efficient to classify the sentiment among the user text. Pandey and Goyal [20, 21] used fuzzy logic for early software fault prediction and improved reliability of software systems.

C. *Research Motivation*

After the review of many papers, we have found many gaps in the area of sentiment classification. As the most of the machine learning methods performs binary classification either classify sentiment or emotion as positive or negative sentiment. Also user generated text on social networking sites are more unstructured and fuzziness in nature.

As fuzzy sets are capable of containing elements that have variable degree of membership and can be used to increase the degree of polarity of the expressed opinions. The opinionated word can be classified more accurately with the help of fuzzy sets, words like “Excellent” and “Good” are given different degree of polarity.

Fuzzy inference system (FIS) shall be used to summarize the opinions with more levels. This paper integrates support vector machines (SVM) algorithm with fuzzy inference system to predict sentiment polarity of textual reviews.

3 Research Background

A. *Data Mining*

Data mining techniques deal with large amounts of data sets and extract knowledge or patterns. In the literature various data mining techniques are discussed like regression, classification, associations and clustering. Classification technique results in a model, which is an automatic way of classifying the data based on predefined class labels for classifying future data points and predicting the characteristics [2].

In the literature, many classification techniques are found which are regression, linear and quadratic discriminant analysis, ID3, C4.5, k-nearest neighbor, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [7, 8]. SVM techniques are most effective used in text classification, and has high levels of performance compared to Naive Bayes [5, 6]. SVM partions positive and negative training vector sets by using hyper-plane with maximum margin.

B. *Opinion Mining*

Opinion mining process deals with identifying, extracting and sentiment classification which is hidden in text messages. Opinion mining consists of three main steps data collection, data preprocessing and data processing followed by opinion summarization. Data processing consists of sub phase’s feature extraction and sentiment classification. Opinion summarization is a method of analyzing and summarizing the mining results.

i. Data collection step consists of collecting text reviews posted in social networking web sites which are repository of web documents. Textual reviews collected in the form of sentences may be about a product opinion or any user interested topic.
ii. In data preprocessing step, text reviews collected from various sources of document is broken into tokens, stop words are eliminated. Natural Language processing techniques like stemming and Parts of speech (POS) Tagging are performed. As a result of POS Tagging process, only opinion holding words are extracted based on their part of speech corresponding verbs, adverbs and adjectives.
iii. Data processing step consists of identifying and classifying the sentiment from text messages. Sentiment classification step classifies the reviews into a positive, negative or neutral. Both supervised learning and unsupervised algorithms are used for classifying the sentiment among opinion holding words.

C. *Fuzzy Set Theory*

Classical or crisp set is a collection of well defined distinct objects. These crisp sets contain objects that satisfy well precise properties of membership. For a crisp set, an element x in the universe X is either a member of some crisp set (A) or not and binary issue of membership can be represented by a characteristic function as:

$$\gamma_A(x) = \begin{cases} 1, & if\ x \in A \\ 0, & if\ x \notin A \end{cases}$$

where, χ_A(x) gives an unambiguous membership of the element, x in a set A.

Fuzzy set is a having clear defined boundary and also contain elements with a partial degree of membership. Suppose Ã is a fuzzy set of A, if an element in the universe, say, x is a member of fuzzy set Ã then this mapping is given by a membership function μÃ(x). The membership function μÃ(x), gives the degree of membership for each element in the fuzzy set Ã and is defined in range [0, 1] where 1 represents elements that are completely in Ã, 0 represents elements that are completely not in Ã, and values between 0 and 1 represent partial membership in Ã. In Zadeh's notation [19], a fuzzy set Ã can be represented as:

$$\widetilde{\mathrm{A}} = \left\{ \frac{\mu_1}{x_1} + \frac{\mu_2}{x_2} + \frac{\mu_3}{x_n} + \cdots + \frac{\mu_n}{x_n} \right\}$$

where, $\mu_{1,}$ $\mu_{2,}$ μ_3 ... μ_n are the membership value of the elements x1, x2 … xn respectively in the fuzzy set Ã. Let FP is a fuzzy set representing collection of features of an entity f1, f2, f3 … fn. Let $\mu_{1,}$ $\mu_{2,}$ μ_3 ... μ_n features are the membership value of features f1, f2 … fn respectively. These membership values represent the degree of polarity of a feature, and a fuzzy set of polarity of feature can be described using Zadeh's notation [19] as:

$$F\acute{P}\left\{\frac{\mu_1}{f_1}+\frac{\mu_2}{f_2}+\frac{\mu_3}{f_3}+\cdots+\frac{\mu_n}{f_n}\right\}$$

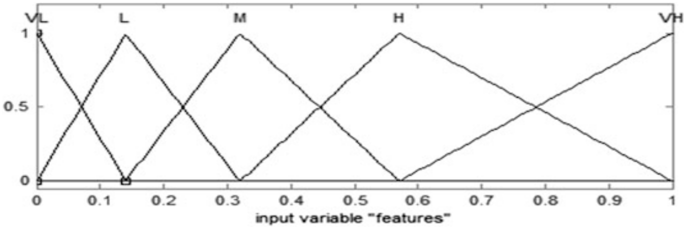
D. *Fuzzy Profile Development*

Features related to a particular product of three kinds branding features, sentiment features and product feature itself. From the given text review, with each identified feature set and extracted opinion holding words for each feature shall be provided as input for fuzzy inference system and output is the degree of polarity.

Membership functions of a product features can be build by using any of the suitable method among rectangular, triangular, trapezoidal and gamma [3].

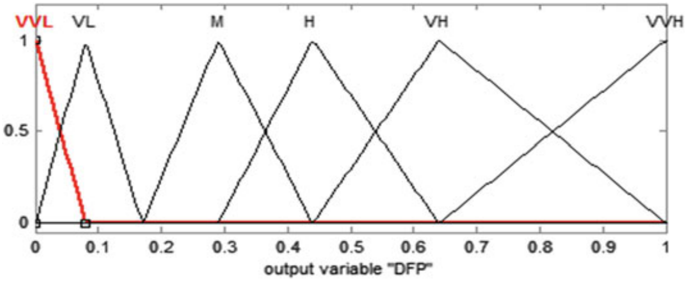
Input variables for this model are opinion features which are of logarithmic in nature and are divided into five linguistic categories such as very low (VL) (0; 0; 0.14), low (L) (0; 0.14; 0.32), medium (M) (0.14; 0.32; 0.57), high (H) (0.32; 0.57; 1.0) and very high (VH) (0.57; 1.0; 1.0). By using these five categories, fuzzy profile ranges (FPR) of opinion features are developed using the following formula:

$$FPR=\left[1-\frac{\{\log_{10}(1{:}5)\}}{\log_{10}5}\right]$$



Output of this model is DFP (Feature-polarity Degree), also follows logarithmic scale and can be divided into seven linguistic categories such as: very very low (VVL) (0; 0; 0.08), very low (VL) (0; 0.08; 0.17), medium (M) (0.17; 0.29; 0.44), high (H) (0.29; 0.44; 0.64), very high (VH) (0.44; 0.64; 1.0) and very very high (VVH) (0.64; 1.0; 1.0). Fuzzy profile range (FPR) of DFP can be developed by using below formula:

$$FPR=\left[1-\frac{\{\log_{10}(1{:}7)\}}{\log_{10}7}\right]$$



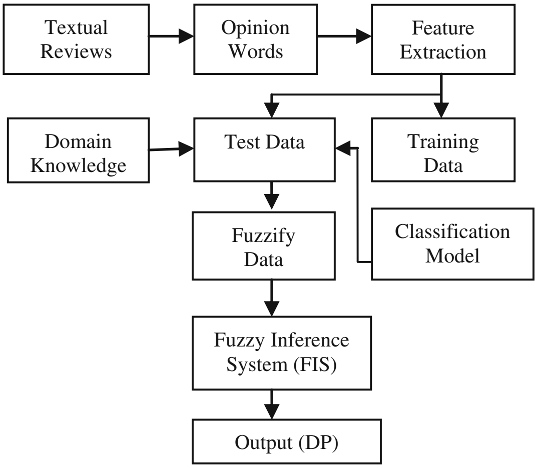
4 Proposed Work

The model architecture is shown in Fig. 1, which uses data mining techniques and fuzzy logic which can be implemented by using R programming language and fuzzy logic toolbox of MATLAB. Steps in this model are identification of input-output (I-O) variables, development of fuzzy profiles by using I-O variables, defining relationship between I-O variables with the help of fuzzy inference system. Features of a particular product are considered as input variables and the output variable is degree of polarity (DP) of the proposed model.

A. *Data-Set Used*

Twitter is a vast area or platform to exchange opinions or information for unstructured text and streaming data. Text reviews in the form of statements are collected from twitter about any product information or any user interested topic. R language packages are used to access twitter data and searched for the keyword

Fig. 1 Proposed model architecture



“Nokia” and “Phone” and 500 recent reviews are captured and tweets which are matching with the key words are captured in a corpus.

B. *Data Preprocessing*

The general concept of data preprocessing has already been discussed in the Sect. 3. This section discusses the pre processing steps used in this proposed model. Text reviews which are matching with the key words are captured in a corpus, broken into tokens and stop words are eliminated. Natural Language processing techniques stemming and Parts of speech (POS) Tagging are used and opinion holding words are extracted and document term matrix is build.

C. *Feature Extraction*

Features of the objects are extracted at every sentence and subjective features are classified. Opinion holding words are extracted and classified as positive and negative and their degree of polarity is assigned with the help of fuzzy sets.

D. *Data Processing*

The general concept of data processing has already been discussed in Sect. 3. In the proposed model, supervised learning algorithms Support vector machines (SVM) and Maximum entropy are used for classifying the sentiment among opinion holding words.

E. *Sentiment Classification*

R programming language packages are used for sentiment classification purpose. Training model is built by using the below steps:

i. Create a document-term matrix
ii. Create a container
iii. Create a model by feeding a container to the machine learning algorithm
iv. Test the model.

F. *Sentiment Prediction*

Fuzzy sets are used with variable degree of membership and degree of polarity of the expressed opinions is improved by summarize the opinions with more levels.

5 Results and Discussion

Results were based on the data set which has already mentioned in Sect. 4 and we examined the support vector machine and maximum entropy algorithms for sentiment classification purpose. Results were compared with the selected data set to classify the sentiment in the text. Compute the quality parameters such as completeness, correctness, efficiency and effectiveness and overall error rate (misclassification).

Also, compared the classification accuracy of the methods and compute training time of learning models. It is found support vector machines are most widely used in sentiment classification and are having high performance classification rate. For precision and recall we use the below equations:

(a) *Precision*:

P = (Correctly extracted features)/(Total extracted features)

(b) *Recall*:

R = (Correctly extracted features)/(Total manually labeled features)

(c) *Fscore*:

F = (2 * Precision * Recall)/(Precision + Recall)

Table 1 shows the results of predictions from trained model by using support vector machine (SVM) and maximum entropy (MAXENT) classification algorithms.

Table 1 Classification results

MAXENTROPY_LABEL	MAXENTROPY_PROB	SVM_LABEL	SVM_PROB
20	0.50050587	19	0.2000935
29	0.54993552	19	0.2036404
16	0.99998085	19	0.2184277
20	0.50045149	19	0.2010367
20	0.50063741	19	0.1965323
16	0.04892506	19	0.2102050
15	0.50101570	19	0.1963483
12	0.39806081	19	0.2305578
19	1.00000000	16	0.2166676
29	0.99979569	19	0.2094373
20	0.50056562	19	0.1986995
20	0.99999939	19	0.2342912
15	0.50101570	19	0.1963483
16	0.99740386	19	0.2137098
31	0.89294403	19	0.2192618
17	0.57887640	19	0.2047166
19	0.38647755	19	0.2061272
16	0.64896279	19	0.2159892
19	1.00000000	16	0.2143635
16	0.62952345	19	0.2048942
1	0.28536967	19	0.1609520
19	0.49972709	19	0.1066619
19	0.49972709	19	0.1066619
19	0.49972709	19	0.1066619
15	0.22119833	19	0.2075367

Table 2 Accuracy results

Algorithm	Precision	Recall	Fscore
SVMM	0.22	1	0.36
MAX	0.20	1	0.33

Table 3 Cross validation results

Algorithm	Mean accuracy
SVM algorithm	0.1417102
MAXENT algorithm	0.07568438

Table 2 shows the summary of results with precision, recall and Fscore and it is clear that SVM has high accuracy when compared to MAXENT algorithm. Table 3 shows mean accuracy of n-fold cross validation results of these used algorithms.

6 Conclusion

A model for feature based Opining mining and Sentiment Analysis is presented in this paper. This model is based on feature extraction of the products and their degree of polarity is converted into fuzzy sets. The proposed model is used to analyze user opinions and reviews posted on social media websites and helps users in decision making to buy products and organizations to recommend products online.

Opinion spam detection or fake review detection and improving reliability of opinion mining could be taken as future work.

References

1. Bing, Liu. 2012. *Sentiment Analysis and Opinion Mining*. USA: Morgan & Claypool publishers.
2. Han, J., and M. Kamber. 2001. *Data Mining: Concepts and Techniques*. USA: Morgan Kaufmann Publishers.
3. Ross, T.J. 2010. *Fuzzy Logic with Engineering Applications*, 3rd ed. India: Willy India Pvt. Ltd.
4. Mukherjee, Arjun, and Bing Liu. 2012. Aspect Extraction through Semi-Supervised Modeling. *Association for Computational Linguistics* 26 (3): 339–348.
5. Dinu, Liviu P., and Iulia Iuga. 2012. *The Naive Bayes Classifier in Opinion Mining. In Search of the Best Feature Set*. Berlin: Springer.
6. Xiuzhen, Zhang, and Yun Zhou. 2011. *Holistic Approaches to Identifying the Sentiment of Blogs Using Opinion Words.* Berlin: Springer 5–28.
7. Soliman, Taysir Hassan A., Mostafa A. Elmasry, Abdel Rahman Hedar, and M. M. Doss. 2012. *Utilizing Support Vector Machines in Mining Online Customer Reviews*. ICCTA.
8. Kwon, Ye Jin, and Young Bom Park. 2011. *A Study on Automatic Analysis of Social Network Services Using Opinion Mining*. Springer, Berlin, 240–248.

9. Sharma, Anuj, and Shubhamoy Dey. 2012. *An Artificial Neural Network Based approach for Sentiment Analysis of Opinionated Text.* USA: ACM.
10. Moghaddam, Samaneh, and Martin Ester. 2012. *On the Design of LDA Models for Aspect-based Opinion Mining.* USA: ACM.
11. Jin, Wei, Hung Hay Ho, and Rohini K. Srihari. 2009. *OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction.* USA: ACM.
12. Yang, Bishan, and Claire Cardie. 2012. *Extracting Opinion Expressions with Semi-Markov Conditional Random Fields.* Association for Computational linguistics.
13. He, Yulan. 2012. *A Bayesian Modeling Approach to Multi-Dimensional Sentiment Distributions Prediction.* USA: ACM.
14. Mudinas, Andrius, Dell Zhang, and Mark Levene. 2012. *Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis.* USA: ACM.
15. Jusoh, Shaidah, and Hejab M. Alfawareh. 2013. *Applying Fuzzy Sets for Opinion Mining.* IEEE.
16. Ding, Xiaowen, Bing Liu, and Philip S. Yu. 2008. *A Holistic Lexicon-Based Approach to Opinion Mining.* USA: ACM.
17. Cheng, Li-Chen, and Hua-An Wang. 2011. *A Novel Fuzzy Recommendation System Integrated the Experts' Opinion*. IEEE International Conference on Fuzzy Systems.
18. Goeuriot, Lorraine, Jin-Cheon Na, Wai Yan Min Kyaing, Christopher Khoo,Yun-Ke Chang1, Yin-Leng Theng, and Jung-Jae Kim. 2012. *Sentiment Lexicons for Health-related Opinion Mining*. USA: ACM.
19. Zadeh, L.A. 1965. Fuzzy Sets. *Information and Control* 8: 338–353.
20. Pandey, A.K., and N.K. Goyal. 2009. A Fuzzy Model for Early Software Fault Prediction Using Process Maturity and Software Metrics. *International Journal of Electronics Engineering* 239–245.
21. Pandey, A.K., and N.K. Goyal. 2010. *Predicting Fault-prone Software Module Using Data Mining Technique and Fuzzy Logic.* International Conference.

Hexagonal Image Processing and Transformations: A Practical Approach Using R

E. Ramalakshmi and Neeharika Kompala

Abstract Hexagonal structure is remarkable in connection to the standard square structure for picture depiction. The geometrical course of action of pixels on hexagonal structure can be portrayed similar to a hexagonal system. Hexagonal structure gives a straightforward way to deal with picture translation and turn information. Winding Architecture is a reasonably new and competent approach to manage machine vision structure. Regardless, all the present hardware for finding picture and for indicating picture are made in light of rectangular building. It has transformed into a noteworthy issue impacting the pushed research on Spiral Architecture. In this paper, another approach to manage Spiral Architecture is presented using R. This duplicate Spiral Architecture for all intents and purposes holds picture assurance and does not present contorting. Also, pictures can be effortlessly and adequately traded between the regular square structure and this new hexagonal structure. R language plays a pivotal role in development of numerical analysis and machine spaces. R language is the best way to create reproducible and high-quality analysis. It has all the flexibility and power needed to be dealing with data.

Keywords Image processing · Hexagonal pixel · Image transformation
Location of pixel · R language

1 Why Is R Used in Processing and Transformations

R language plays a major role in development of numerical analysis and machine spaces. R language is the best way to create reproducible and high-quality analysis. It has all the flexibility and power needed to be dealing with data. Also to implement the concept of parallelism, R language provides many opportunities.

Given the sort of picture of value extending from great to terrible, the fundamental preparing begins from rebuilding of the information from picture to a configuration satisfactory by the utilized calculation. Straight square relapse at that point forms the present information to show the picture such that it can be adequate [1].

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_9

Once a satisfactory stage has been achieved we advance into thresholding and inclination boosting of the picture to further make the information into a well respectable way, where issues in regards to edge based thresholding are secured [2, 3]. Assist on components with respect to versatile thresholding are altogether required in preparing.

2 Hexagonal Grid: Compendium

An electronic picture contains an enormous number of pixels to address this present reality and when we touch the expression "pixel" as of not long ago, that infers a rectangular box in a photo. All the past picture get ready and picture examination look into relies on upon this customary picture structure. The upsides of using a hexagonal system to address digit pictures have been investigated for more than thirty years. The centrality of the hexagonal depiction is that it has exceptional computational segments that are related to the vision method. Many reports portraying the advantages of using such a system sort have been found in the composition. The hexagonal picture structure has segments of larger amount of circuitous symmetry, uniform accessibility, more conspicuous exact assurance, and a diminished need of limit and computation in picture taking care of operations. Its computational power for canny vision pushes forward the photo get ready field involves the progressive units of vision. Despite its different great conditions, hexagonal lattice has so far not yet been comprehensively used as a piece of PC vision and outlines field. The crucial issue that limits the use of hexagonal picture structure is acknowledged in light of nonattendance of hardware for getting and demonstrating hexagonal-based pictures. A hexagonal lattice is a combination of seven other hexagons performs same action as a square pixel that is 3x3 as shown in Fig 1. On a hexagonal picture structure, each pixel has only six neighboring pixels which have a comparable partition to the central pixel. In this report, we will build up a hexagonal structure that is changed over from the standard square structure viably and quickly using R [1, 4].

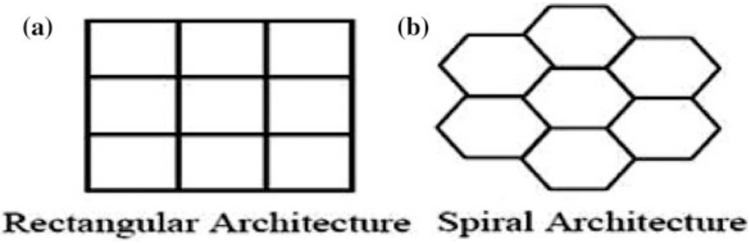


Fig. 1 Unit of vision in two image architectures

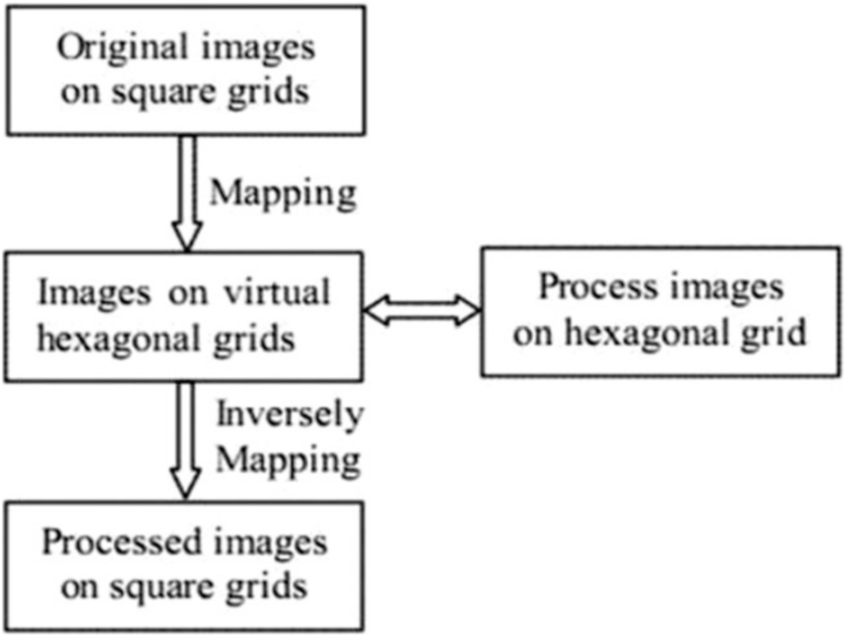

Original images on square grids
Mapping
Images on virtual hexagonal grids
Process images on hexagonal grid
Inversely Mapping
Processed images on square grids

Fig. 2 Image processing on virtual spiral architecture

3 Structure and Addressing

Hexagon framework is an option decoration plot other than the customary square lattice for testing and speaking to pictures. Hexagon pixel is profitable over square pixel in light of its higher symmetry, higher inspecting productivity, equi-distance, more prominent rakish determination, less associating impact, predictable availability. On referencing to Fig. 2.

3.1 Tessellations

To tile a plane which is regular and whose samples do not overlap with each other and with its gaps, there exist three possible tessellation schemes which are

1. Triangular Tessellation
2. Square Tessellation
3. Hexagonal Tessellation.

Any other tessellation will result in inconsistency in neighboring connectivity, gaps, overlapping among samples (Figs. 3 and 4).

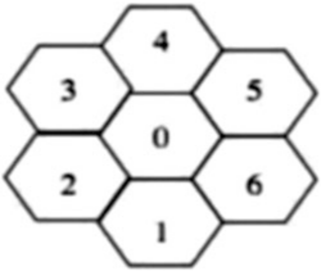


Fig. 3 A collection of seven hexagons with unique addresses

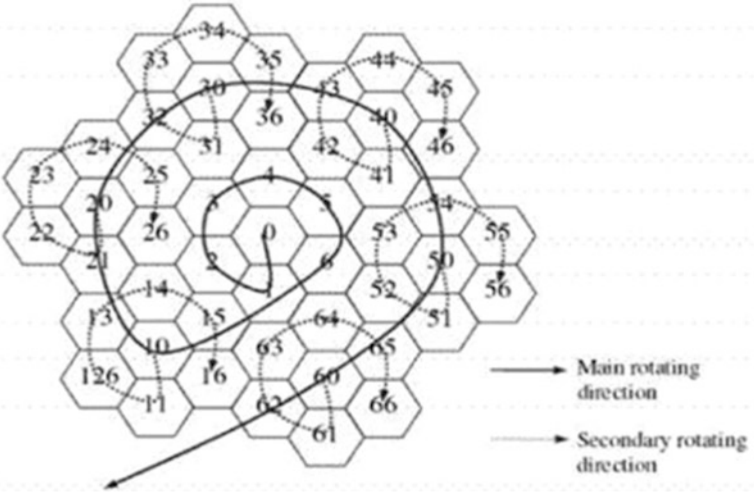


Fig. 4 Spiral architecture with spiral addressing

4 Operations On The Grid

4.1 Construction of Hexagonal Pixels

To develop hexagonal pixels, each square pixel is first isolated into 77 littler pixels, called sub-pixels [3]. To be basic, the light power for each of these sub-pixels is set to be the same as that of the pixel from which the sub-pixels are isolated. Each virtual hexagonal pixel is framed by 56 sub-pixels orchestrated as appeared in Fig. 5. To be straightforward, the light force of each developed hexagonal pixel is registered as the normal of the powers of the 56 sub-pixels framing the hexagonal pixel. A hexagonal pixel, called a hyperpel, is mimicked utilizing an arrangement of many square pixels. The R work hypel is utilized to reenact a hexagonal pixel on a square framework as per Fig. 5 [5, 6].

		X	X	X	X	X		
	X	X	X	X	X	X	X	
	X	X	X	X	X	X	X	
X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X
	X	X	X	X	X	X	X	
	X	X	X	X	X	X	X	
		X	X	X	X	X		

Fig. 5 The structure of a single hexagonal pixel

									4	4	4	4	4									
								4	4	4	4	4	4	4								
								4	4	4	4	4	4	4								
							4	4	4	4	4	4	4	4	4							
		3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5	5	5	5		
	3	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	
	3	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	
3	3	3	3	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	5	5	5	5
3	3	3	3	3	3	3	3	3	0	0	0	0	0	5	5	5	5	5	5	5	5	5
	3	3	3	3	3	3	3	0	0	0	0	0	0	0	5	5	5	5	5	5	5	
	3	3	3	3	3	3	3	0	0	0	0	0	0	0	5	5	5	5	5	5	5	
		3	3	3	3	3	0	0	0	0	0	0	0	0	0	5	5	5	5	5		
		2	2	2	2	2	0	0	0	0	0	0	0	0	0	6	6	6	6	6		
	2	2	2	2	2	2	2	0	0	0	0	0	0	0	6	6	6	6	6	6	6	
	2	2	2	2	2	2	2	0	0	0	0	0	0	0	6	6	6	6	6	6	6	
2	2	2	2	2	2	2	2	2	0	0	0	0	0	6	6	6	6	6	6	6	6	6
2	2	2	2	2	2	2	2	2	1	1	1	1	1	6	6	6	6	6	6	6	6	6
	2	2	2	2	2	2	2	1	1	1	1	1	1	1	6	6	6	6	6	6	6	
	2	2	2	2	2	2	2	1	1	1	1	1	1	1	6	6	6	6	6	6	6	
		2	2	2	2	2	1	1	1	1	1	1	1	1	1	6	6	6	6	6		
							1	1	1	1	1	1	1	1	1							
								1	1	1	1	1	1	1								
								1	1	1	1	1	1	1								
									1	1	1	1	1									

Fig. 6 A cluter of seven hexagonal pixel

Figure 5. The structure of a solitary hexagonal pixel.

Take note of that the extent of each developed pixel is greater than each square pixel. Henceforth, the quantity of hexagonal pixels is 12.5% not as much as the quantity of square pixels to cover a similar picture. As a result of this rate, the hexagonal pixels built in the route above won't lose picture determination if legitimate light forces of hexagonal pixels are allocated or added (Fig. 6).

5 Simulations

R functions sprl2rect is used to simulate a hexagonal image represented in spiral addressing scheme. Figures 7, 8 and 9 are R images of hexagonal images given as one dimensional array. The R script is given at the end. Figure 10 shows a simple conversion of an image from rectangular architecture to spiral architecture. More complex examples are not considered due to high computation power requirement. The R script is given at the end [3, 7].

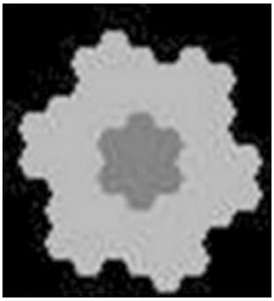
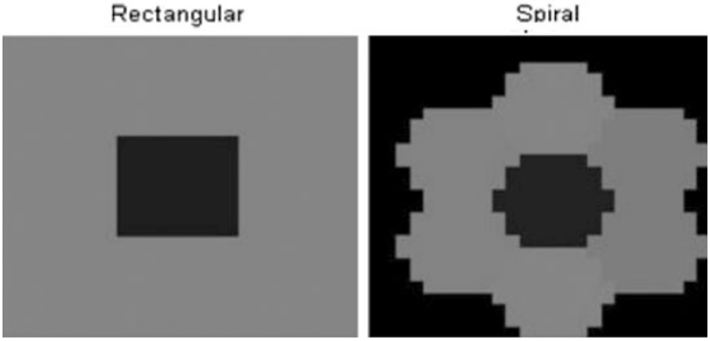


Fig. 7 A cluster of $7^2 = 49$ hexangonal pixels



Rectangular

Spiral

Fig. 8 A simple conversion from rectangular to spiral architecture

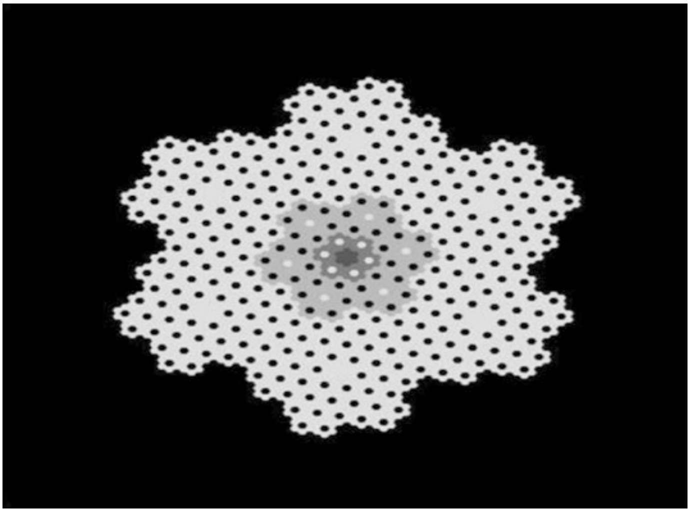


Fig. 9 A cluster of $7^3 = 243$ hexangonal pixels

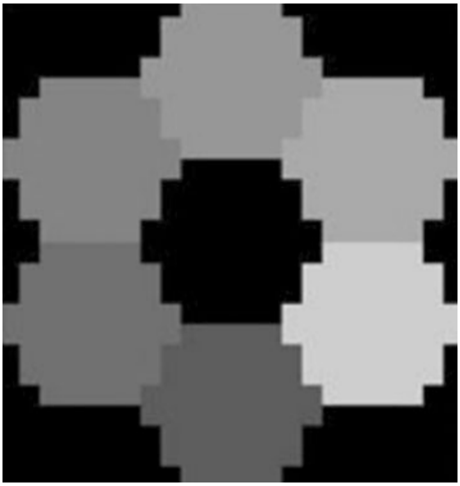


Fig. 10 A cluster of 7 hexangonal pixels

6 Results

The square pixel shows 4-overlap symmetry while hexagonal pixel displays 6-overlay symmetry. Numerous morphological operations are produced by Serra which are been generally utilized as a part of picture handling [6]. He concentrated this for various matrices and found that hexagon framework has higher symmetry.

In square pixel structure, adjacent pixels are separated by 90° which cause jaggiest in curved images. In hexagonal pixel structure, adjacent pixels are separated by 60°. Human eye have special visual preference of seeing lines which are at oblique angle because of its resemblance to that of the photoreceptors in human eye [6]. This results in better angular resolution.

7 Conclusion

In this report, a novel technique to build or copy the Spiral Architecture has been actualized in R. This built hexagonal structure does not change the picture determination and present picture mutilation. It holds the benefits of the genuine hexagonal framework, for example, higher level of symmetry, consistently associated and shut pressed shape. This structure together with the light powers can't be shown and it exists just in the PC memory amid the system of picture handling. Picture preparing in light of a hexagonal structure can be actualized utilizing this structure. The development of this new copy structure does not require complex calculation for deciding the districts of hexagonal pixels, and does not require to manufacture an extensive table put away in the PC memory to record the pixel areas. The area of every pixel can be effectively and immediately decided and figured utilizing scientific formulae.

The utilization of hexagonal pixel based pictures has increased much consideration as of late in picture engineering. From above clarifications, obviously there is change and better representation with hexagonal inspecting. Since there is absence of equipment for catching hexagonal based pictures, winding tending to gives great approach for reenacting hexagon picture handling. Hexagonal pixel structure has three overwhelming tomahawks which are 60° separated which implies little point of revolution hexagon pixels speaks to pictures superior to square pixels.

References

1. Takeda, Hiroyuki, Sina Farsiu, and Peyman Milanfar. 2007. *Kernel regression for image processing and reconstruction*. IEEE Transactions on Image Processing 2007.
2. Yeh, Jen-Chieh, Lin, Chi-Hung, and Liu, Chun-Nan. 2014. *Multi-core System Performance Prediction and Analysis at the ESL.*
3. Groenewald, A.M., E. Barnard, and E.C. Botha. 2003. *Related approaches to gradient-based thresholding*. Department of Electrical and Electronic Engineering, University of Pretoria, Pretoria, South Africa.
4. Chen, Yen-Kuang, X. Tian, Steven Ge, and M. Girkar. 2004. *Towards efficient multi-level threading of H.264 encoder on Intel hyper-threading architectures*. Proceedings of the 18th International Parallel and Distributed Processing Symposium, 2004, pp. 63.
5. Pau, Grégoire, Florian Fuchs, Oleg Sklyar, Michael Boutros, and Wolfgang Huber. 2007. *EBImage—an R package for image processing with applications to cellular phenotypes*. 8. Scheffer, R. 2007. *Uma visaoGeralsobre Threads*. RevistaCampoDigitl, vol. 2. N$_{u}$mero1. *Paginas*: 7–12.

6. Singh, Illa, and Ashish Oberoi. Comparison between Square Pixel Structure and Hexagonal Pixel Structure in Digital Image Processing. 2015. *International Journal of Computer Science Trends and Technology* (IJCST) 3(1): 176–181.
7. Huber, W., V.J. Carey, R. Gentleman, S. Anders, M. Carlson, B.S. Carvalho, H.C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K.D. Hansen, R.A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A.K. Oles, H. Pagès, A. Reyeś, P. Shannon, G.K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. 2015. *Orchestrating high-throughput genomic analysis with Bioconductor.*

EEG Based Emotion Recognition Using Wavelets and Neural Networks Classifier

S. Thejaswini, K. M. Ravi Kumar, Shyam Rupali and Vijayendra Abijith

Abstract Emotions have a vital role in the day-to-day life of human beings, the need and importance of emotion recognition systems have increased with the role of human computer interface applications. In this paper, machine learning methods are used to model a relationship using the publicly available dataset SEED (SJTU Emotion EEG Dataset) which contains EEG signals of 15 participants recorded when excited to video stimuli. The signal processing techniques in time domain and time-frequency domain (Wavelet analysis) are used to extract the desired features. The discrete wavelet transforms are used to extract frequency bands. The features such as Statistical features, Hjorth parameters, differential entropy, and the combination on symmetric electrodes (differential asymmetry DASM and rational asymmetry RASM) are extracted. Artificial neural networks and Support Vector Machine (SVM) are applied on the feature set to develop prediction models to extract the emotion information carried by the participant from emotional characteristics exhibited in different frequency bands. These models are evaluated on the dataset and emotions are classified using ANN into three different states such as positive, negative and neutral states with an accuracy of 91.2%.

Keywords Affective computing · Electroencephalogram · Emotions
SVM · ANN

1 Introduction

Emotion recognition through EEG is used largely in the field of affective computing. It involves the use of emotions in HCI (Human-computer interaction) systems giving machines a certain degree of emotional intelligence. Various methods have been proposed for these machine learning systems such as (i) the use of multimedia environments that recognise the emotions of the users, (ii) recommendation and tagging systems, (iii) gaming and films that respond to user emotions, and (iv) biofeedback devices that have been worn in the manner of headsets and might help users gain control over their emotional states [1].

S. Gurumoorthy et al., *Cognitive Science and Artificial Intelligence*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6698-6_10

Emotions can be detected by a number of ways includes facial expressions, body language, speech, brain waves or a combination of them. Since the facial, body language and speech methods are under the voluntary control of the subject under consideration and can be faked whereas the brain waves are generated for the emotions elicited and cannot be faked. Hence the use of EEG signals for emotion recognition is an ideal choice. The electroencephalogram (EEG) is a non-invasive method to record the electrical activity in the brain from the scalp. The recorded waveforms show the cortical electrical activity of the subject. EEG based emotion detection is widely popular since the machine is able to recognize human emotion states and interact based on users feelings and emotion states. The applications of emotion detection are widely used in many fields like health monitoring, entertainment, e-learning, marketing etc.

EEG signals are acquired using standard 10–20 International system. EEG signals amplitudes are measured in microvolts. These low amplitude signals are developed as a response to certain stimuli. The most well-known ways of evoking emotions are by presenting images, playing music or watching a part of a film.

In this paper, an efficient algorithm is developed to recognize positive, negative and neutral emotions using EEG signal. Publicly available dataset SEED [2] is used to detect and classify the emotion into three emotion states.

Our aim is to improve the accuracy rate of the classifier using the time domain, time–frequency domain and combination of symmetric electrodes. The SVM and ANN classifiers are used to classify the emotions in two categories: Binary and Multiclass. The accuracy of the system for two state (positive and negative) responses using a binary SVM classifier networks and three state (positive, negative and neutral) responses is compared using multiclass SVM and ANN classifiers.

This paper is organised as follows: Sect. 2 briefs out the related work for emotion recognition. In Sect. 3 the proposed methodology, extracted features and classifiers are explained in detail, followed by the performance of our proposed system in Sect. 4. The conclusion is given in Sect. 5.

2 Related Work

From the literature survey a various methods of signal processing are often used for emotion recognition with EEG signal.

Raja et al. [3], proposed a LPP-based feature extraction algorithm from EEG signals to recognize the four emotions happy, calm, sad and scared using IAPS database. The data was recorded using Emotive epoch headset using 16 channels (AF3, F7, F3, FC5, T7, CMS, P7, O1, O2, P8, DRL, T8, FC6, F4, F8, and AF4). The artifacts were removed by applying after the ERP method in MATLAB followed by band pass filtering on all EEG channels. The extracted statically features and frequency domain features of each band are classified using KNN and SVM classifiers with an average recognition rate of all the subjects was 55 and 58% respectively.

Liu et al. [4], DEAP dataset EEG data is used by pre-processing and down-sampling to 256 Hz. They also used high-pass filtering with a cut-off frequency of 2 Hz and then time domain, frequency domain, time-frequency domain and multielectrode features were extracted. They used KNN and RF to classify emotions into valance and arousal states. The performance accuracy of the classifiers for valance and arousal states is 66.1 and 65.7% respectively. By using MMR, the performance was increased to 71.23 and 69.97% respectively.

Goshvarpour et al. [5], EEG signals from eNTERFACE 06 are examined and Recurrence Quantification Analysis (RQA) is carried out, using MATLAB Toolbox. Also, the t-test ($p < 0.05$) is used to evaluate whether the difference between the extracted features is significant. They were able to recognize arousal/valance with the rates of 73.06, 62.33 and 45.32% for 2, 3, and 5 classes, respectively.

Atkinson et al. [6], proposed a novel approach that combines minimum-Redundancy Maximum-Relevance (mRMR) based feature selection tasks and kernel classifiers for emotion recognition. It trained a multi-class Support Vector Machine (SVM) classifier.

Soleymani et al. [7], made use of a face tracker to track 49 facial fiducially points. MAHNOB-HCI database was used, which is a publicly available database for multimedia implicit tagging. Regression models like linear regression, support vector regression (SVR), continuous conditional random fields (CCRF) and recurrent neural networks were used.

Bhatti et al. [8], emotions are elicited using audio tracks a dataset is created using a single channel EEG headset (Neurosky). Thirteen features from different domains are extracted, which are then classified into four different emotions (happy, sad, love and anger) using k-NN and SVM classifier.

Zhang et al. [9], developed STRNN novel deep learning framework to recognize emotion states for SEED database using differential entropy calculated for 5 bands. Their algorithm performed well for 4 kinds of expressions like anger, happiness, sadness and surprise with an accuracy rate of more than 90%.

In order to improve the recognition rate, combinations of the above feature extraction methods have been proposed.

3 Methodology

The machine learning models developed for emotion classification are generally supervised learning tasks since labels have already been assigned to the data by humans. Even so, clustering methods have additionally been employed for these purposes [10].

The SEED pre-processed data are used to extract features using wavelet transform. The wavelet transform considers both time and frequency domain components and the best combination of features are selected and applied to neural networks and SVM classifiers. The proposed methodology shown below in Fig. 1 gives the overview of the system developed to detect the emotional states.

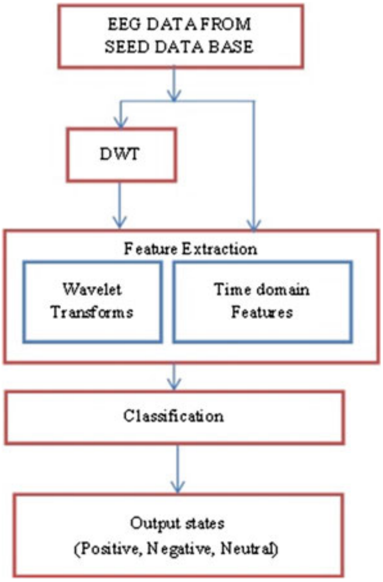


Fig. 1 Proposed methodology

3.1 SEED Database

Videos are one of the forms of visual art that can elicit emotions in subjects. Watching videos involves the sense of sight as well as hearing which acts as combined stimuli for the brain. The publicly available dataset SEED [2] is open to research communities and researchers and have been encouraged to validate their analysis on this dataset. It consists of EEG data acquired from 15 subjects, while watching emotional film videos. In order to investigate the neural signatures and stability of the patterns across sessions and individuals, each subject undergoes 15 trials in three different sessions with a gap of one week or even more between sessions. Hence are available along with the labels for each of them.

EEG signals were recorded using an ESI Neuro-Scan System at a sampling rate of 1000 Hz from the 62-channel active silver-chloride electrode cap according to the conventional international 10–20 system. The duration of each film clip is about 4 min [2]. The detailed protocol used in SEED database for data acquisition is shown in Fig. 2.

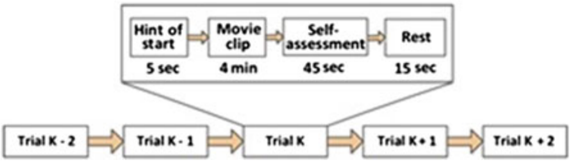


Fig. 2 Protocol used for data acquisition

The data was down-sampled to 200 Hz and a band pass frequency filter from 0 to 75 Hz was applied. EEG segments corresponding to the duration of each of the movie clips were extracted.

3.2 Feature Extraction and Selection

From the available pre-processed data, the features in time domain, time-frequency domain and a few multiple electrode features are extracted such as DASM and RASM.

From the literature survey, it is found that Wavelet transform provides better resolution in the time and frequency domain as compared to FFT and STFT. Hence, DWT is used to extract different frequency bands as shown in Fig. 3. In the time-frequency domain Discrete Wavelet Transform with 4 levels of decomposition and wavelet function “db8” is used to obtain signals in the desired frequency bands delta (1–3 Hz), theta (4–7 Hz), alpha (8–13 Hz), beta (14–30 Hz), gamma (31–50 Hz) bands as shown in Fig. 3.

Features such as energy spectrum, differential entropy and its combination on symmetrical electrodes [11] are extracted in each of the five frequency bands. It is seen that Differential Entropy and its combination on symmetrical electrodes yield better performance than ES features.

Average energy of each band i.e. alpha (e_a), beta (e_b), gamma (e_g), delta (e_d) and theta (e_t) are calculated. Energy Spectrum is the average energy taken in the five bands. Differential Entropy is taken as the logarithm of the energy spectrum for the respective bands. It can be defined as,

$$\begin{aligned} G(x) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}\right) dx \\ &= 1/2 \log(2\pi e\sigma^2) \end{aligned} \tag{1}$$

The combination of 27 pairs of symmetrical electrode channels as shown in Table 1 is used to calculate Differential Asymmetry (DASM) and Rational Asymmetry (RASM).

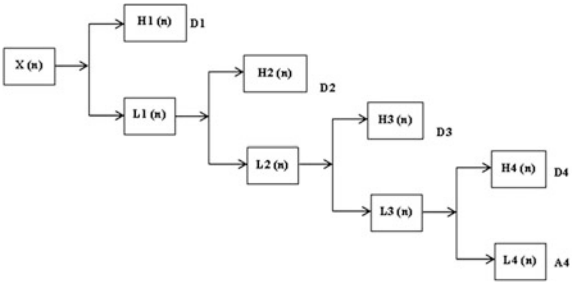


Fig. 3 The corresponding frequency bands for alpha, beta, gamma, theta and delta

Table 1 Twenty seven pairs of asymmetry electrodes

Pair No.	1	2	3	4	5	6	7
Left	FP1	F7	F3	FT7	FC3	T7	P7
Right	FP2	F8	F4	FT8	FC4	T8	P8
Pair No.	8	9	10	11	12	13	14
Left	C3	TP7	CP3	P3	01	AF3	F5
Right	C4	TP8	CP4	P4	02	P04	F6
Pair No.	15	16	17	18	19	20	21
Left	F7	FC5	FC1	C5	C1	CP5	CP1
Right	F8	FC6	FC2	C6	C2	CP6	C22
Pair No.	22	23	24	25	26	27	
Left	P5	P1	P07	P05	P03	CB1	
Right	P6	P2	PO8	P06	P04	CB2	

$$\mathrm{DASM} = \mathrm{G}(\mathrm{X}_\mathrm{i}^{\mathrm{left}}) - \mathrm{G}(\mathrm{X}_\mathrm{i}^{\mathrm{right}}) \tag{2}$$

$$\mathrm{RASM} = \mathrm{G}\left(\mathrm{X}_\mathrm{i}^{\mathrm{left}}\right)/\mathrm{G}\left(\mathrm{X}_\mathrm{i}^{\mathrm{right}}\right) \tag{3}$$

For additional features, the time domain Statistical features such as mean, standard deviation, first difference, normalized first difference, second difference, normalized second difference, kurtosis, Hjorth parameters-complexity and mobility are also considered.

The features are defined as follows.

$$\text{Mean}, \mu = \frac{1}{T}\sum_{t=1}^{T} s(t) \tag{4}$$

Standard deviation Measures the deviation of electrodes potential from its mean value over different emotional EEG signals.

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(s(t) - \mu\right)^2} \tag{5}$$

First difference is given by,

$$\delta = \frac{1}{T-1}\sum_{t=1}^{T-1} |s(t+1) - s(t)| \tag{6}$$

Normalized first difference is the first difference by standard deviation and is calculated as

$$\delta' = \frac{\delta}{\sigma} \tag{7}$$

Second difference is given by,

$$\gamma = \frac{1}{T-2}\sum_{t=1}^{T-2} |s(t+2) - s(t)| \tag{8}$$

Normalized second difference is the second difference by standard deviation and is calculated as

$$\gamma' = \frac{\gamma}{\sigma} \tag{9}$$

Kurtosis is the sharpness of the peak of a frequency-distribution curve and is calculated as,

$$k = E\left(\frac{t-\mu}{\sigma}\right)^4 \tag{10}$$

The Hjorth parameters are used to calculate features such as complexity and mobility. The activity feature is not used as it is not useful in emotion detection [2]. Complexity and mobility are simple statistical features computed using variance and mean of each channel.

$$\text{MOBILITY} = \sqrt{\frac{variance(\dot{s}(t))}{variance(s(t)}} \tag{11}$$

$$\text{COMPLEXITY} = \sqrt{\frac{mean(s(\dot{t}))}{mean(s(t))}} \tag{12}$$

The above-mentioned features have been computed for all the 62 channels [2] and averages of all the channels are taken.

3.3 Classification

The selected feature vector of size 291 is applied to two kinds of classifiers-Multi-Class Support Vector Machine (SVM) and Artificial Neural Networks (ANN) to classify the features into three states. The SVM classifier is a binary classifier and hence to classify more than two states we make use of multiclass SVM using dendogram function. It generates a dendogram plot of the hierarchical binary cluster tree. A dendogram is made up of many U-shaped lines that connect

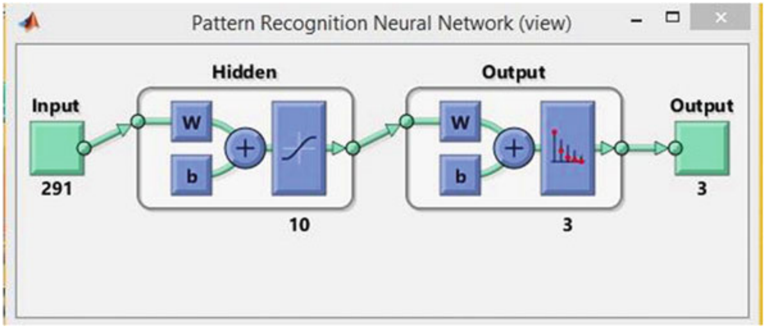


Fig. 4 Neural network architecture

data points in a hierarchical tree. The height of each of the U shaped lines represents the distance between the two data points being connected. This tree structure approach SVM classifier gives us an accuracy of 40.4%.

The ANN classifiers are an ideal method of classification for multiple classes using neural pattern recognition. In this paper ANN classifier is implemented using neural networks toolbox to perform classification. To implement ANN classifier, “nnstart” a neural network start function in MATLAB is used to obtain architecture as shown in Fig. 4. The algorithm performs nonlinear classification and clustering on the extracted features using multilayer perceptron. From the features extracted we use 60% of the data for training the model and the remaining is used for testing and validation. A 5-fold cross validation is used for better performance. An accuracy of up to 91.2% is obtained.

The same approach is used to classify the features into two states (positive and negative) by removing the neutral component and the results obtained for both classifiers are compared with the results of multi class classifier (3 states).

The average energy which is calculated for each band is applied for ANN classifier. It is seen that the energy in alpha and gamma are more predominant than other bands.

3.4 Confusion Matrix

The confusion matrix is developed by comparing the predicted classes and target classes. The accuracy of each of the emotional states and the total accuracy of the model can be determined using the confusion matrix. The diagonal elements represent the predicted values corresponding to each emotional state to their respective target classes. For good results, the diagonal elements value should be higher than surrounding values. As shown in Figs. 5 and 6, the bottom-right corner element represents the overall accuracy of the model.

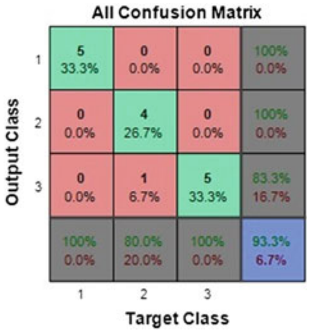


Fig. 5 Confusion matrix for three states using neural network for one subject

4 Results

In this paper, 15 subjects performing three sessions from SEED Database are used. The 291 features are calculated for each subject and supplied for the classifier. The results of the various classification processes are calculated by constructing a confusion matrix for each of the classifiers. The confusion matrix gives us the accuracy of the model.

Figures 5, 6, and 7 shows the confusion matrix for each of the classifiers. Figure 5 shows an accuracy of 93.7% for the three emotional states neural network classifier for one subject. Fivefold cross validation is done on training and testing data to obtain the most accurate outputs. The accuracy obtained for each of the classifiers is shown in the form of a bar graph in Fig. 8.

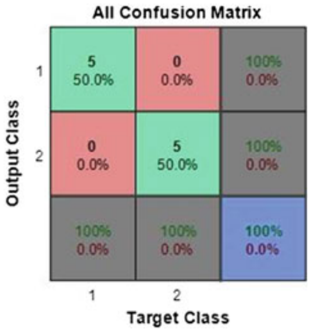


Fig. 6 Confusion matrix for two states using neural network

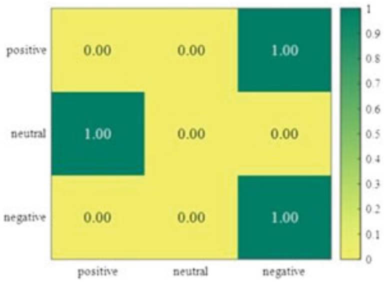


Fig. 7 Confusion matrix for three states using multiclass SVM classifier

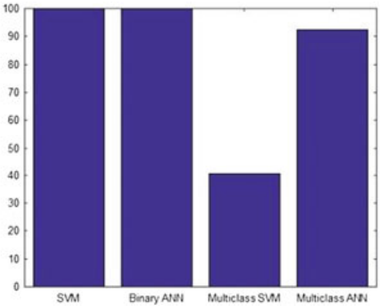


Fig. 8 Accuracy of each classifier for two and three states

Table 2 Classifier performance of 10 subjects

Subject	Positive	Neutral	Negative	Overall accuracy
1	33.3	26.7	33.3	93.3
2	26.7	26.7	33.3	86.7
3	33.3	26.7	33.3	93.3
4	33.3	33.3	26.7	93.3
5	33.3	20	33.3	86.6
6	33.34	33.33	33.33	100
7	26.7	26.7	33.33	86.73
8	33.3	26.7	33.33	93.33
9	33.3	26.7	33.33	93.33
10	33.3	33.3	20	86.6

Table 2 and Fig. 9, shows the classifier performance for ten subjects and its distribution among the positive, neutral and negative emotional states using ANN classifier. A mean accuracy of 91.2% is obtained for all the subjects. It is seen that the neural networks classifier provides much better accuracy compared to the

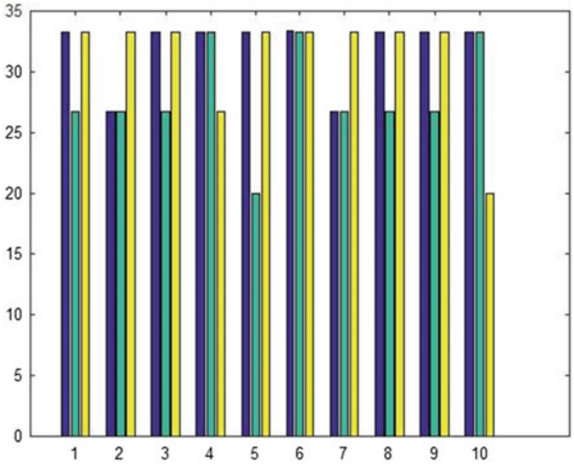


Fig. 9 Accuracies for ten subjects in each of the emotional states

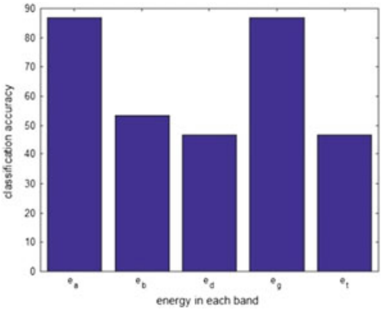


Fig. 10 Average energy of each band

multiclass SVM for three emotional states. However, for binary classification both classifiers provide the best results of performance.

The performance accuracy of average energy for each band is shown in Fig. 10. It is shown that emotion states are predominant in alpha and gamma frequency bands compared to other bands.

5 Conclusion

In this paper, an efficient algorithm to detect emotional states using the SEED dataset is developed. It combines the features obtained from wavelet transform and multi-electrode features and is applied to SVM and ANN classifiers to detect emotion states. The performance accuracy for multiclass classifiers using SVM and ANN are 40.4 and 91.2% respectively. The performance accuracy for binary classifiers using SVM and ANN are above 94%. It is observed that the classification accuracy of ANN is better compared to SVM for SEED Dataset. The alpha and gamma bands are predominant for emotion detection.

This model works best for the downloaded SEED dataset and in the future the model can be tested on real time EEG signals and results can be compared. It can further be improved by adding more features and combining the results of different classification methods to provide best results. Different combinations of features and classifier can be tested and based on the accuracies the most efficient combination can be applied.

References

1. Liu, Jingxin, Hongying Meng, Asoke Nandi, and Maozhen Li. 2016. Emotion Detection from EEG Recordings. In *12th International IEEE Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).*
2. Zheng, Wei-Long, and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development (IEEE TAMD)* 7 (3): 162–175.

3. Mehmood, Raja Majid, and Hyo Jong Lee. 2016. A Novel Feature Extraction Method Based On Late Positive Potential For Emotion Recognition In Human Brain Signal Patterns. *Computers and Electrical Engineering Journal* 53: 444–457.
4. Liu, Jingxin and Hongying Meng. 2016. Emotion Detection from EEG Recordings. In *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*.
5. Goshvarpour, Ateke, Ataollah Abbasi, and Atefeh Goshvarpour. 2016. Recurrence Quantification Analysis and Neural Networks for Emotional EEG Classification. *Applied Medical Informatics* 38(1): 13–24.
6. Atkinsona, John, and Daniel Campos. 2016. Improving BCI-Based Emotion Recognition By Combining EEG Feature Selection And Kernel Classifiers. *Expert Systems with Applications Journal* 43: 35–41.
7. Soleymani, Mohammad (Member, IEEE), Sadjad Asghari-Esfeden (Student Member, IEEE), Yun Fu (Senior Member, IEEE), and Maja Pantic (Fellow, IEEE). 2016–2017. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions On Affective Computing* 7(1): 17–28.
8. Bhatti, Adnan Mehmood, Muhammad Majid, and Syed Muhammad Anwar . 2016. Human Emotion Recognition And Analysis In Response To Audio Music Using Brain Signals. *Computers in Human Behavior Journal* 65: 267–275.
9. Zhang, Tong, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. 2014. Spatial-Temporal Recurrent Neural Network for Emotion Recognition. *Journal of Latex Class Files* 13(9): 1–8.
10. Valenzi, S., T. Islam, P. Jurica, and A. Cichocki. 2014. Individual Classification of Emotions Using EEG. *Journal of Biomedical Science and Engineering* 7: 604. (BMC Bioinformatics BMC series).
11. Duan, Ruo-Nan, Zhu, Jia-Yi, and Bao-Liang Lu. 2013. Differential Entropy Feature for EEG-based Emotion Classification. In Proceedings of the 6th International IEEE EMBS Conference on Neural Engineering (NER), 81–84.