SPECIAL ISSUE

THE IIOAB JOURNAL

VOLUME 6 : NO 4 : DECEMBER 2015 : ISSN 0976-3104

Institute of Integrative Omics and
Applied Biotechnology Journal

Dear Readers,

It is my utmost pleasure to extend a heartfelt welcome to each of you as we embark on a thrilling journey into the realm of our scientific journal dedicated to the diverse applications of soft computational approaches in bioinformatics.

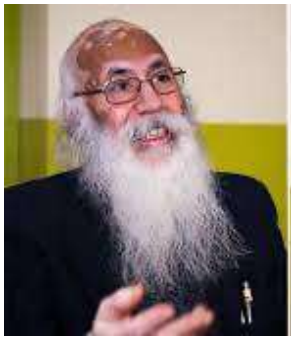
As the Editor of this IIOAB Journal issue, under the topic of Embracing the Boundless Horizons of Soft Computational Approaches in Bioinformatics, I am thrilled to witness the convergence of innovative methodologies and cutting-edge technologies in the realm of bioinformatics. Our journal serves as a conduit for the dissemination of pioneering research, transformative algorithms, and novel applications that harness the power of soft computational techniques in unraveling the mysteries of biological systems.

Your expertise, dedication, and scholarly contributions play an integral role in shaping the future landscape of bioinformatics. Through your research, you illuminate the path toward understanding complex biological phenomena, deciphering genetic codes, and exploring the intricate networks of biomolecular interactions.

The interdisciplinary nature of this field not only drives innovation but also holds immense promise for revolutionizing healthcare, agriculture, pharmaceuticals, and beyond. Your endeavors in this field are invaluable, propelling us towards groundbreaking discoveries and applications with profound implications for society.

I encourage each of you to share your groundbreaking insights, submit your pioneering research, and engage in vibrant discussions within our journal's pages. Let us collaboratively nurture an environment where ideas flourish, collaborations thrive, and knowledge propels us forward.

Thank you for your unwavering commitment to advancing the frontiers of bioinformatics through the integration of soft computational approaches. I eagerly anticipate the wealth of knowledge and transformative insights that will emanate from your contributions.



Warm regards,
Guest Editor:
Prof. S. Arun Kumar and
Prof. T. Arun Kumar

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Sangaiah and Thangavelu

THE IIOAB JOURNAL

EDITORIAL OPEN ACCESS

APPLICATIONS OF SOFT COMPUTIONAL APPROACHES IN BIOINFORMATICS

Arun Kumar Sangaiah* and Arun Kumar Thangavelu

School of Computing Science and Engineering, VIT University, Vellore, INDIA

In today's world, the use of soft computational approaches has been increased to a variety of bioinformatics application. Bioinformatics approaches – from soft computational methods such as fuzzy system, neural networks, evolutionary algorithms, statistical model algorithms are currently attractive topic in wide variety of bioinformatics research areas. Usually, biological systems and its objects have intrinsically fuzzy as their features and behaviors consist of randomness or uncertainty in nature. Subsequently, only limited computational intelligence approaches are available to handle bio-informatics problems. In order to address the research gap, this special issue identifies the novel work that focuses on the recent advances in bioinformatics, bio-inspired methods, and contribution of computational intelligence approaches to solve biological problems. In the following section, we start by providing the readers of this special issue a brief overview of these research papers.

Published on: 8th –Aug-2015

KEY WORDS

Bio informatics; Soft Computing, Optimization Methodology

***Corresponding author: Email:** arunkumarsangaiah@gmail.com, **Tel.:** +91-9842935634

EDITORIAL: APPLICATIONS OF SOFT COMPUTATIONAL APPROACHES IN BIOINFORMATICS (ASCB)

The first paper in this special issue, ''Diagnosis of diabetic retinopathy from fundus image using fuzzy c-means clustering algorithm'', Mohan Jagannath and Kanagasabai Adalarasu presents novel spatially weighted Fuzzy C-Means clustering algorithm for vessel detection in ocular fundus images.

The second paper by Pradeep Reedy et al. intend presenting a novel on-demand multicast routing protocol generally known as Cross-Layered Ant Colony Optimization Multicast Routing Protocol (CLAMR). The result of this selected protocol has been improved and enhanced version of the already existing on-demand multicast routing protocol (ODMRP).

The next paper by Sasi Kumar and Tripathy describes the improvement in automated analysis for human brain signal processing to finding alzheimer's disease using intelligent techniques. Also they presented an improved fuzzy firefly algorithm, which would enhance the classification of the brain signal efficiently with minimum iteration.

The predictive analytical approach towards improving the crop growth yield using fuzzy cognitive maps is the focus of the next paper. Gunasundari Anantharaj et al. describe an important case study to multiple rethegorical factors from the basis of improved crop yield. Moreover, this paper proposes Fuzzy Cognitive Map approach to help in decisive making and suggesting an optimized solution of improving crop yield as well compared with existing approaches such as Genetic Algorithm and Artificial Neural Networks (ANN).

The subsequent two papers deal with a recent wireless sensor network research issues. The paper by Roselin Jones et al. focuses an novel redeployment strategies for balanced coverage in Wireless Sensor Network (WSN). Moreover, the proposed redeployment strategies significantly result in balanced point coverage, which in turn, improves the overall lifetime of the network. Likewise, next paper by Rajesh et al. proposing Chaos Theory based Data Aggregation (CTAg) based approach to reduce redundant information and number of packet transmission between aggregator and sink node in WSN.

The next paper by Maheswari Arumugam and Arun Kumar Sangaiah gives a brief review on various methods applicable for feature extraction and classification of bio signals for cardiac arrhythmia. Similarly, the paper

THE IIOAB JOURNAL
ISSN: 0976-3104

complemented by Challa Anusha et al. determines the evaluation criteria of social network usage from the perspective of student academic performance, authors have integrated multi-criteria decision making (MCDM) approaches: (a) Quality Function Deployment (QFD), (b) Decision Making Trial and Evaluation Laboratory (DEMATEL) and (c) Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) under fuzzy environment is proposed.

The next paper by Suresh Thangakrishnan and Kadarkaraiyandi Ramar presents biometrics for identifying a person in the world by the physiological features in the human body. Various biometric techniques include features in the human body like the facial, iris, gestures, fingerprint, gene, key stroke biometrics, etc. Subsequently, next paper by Prabhavathy Paneer, BalakrishnaTripathy tunes the existing rough c-means and fuzzy c-means and integrates them into a tuned hybrid soft clustering algorithm termed as the tuned rough-fuzzy c-means algorithm. In addition, the proposed algorithm performance is compared with the existing rough c means, fuzzy c-means, and rough fuzzy c –means approaches.

Dynamic fuzzy clustering algorithm on cancer data and identify the changes in cluster structures for every incoming new data set with respect to previous data is the focus of a next paper. Chatti Subbalakshmi et al. elaborates fuzzy c-means clustering or soft clustering algorithm on cancer data to get initial clustering solution, after that for each cycle added into the new cases for apply conditions to identify the changes.

The next paper by Ramsingh, and Bhuvaneswari presents hybridized approach for biomarker discovery using Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). The proposed approach is found to be better for identifying optimal gene is verified based on biological validation. Further, next paper by Nagaraju and Tripathy presents the rough equality and rough equivalence in the context of covering based optimistic multi-granular rough sets and establish their properties in the general form as well as in the replacement form.

Final paper by Suja et al. sums up the Extreme Learning Machine (ELM) has been simple and effective algorithm for single-hidden layer feed forward neural networks (SLFN) which automatically classifies the EMG signal into healthy, myopathic or neuropathic.

EDITORS' BACKGROUND

Arun Kumar Sangaiah *obtained his Doctor of Philosophy (Ph.D.) in Computer Science and Engineering from VIT University, Vellore, Tamil Nadu, India. He is presently working as a Associate Professor in School of Computing Science and Engineering, VIT University, India. His experience and areas of interest focus on Software Engineering, Soft Computing, Wireless Networks, Bio-Informatics, and Embedded Systems. He has author of more than 100 publications in different journals and conference of National and International repute. Dr.Arun Kumar is member of international advisory board of IJIIT (IGI publisher) and editorial board member of IJIES, Japan. Also, he has served a special guest editor of various international journals (Inderscience, Hindawi, IGI publishers). He is active member for Compute Society of India. He has guided many research students and post-graduate students in the field of software engineering, communication networks, ad hoc networks, database, and soft computing techniques.*

Arunkumar Thangavelu *is associated with Vellore Institute of Technology as Senior Professor in School of Computing Sciences. His area of research interest focus on soft computing approaches, adhoc high-performance networking (in MANET/ VANET/ PANET), data mining / analysis, Big Data Analysis, ambient intelligence, aspect based network management, context aware middleware and Internet of Things (IoT). He has published more than 90 works in multiple international conferences and journals. He serves as chair and program committee member in organizing numerous international / national conferences, in which he had delivered key note lecturers. He had authored research books and review research works from International journals such as IEEE, Science Direct and Wiley publications. He is also fond of giving lecturers and helping the research community to achieve and potential towards active participation of research.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Jagannath and Adalarasu

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

DIAGNOSIS OF DIABETIC RETINOPATHY FROM FUNDUS IMAGE USING FUZZY C-MEANS CLUSTERING ALGORITHM

M. Jagannath[1] and K. Adalarasu[2*]

[1]*School of Electronics Engineering, VIT University Chennai, Tamil Nadu, INDIA*
[2]*Department of ECE, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, INDIA*

ABSTRACT

Diabetic retinopathy is a chronic disorder which is considered as a major source of vision loss in patients suffering from diabetes. It is characterized by the destructive of blood vessels that nourish the retina. However, early detection of such disorder through regular diagnosis, vision loss can be avoided. In order to reduce the diagnosis cost and enhance the automated analysis, modern image processing tools are used to detect the existence of disorders in the retinal images acquired during the initial process of screenings. This paper presents a methodology for the extraction of exudates within blood vessels from fundus images using Fuzzy c-Means (FCM) clustering algorithm. Matched filter was applied for vessel extraction with the help of adaptive histogram equalization, thresholding method and segmenting method, which incorporates spatial neighborhood information into the FCM clustering algorithm. A standard diabetic retinopathy database was used in this study to test the proposed algorithm. This methodology showed improved sensitivity and accuracy of the segmented result. The proposed method seems to be promising as it can also detect the very small areas of exudates. Such an image processing technique can reduce the work of ophthalmologists and help in patient screening, treatment and clinical studies.

Received on: 18th-March-2015
Revised on: 15th-May-2015
Accepted on: 23rd-May-2015
Published on: 8th-Aug-2015

KEY WORDS

Diabetic retinopathy; Fuzzy c-Means; digital fundus image; exudates

***Corresponding author:** **Email:** adalbiotech@gmail.com, **Tel.:** +91-9965718013; **Fax:** +91-451-2554249

INTRODUCTION

Diabetic retinopathy is an emerging cause of vision loss in both developed and developing countries. The World Health Organization (WHO) has projected the number of adults with diabetes in the world would increase exceptionally: from 135million in 1995 to 300 million in 2025. In India, this condition is expected to be the greater; (i.e.) nearly 195% from 18 million in 1995 to 54 million in 2025 [1-3].

Diabetic retinopathy is a chronic disorder which is considered as a major source in patients suffering from diabetes [4]. It is characterized by the destructive of blood vessels that nourish the retina. It is of two categories: category 1 primarily initiated by autoimmune pancreatic b-cell damage and described by absolute insulin deficiency, and category 2 described by insulin conflict and relative insulin deficiency [5]. For example, in the USA; it was estimated that nearly 7% of the US population achieved the diagnostic criteria for diabetes [6]. In Saudi Arabia the occurrence of diabetes mellitus was estimated at around 31% and it grows with age [7]. It is the general cause of vision loss between the age-group of 20 and 65 years. Recently an extensive study had been performed in various aspects of diabetic retinopathy.

The most effective treatment is detecting the condition in premature stage through regular screenings [8, 9]. During the screenings, color fundus images are obtained using fundus camera. However, as a result large number of fundus images are produced which requires physical investigation and diagnosis. Moreover, medical experts like ophthalmologists need to spend their time and energy to analysis these images. It would be more affordable if the initial process of analyzing the images can be done automatic so that merely the abnormal fundus images need to be diagnosed by the ophthalmologists [8, 10].

On the other hand, diabetic retinopathy which consequences from long-term diabetes mellitus, it is a common disease that clues to choroidal neo-vascularization [2]. The choroidal neo-vascularization is a significant stage that leads to vision loss. It decreases the amount of blood supplied to the retina. One of the treatment strategies is that

THE IIOAB JOURNAL ISSN: 0976-3104

the affected areas of the retina are photocoagulated with the help of lasers. To achieve satisfactory results, the expert has to identify the choroidal neo-vascularization and cauterize it completely. Optic disc and blood vessels should be carefully avoided while radiating the area of acute vision.

Retina is a thin clear structure including of several layers. The cells within the retina includes three major components: (1) neuronal component which contribute the retina its visual function by converting light to electrical signals; (2) Glial components are the supporting column of the retina; and (3) Vascular components which delivers the inner retina while the outer retinal is being delivered by diffusion from choroidal circulation [5]. Diabetes will produce its result on both neuronal and vascular components of the retina.

Several issues were found to influence diabetic retinopathy including chronic characteristics of the disease, age, pregnancy, blood pressure, hyper-viscosity, kidney failure and anemia. Hyper-viscosity of the blood [11] influences the diabetic retinopathy.

More accurate and reliable solution through image processing and artificial intelligence tools included Genetic Algorithm (GA) and Artificial Neural Network (ANN) [12, 13]. Currently numerous clustering algorithms have been developed for image segmentation. Artificial intelligence and fuzzy based method such as Adaptive Neuro-Fuzzy Inference System (ANFIS), K-Means, Fuzzy c-Means (FCM) [14-17] are considered to be effective in image segmentation. Khalida et al. [18] suggested FCM as a clustering technique for segmentation of color image based on the elementary region developing method and used membership grades of pixels to categorize pixels into suitable segments.

The research framework of this paper is to perform FCM algorithm on fundus image to extract the exudates within blood vessels. The organization of the rest of this paper is as follows: Section 2 presents our methods, including the overview of the methodology, detection, filtering, template matching and descriptions of clustering algorithm's structure. Section 3 discusses the results and discussions. Finally, the paper ends with a short summary in Section 4.

MATERIALS AND METHODS

The proposed research framework starts with pre-processing which is mainly to enhance the image quality for the stages that follows. The image pixel values were permanently altered and enhanced data was used for further analysis. Pre-processing suppressed the undesired information and enhances the desired features. Pre-processing involved brightness correction, detection of edges, histogram equalization and so on.

Detection

Diabetic patients require routine eye examinations so that interrelated eye problems can be identified and treated efficiently. Most diabetic patients are normally observed by an endocrinologist who works closely with the ophthalmologist.

Filtering

A fundus camera delivers fundus image in digital form which can be effectually used for the computer based automated detection of diabetic retinopathy. The performance on one of the normal fundus image is shown in **Figure– 1**.

One of the most important components of our system is the filters, which extracts the ideal data for the diagnosis. Matched filter was used for vessel segmentation that acquires more computational time than other edge detector for the same purpose [1, 19].

Unfortunately due to uneven illumination it may appear darker because macula centered fundus images are often captured on both right and left eye [9]. To rectify this problem we have to use illumination equalization to normalize the luminosity across the image. Adaptive histogram equalization enhances the blood vessels in vessel segmentation [21]. Since blood vessels usually have lower reflectance when compared with the background, these types of contrast enhancement improve the vessel segmentation.

Template matching

The normal and healthy fundus images were taken and kept as standard to separate the abnormalities in the test image. This standard image acted as the template. Both the standard image and test images were transformed from RGB to gray levels and later by point processing (pixel by pixel) both the images were assessed.

During assessment, the further objects present in the test image got separated and made clearly visible in the result. While comparing, if the test image matches with a normal one, then it gets cancelled as there is no difference in pixel values between

THE IIOAB JOURNAL
ISSN: 0976-3104

the two images. The basic requirement of the proposed method is that, the references of the normal fundus image and the test images must be taken in the same orientation with same lighting, background and so on [21, 22].

Proposed methodology

The proposed methodology is composed of four steps. Since blood vessels usually have lower reflectance compared with the background, a green component of fundus image will be separated. It was preprocessed to reduce the noise. Then we applied the matched filter to enhance blood vessels. The spatially weighted Fuzzy c-Means (FCM) clustering algorithm is used to distinguish between vessel segments and the background of the matched filter response (MFR) image [3]. A label filtering technique is used to remove the misclassified pixels.

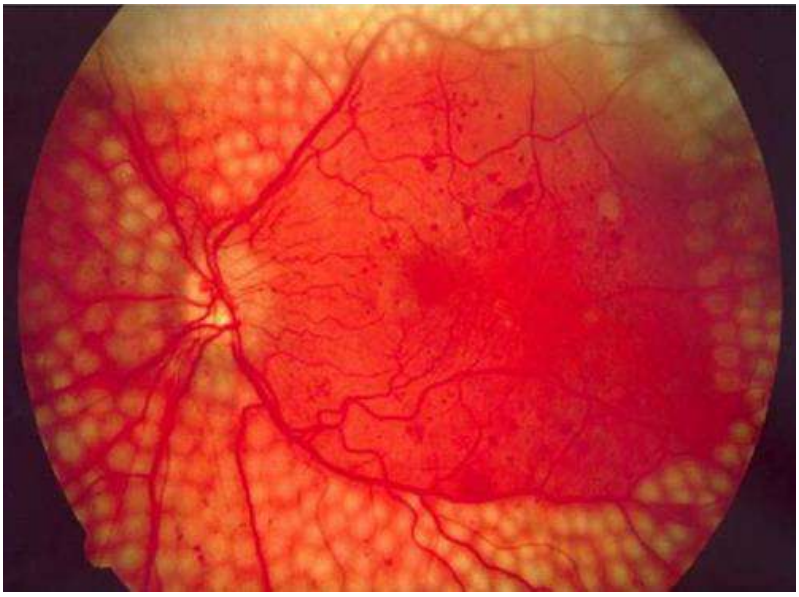


Fig: 1. Vessel segmentation of original fundus image.

In order to extract the enhanced segments in the matched filter response images, an effective thresholding scheme is necessary to avoid complicated relationships or overlap between foreground and background. Hence thresholding based on spatially weighted FCM clustering algorithm was implemented.

The spatially weighted FCM clustering algorithm was formulated by combining the information of spatial neighboring into the FCM algorithm [3]. The weight in the algorithm was modified by bearing in mind the neighborhood influence on the central pixel to improve the performance of image thresholding.

The computation time of the proposed method is very fast as compared to the conventional techniques since the gray level histogram of image was used as an alternative of the whole data of image to compute the parameter for the FCM algorithm [22]. Due to the concern of the neighborhood information, the proposed method seemed to be noise resistant.

Clustering algorithm

In this paper, a prior knowledge about spectral information for certain land cover classes is preferred, in order to classify image in fuzzy logic manner. Steps to implement the algorithm are given below.

Step 1: Assign the number of clusters.
Step 2: Randomly allocate input vector to the cluster. Create partition.
Step 3: Compute a cluster as center, the mean of each vector component of all vectors assigned to that cluster. Repeat for all clusters.

Step 4: Estimate the distance between each and every center and input vector.
Step 5: Update partition by assigning each input vector to its nearest cluster center.
Step 6: Stop if center do not move any more otherwise loop to Step 4.

The proposed method uses an iterative clustering method that yields a best c partition, Since FCM algorithm is iterative and time consuming, the gray level histogram of image was used to the algorithm [23].

THE IIOAB JOURNAL
ISSN: 0976-3104

RESULTS AND DISCUSSIONS

In order to evaluate the performance, we compare our simulation results with the state-of-the-art results obtained from piecewise threshold probing, local entropy thresholding, well known Otsu thresholding and hand-labeled ground truth segmentations [1, 24]. The vessel detection from Otsu thresholding is depicted in **Figure– 2**.

In this section, the results obtained from our proposed algorithm are presented. Results are compared with the well-known Otsu thresholding. Moreover, since the algorithm behavior should be image size independent, a statistical study of the measures variation was performed out. The study was based on the processing of 20 normal retinal images of each class and on three sets of measurements: one for the original images, one associated to halved area versions of the original images, and another for doubled area versions of the original images. The results of the present study revealed that the proposed algorithm is robust to differences in resolution of the image.

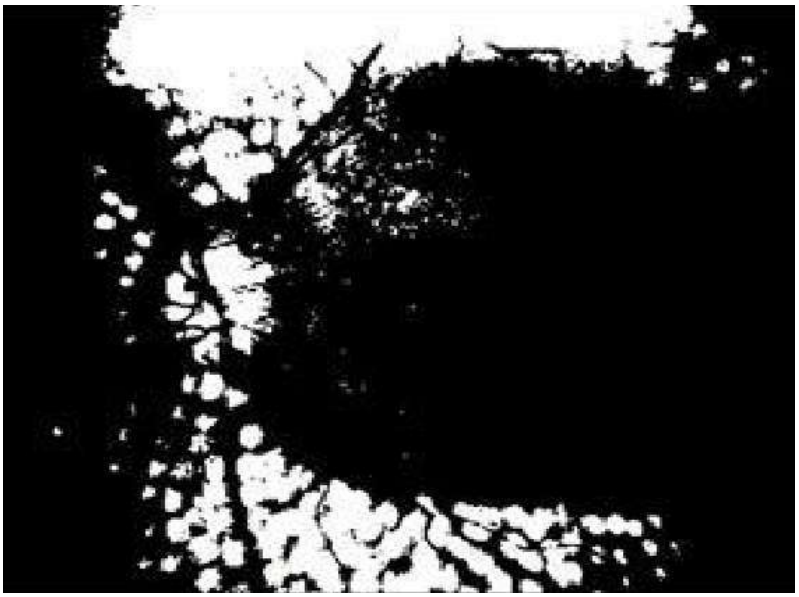


Fig: 2. Vessel detection from Otsu thresholding.

The proposed method preserves the computational ease and also can achieve accurate segmentation results in the case of normal fundus images and abnormal images with obscure blood vessel appearance. Blood vessel width is deliberated by interpolation technique and award elevated disparity among normal and abnormal images.

The algorithm was implemented in MATLAB version 2009b and is run on a 1.7GHz Core 2 Duo personal computer with a memory of 8 GB. Test fundus images are taken from the stare database. Normal retinal images and affected images are used for the experiment. The computational time for the complete procedure of the proposed algorithm just takes about one minute for each fundus image. Among the 30 images of standard diabetic retinopathy Database, 20 normal images with no pathology (normal) and 10 abnormal images including pathology that obscures or even confuses the blood vessel appearance in varying positions of the image (abnormal) are taken for analysis.

Detection of vessel in fundus images produces unconnected parallel edges to let extraction of blood vessels using spatial weighted FCM **[Figure– 3]**. The edge detection techniques produce better results only when the edges are distinct and sharp. The proposed methodology performs better by segmenting even the smaller blood vessels. The proposed methodology segments the blood vessels very well from the background and extracts the exudates to classify fundus images. The proposed method has achieved 94.5% accuracy in identifying all the retinal images with exudates, and 86% accuracy in classifying normal retinal images as normal. But it shows that the major obstacle of this approach is the presence of lesions in the abnormal Images. From the above experimentation it is found that Global Otsu thresholding algorithm for vessels segmentation is best suited because it is compatible with template matching algorithm.

THE IIOAB JOURNAL
ISSN: 0976-3104

Early diagnosis of diabetic retinopathy was suggested based on decision support system by Kahai et al. [25]. Normal and prolific diabetic retinopathy phases were automatically classified with the help of dimensions of the RGB components of the blood vessels coupled with a neural network technique. Nayak et al. [26] have investigated exudates and blood vessel area along with texture parameters together with neural network to classify fundus images into normal and diabetic retinopathy. Recently, Acharya et al. [27] used support vector machine (SVM) classifier to categorize the fundus image into normal and prolific diabetic retinopathy phases. They have established an average accuracy of 82%. Larsen et al. [28] have developed an automated system, which identified around 90% of patients with diabetic retinopathy and around 82% of patients without diabetic retinopathy, when implemented in a screening population involving of patients with untreated diabetic retinopathy.

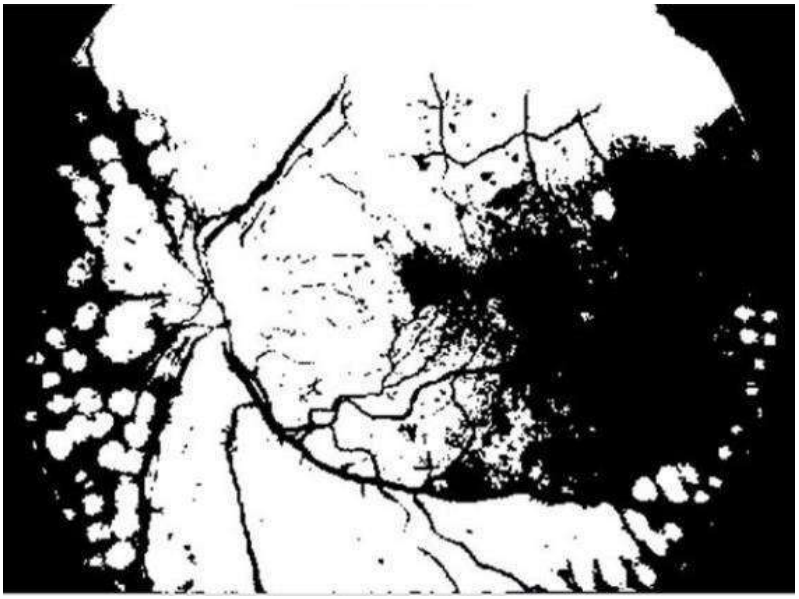


Fig: 3. Vessel detection from spatial weighted Fuzzy c-Means clustering algorithm.

With the proposed method, automatic detection of exudates due to diabetic retinopathy is achieved within short computational period. The accuracy is seemed to be increased using Fuzzy c-Means compared to the results from existing literatures. In order to support the sophisticated operations imposed in the automated image diagnosis by sophisticated embedded systems can be used. These automated systems can be attached with an ophthalmoscope from which the fundus image is entered and a separate visual device can be used to display the output in terms of the normal and abnormal conditions of the disease. Further the above algorithms can be enhanced to produce results including the grade and severity of the disease.

CONCLUSION AND FUTURE WORK

In the present study, a spatially weighted Fuzzy c-Means clustering algorithm for vessel detection in ocular fundus images is proposed. The proposed method not only considered the advantage of the fuzzy system, but also deliberates the spatial neighborhood relation among pixels. The way the weight in the algorithm used plays a key role in improving the performance of the clustering algorithm. The proposed method maintains the computational simplicity and also achieves accurate segmentation results in the case of normal fundus images and images with obscure blood vessel appearance. The proposed method can be further amended by considering the optic disc region, vessel curving and crossings of the retinal image. However, obtained results of proposed method show that the method is able to perform the optic cup and disc detection, but it involves further improvement and parameter tuning to be incorporated to this specific purpose. The future work also aims at applying the shape analysis and classification strategies to the segmented vessels produced by method described in this study. Because of its simplicity and general nature the proposed algorithm seems to be applicable to a variety of other applications. An automated process for the early diagnoses and intervention can hence be of great aid to the patient and ophthalmologist alike in the appropriate supervision of this widespread disease.

ISSN: 0976-3104

ACKNOWLEDGEMENT

Authors acknowledge the great help received from the scholars whose articles cited and included in references of this manuscript. The authors are also grateful to authors/ editors/ publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

CONFLICT OF INTERESTS

We declare that we have no conflict of interest with any of the suggested reviewers and that the paper has not been supported by any grant to declare and have no personal relationships with other people or organizations that could inappropriately influence (bias) their work.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Hoover A, Kouznetsova V, Goldbaum M. [2000] Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3): 203–210.

[2] Niemeijer M, Staal J, van Ginneken B, Loog M, Abramoff MD. [2004] Comparative study of retinal vessel segmentation methods on a new publicly available database. In J. Michael Fitzpatrick and M. Sonka, editors, *SPIE Medical Imaging*, 5370: 648–656.

[3] Asyali MH, Alci M. [2005] Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. *Bioinformatics*, 2(5): 644–649.

[4] Rubini SS, Kunthavai A. [2015] Diabetic retinopathy detection based on Eigenvalues of the Hessian matrix. *Procedia Computer Science*, 47: 311–318.

[5] Abdulrahman AA. [2011] Diabetic retinopathy – An update. *Saudi Journal of Ophthalmology*, 25: 99–111.

[6] Roy MS, Klein R, O'Colmain, BJ, et al. [2004] The prevalence of diabetic retinopathy among adult type 1 diabetic persons in the United States. Arch. *Ophthalmol,* 122: 546–551.

[7] Alqurashi KA, Aljabri KS, Bokhari SA. [2011] Prevalence of diabetes Mellitus in Saudi community. *Ann. Saudi Med*, 31(1): 19–23.

[8] Olson JA, Strachana FM, Hipwell JH, Goatman KA, McHardy KC, et al. [2003] A comparative evaluation of digital imaging, retinal photography and optometrist examination in screening for diabetic retinopathy. *Diabetic Medicine*, 20(7):528–534.

[9] Walter T, Klein JC, Massin P. [2002] A contribution of image processing to the diagnosis of diabetic retinopathy- detection of exudates in colour fundus images of the human retina. *IEEE Transaction on Medical Imaging*, 21: 1236–1243.

[10] Gonzalez RC, Woods RE, Eddins SL. [2004] *Digital Image Processing using Matlab*. Pearson Prentice Hall.

[11] Alghadyan A. [1993] Retinal vein occlusion in Saudi Arabia. Possible role of dehydration. *Ann. Ophthalmol.,* 25(10): 394–398.

[12] Saini HK, Chand O. [2013] Skin segmentation using RBG colour model and implementation of switching conditions. *International Journal of Engineering Research and Applications,* 3(1): 1781–1787.

[13] Jagadish, N, Rajendra AU, Subbanna P, Nakul BS, Teik CL. [2009] Automated diagnosis of Glaucoma using digital fundus images. *Med Syst.,* 33: 337–346.

[14] Yong Y, Huang S. [2012] Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term. *Computing and Informatics*, 26: 17–31.

[15] Bhoyar K, Omprakash K. [2010] Colour image segmentation using fast fuzzy c-means algorithm. *Electronic Letters on Computer Vision and Image Analysis*, 9(1): 18–31.

[16] Sangaiah AK, Thangavelu AK. [2014] An adaptive neuro-fuzzy approach to evaluation of team-level service climate in GSD projects. *Neural Computing and Applications*, 25(3-4): 573–583.

[17] Sangaiah AK., Thangavelu AK, Gao XZ, Anbazhagan N, Durai MS. [2015]. An ANFIS approach for evaluation of team-level service climate in GSD projects using Taguchi-genetic learning algorithm. *Applied Soft Computing*, 30: 628–635.

[18] Khalida NEA. Noora NM. Ariffa NM. [2014] Fuzzy c-Means (FCM) for optic cup and disc segmentation with morphological operation. *Procedia Computer Science*, Vol. 42: 255–262.

[19] Chaudhuri S, Chatterjee S, Katz N, Nelson M, Goldbaum M. [1989] Detection of blood vessels in retinal images using two dimensional matched filters. *IEEE Transactions on Medical Imaging*, 8(3): 263–269.

[20] Osareh A, Mirmehdi M, Thomas B, Markham R. [2003] Automated identification of diabetic retinal exudates in digital colour images. *British Journal of Ophthalmology*, 87(10): 1220–1223.

[21] Wu D, Zhang M, Liu JC, Bauman W. [2006] On the adaptive detection of blood vessels in retinal images. *IEEE Transaction on Biomedical Engineering*, 53(2): 341–343.

[22] Staal J, Abramoff MD, Niemeijer M, Viergever MA, van Ginneken B. [2004] Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4): 501–509.

[23] Winder RJ, Morrow PJ, McRitchie IN, Bailie JR, Hart PM. [2009] Algorithms for digital image processing in diabetic retinopathy. *Computerized Medical Imaging and Graphics*, 33: 608–622.

[24] Cheng HD, Jim-Rong Chen, Jiguang Li. [1998] Threshold selection based on fuzzy c-partition entropy approach. *Pattern Recognition*, 31(7): 857–870.

[25] Kahai P, Namuduri KR, Thompson H. [2006] A decision support framework for automated screening of diabetic retinopathy. *Int. J Biomed. Imag.,* 2006: 1–8.

[26] Nayak J, Bhat PS, Acharya UR, Lim CM. Kagathi M. [2008] Automated identification of different stages of diabetic

retinopathy using digital fundus images. *J Med. Syst*., 32(2): 107–115.

[27] Acharya UR, Tan PH, Subramaniam T, et al. [2008] Automated identification of diabetic type 2 subjects with and without neuropathy using wavelet transform on pedobarograph. *J Med Syst.*, 32(1): 21–29.

[28] Larsen M, Godt J, Larsen N, et al. [2003] Automated detection of fundus photographic red lesions in diabetic retinopathy. Invest. *Ophthalmol Vis Sci*., 44(2): 761–766.

ABOUT AUTHORS

Dr. M. Jagannath *is an Associate Professor in the School of Electronics Engineering at VIT University, Chennai, India. Prior to joining VIT University, he was heading the Department of Biomedical Engineering at SMK Fomra Institute of Technology, Chennai, India. He obtained his Ph.D. from IIT Madras, Chennai in the year 2012. He has served the position of Senior Project Officer at Indian Institute of Technology Madras, Chennai, India. He has taught at Sri Sai Ram Engineering College, Chennai; Madras Medical Mission, Chennai; Vellore Institute of Technology, Vellore. He received IndiraGandhi Sadbhavna Gold Medal Award for Individual Achievement and Service to the Nation from Global Economic Progress and Research Association, India, 2014. He has more than 50 research articles published in various reputed conferences and journals. His research interests are ergonomics, biomedical instrumentation systems, biomechanics, control systems, mechatronic systems and robotics.*

Dr. K. Adalarasu *received the B.E. degree in Electronics and Instrumentation Engineering from University of Madras, Tamilnadu, India, in 1998 and the Ph.D. degree in driver fatigue measurements from Indian Institute of Technology Madras, India, in 2010. At present, he is a Professor in the Department of Electronics and Communication Engineering at PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India. He has published his research work in number of national and international journals and conferences.His research interests include cognitive neuroscience, industrial human safety and ergonomics testing of vehicles.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Reddy et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

CLAMR-AN ENHANCED BIO-INSPIRED ROUTING PROTOCOL FOR WIRELESS ADHOC NETWORKS

Pradeep Reddy[1*], Jagadeesh Gopal[1], Arun Kumar Sangaiah[3]

[1]*School of Information Technology and Engineering, VIT University, Vellore, TN, INDIA*
[2]*School of Computer Science and Engineering, VIT University, Vellore, TN, INDIA*

ABSTRACT

Establishing communication through portable devices without the dependence on or constraints of any central infrastructure is possible in Ad hoc networks. But this very feature of the absence of a central infrastructure and the random mobility of the devices also give rise to multiple problems involving security and routing. The challenge encountered by multicast routing protocols in this ambience is to envisage creating a multi-hop routing within the constraints of the mobility of the host and the bandwidth. The role of multicast routing in wireless ad hoc networks is very significant and pivotal. We intend presenting in this paper a novel on-demand multicast routing protocol known as Cross-Layered Ant Colony Optimization Multicast Routing Protocol (CLAMR). This selected protocol is an improved and enhanced version of the already existing on-demand multicast routing protocol (ODMRP) with the enhancement of Bio-inspired Ant Colony mechanism. The fundamental mechanism of ODMRP is utilized in the proposed method along with the enhanced features of cross layer, applying ant colony optimization. This CLAMR is a robust and efficient protocol with minimal overhead. The scalability and improved performance levels of the proposed algorithm at higher traffic load in comparison to the existing algorithms are evident from the simulations performed using network simulator 2.

Received on: 17th-Feb-2015
Revised on: 24th-April-2015
Accepted on: 12th– May-2015
Published on: 8th–Aug-2015

KEY WORDS

Cross layered, Ant Colony, ODMRP, Wireless Networks

***Corresponding author: Email:** pradeep1417@gmail.com; **Tel.:** + 91-8056600142

INTRODUCTION

Several routing schemes have been projected for the purpose of providing adequate and efficient performance of ad hoc networks. Basing on time based determination of routes; ad hoc routing is classified as proactive [1-4] or reactive routing [1-6]. Routing decisions are continuously made in proactive routing in such a way that the routes are instantly available when transmission of packets is required. Some existing proactive protocols are DBF, DSDV [1] and WRP. Routes in reactive routing protocols [2] are determined on the basis of as-needed i.e., the node requests for a route when there is a need to transmit a packet. TORA, DSR [3], AODV [1], ABR and RDMAR are a few reactive routing protocols. Exchange of routing information in proactive routing consumes a large amount of radio resources. In addition, the validity of the pre-determined routes in an ad hoc network may rapidly be lost corresponding to its rapidly changing topology. Prior studies prove that the reactive protocols are effective in performance when compared with the proactive protocols. In recent years several multicast [6-8] routing protocols along with the existing unicast routing protocols have been proposed for ad hoc networks. Some of the proposed multicast routing protocols are capable of supporting unicast routing also, which itself is a special form of multicast. The two-fold classification of the proposed multicast routing is as tree-based and mesh-based. Tree-based protocols [7, 8, 10] create a tree connecting all its multicast members. Generally, tree-based protocols are considered more efficient in comparison to the mesh-based protocols. But, the absence of an alternate path between the source and destination conditions it as less robust in the contexts of changing topologies. Consequently, failure of every link in a multicast tree may trigger a series of exchanges of control messages for tree re-build. In contrast to the tree-based protocols having only on existing path between any two nodes, mesh-based protocols permit existence of redundant paths between the nodes with their built-in mechanism of provision for alternate paths and as a result failure of a link need not require or initiate a re-computation of a mesh. Mesh-based protocols are proved to be robust as per the previous studies.

ODMRP [6] is a multicast mesh-based ad hoc routing protocol. The source node in ODMRP protocol periodically broadcasts Join Request whenever it has data to be sent in order to discover and maintain the multicast routes. Rest of the nodes, when they receive non-duplicate packet, re-broadcast it. On receiving Join Request, the

THE IIOAB JOURNAL ISSN: 0976-3104

multicast group member node replies with a Join Reply. A route is established through all the subsequent replies by the nodes in the reverse path. Use of soft states in ODMRP marks leaving a group in an automatic timeout. As illustrated, reliance of ODMRP on frequent network-wide flooding may result into a problem of scalability particularly when the source nodes are more. More prominent control packet overhead can be perceived in the context of small multicast group in comparison to the total network.

RELATED WORKS

In recent studies, there is an extensive focus on the discovery of multipath and extension of QoS of the on-demand routing protocol in such a way as to address the issues pertaining to the single path routing protocols such as AODV [1, 2, and 9] and DSR [1-6]. Ant colony-based multi-path QoS-aware routing (AMQR) [11, 13, 14, and 15] integrates link-disjoint multipath routing and swarm intelligence to choose multiple paths for providing QoS services [16, 17, 18, 19]. The nature of the ad hoc routing protocols whether proactive or reactive depends on the factor of when the routes are determined. With the process of a continuous exchange of routing information among [20] nodes, the routes to destination are pre-determined in proactive routing protocols. When packets need to be transmitted, immediate availability of routes is ensured. Wireless Routing Protocol (WRP) [1-6] and the Destination Sequenced Distance Vector (DSDV) [2] routing protocol are a few instances of proactive routing protocols [1-6]. In contrast, routes are determined on-demand only in reactive routing protocols. A node sends a query for a route to the network when required to transmit a packet. TORA, DSR and AODV [9] are some much known reactive ad hoc routing protocols. But exclusively proactive or reactive routing protocols do not suit the requirements of MANETs. More frequent changes in network topology than the routing requests are possible in MANETs. In such circumstances, the information of routing generated by the proactive routing protocols stays stale. On the other hand, the reactive routing protocols [1-6] conduct a global search for on-demand route discovery process requiring considerable traffic control. This can easily result in saturation in the bandwidth constrained MANETs in a very short interval of time. So, the literature proposes a number of ad hoc multicast routing protocols which are much more than being just extensions of the unicast routing protocols under reference. This is exactly the opposite of what is experienced in wired networks with the several multicast routing protocols acting just as multicast extension counterparts of their corresponding unicast routing protocols. An ad hoc routing protocol to be robust and efficient, must not depend or rely on any implied unicast routing protocol for determining the route or updates. Extensive work has been done for evaluating the performance of the protocols of routing in ad hoc networks.

PROPOSED MODEL

CLAMR can be interpreted as a QoS enabled multipath routing protocol developed on the basis of the foraging behavior of ant colony [15, 21]. The source generates both the ant agents (called reactive FW_ANT) to find multiple paths to the destination and BW_ANT to set up return paths [18]. The respective qualities of the paths are indicated in the form of pheromone table. NHA is considered as metric during the route discovery phase to assess the goodness of the highly available links and nodes. The NHA is considered for finding the path. The metrics used for the computation of NHA [15, 19] are the availability of nodes and links. The NHA can be explained as the probability to discover the next hops [19], which is the node and link availability for routing on a path.

Next hop availability = Link probability × node probability ----------(1)

At the phase of route discovery, a source node desirous of transmitting information to the destination node first verifies the trusted neighbors. Then such nodes with greater NHA [21] than threshold are selected. Next, the source node initiates broadcasting FANT to all its neighboring trusted nodes having the NHA in order to control the routing overhead.

Any intermediate node receiving the FW_ANT, first verifies if its own address figures in the path field. If the address is present, it discards the FW_ANT at that stage itself to eliminate further loops. Or else it attaches its address to FW_ANT and then initiates broadcasts by its NHA values to all its trusted and stable neighbors. In the process of searching for the destination, the FW_ANT collects delay of transmission of each link, delay of processing at each node, each link's available capacity and visited number of hops. On FW_ANT arriving at the destination node d, the destination node first computes the path preference value, by employing end-to-end delay parameter for only such paths meeting the threshold values prescribed by the user and thus generate BW_ANT.

Visiting nodes in the path are stacked in the FW_ANT. By popping up such nodes present in the stack, the BW_ANT is unicasted to the source node.

The pheromone value [21] gets updated when the backward ant reaches any middle node other than the destination node. The pheromone value gets updated in the routing pheromone table of node i as $T_{i,n}$ [15] and it is updated as

ISSN: 0976-3104

$$T_{i,n} = \left(1 + T_{i,n}\right) P(K)_d \quad \text{---------- (2)}$$

Where $P(K)_d$ [15] is the path preference value of the kth path that satisfied the QoS requirements for the destination d.

Ant colony optimization for ODMRP

In the mesh based ODMRP multicast, packets are transmitted to the destination by employing the concept of a forwarding group. It can be termed a 'reactive protocol' as it makes use of 'on-demand' procedures for dynamically building up routes and maintaining multicast group membership. The drawbacks inherently present in tree protocols, such as intermittent connectivity, configurations of tree and computations of shortest paths can be avoided through the use of mesh topology. For the maintenance of multicast group members, ODMRP makes use of a soft-state approach. Explicit control information is not required to desert the group. We intend to provide an elaborate explanation of the functioning the process of mesh creation of enhanced ODMRP duly considering the ant colony optimization. The network is flooded with FW-ANT message whenever the source node has data to transmit. The intermediate node, after receiving a (non-duplicate) FW-ANT packet, stores and updates the information regarding the upstream node. The Intermediate nodes continue to flood the FW-ANT packet. After receiving the FW-ANT packet, the receiver builds the PHEROMONE TABLE and transmits the BW-ANT to its neighbors. Then, all the nodes receiving the BW-ANT packets verify their individual IDs against the ID contained in the BW-ANT packet. If a matching is found, the node becomes the forwarding member of the group by setting the FG_Flag (Forward Group Flag). Further propagation of FW-ANT packet continues, till it arrives at the source. Through the broadcasting of FN-ANT [15, 21] packets at regular intervals, the senders maintain an update of the multicast group. The following data structures are required to be maintained by each host that runs ODMRP:

- **Routing table** – Each node reactively creates a route table and maintains it. When a non-duplicate FW-ANT containing high pheromone value is received, corresponding entry is inserted or updated. The node stores the information regarding the destination and the subsequent hops to the destination. Next hop information during the transmission of BW-ANT is provided by the route table.

- **Forwarding group table** – The multicast group information is stored in the forwarding group table by the node acting as a forwarding group node of a multicast. It records the group ID of the multicast and the time stamp.

- **Trust Pheromone table** – It is a table that contains the trust value of the neighbors depending on the threshold value from end-to-end delay.

- **Neighboring Information table** – This table contains the required information about all the set of neighbors in any specific network.

This table contains the required information about all the set of neighbors in any specific network.

Cross layer model

MAC and NETWORK layers are fused in the proposed cross layered model in order to achieve a cross layer factor. The load of the node and the bandwidth available for a specific node are calculated. We derive a cross layer factor *clfact* [16] on the basis of these calculations. This cross layer factor can be used as a metric for determining the path from source to destination. Use of the enhanced ODMRP mechanism with Ant Colony Optimization [20, 21] can find all the source- to- destination corresponding routes in network layer. We compute the length of the node and the available bandwidth in each node of the specific routes.

Let B(s) be the available bandwidth at source node s. and B(r) is the available bandwidth at receiver node r, then

B(s,r) = Min(B(s), B(int1), B(int2)……B(r)) ----------(3)

Where B(s,r) is the available bandwidth for the entire link between source node and receiver node.

Then, apply this mechanism of cross layer to the paths provided by the network layer and select such route having effective *clfac* for the given data rates. In comparison to the original ODMRP mechanism, we get more efficiency and reduced overhead.

1. Calculate the node load for a node i in the network.

$$nodeload_i = \frac{queue\_len_i}{queue\_len\_nodes} \quad \text{----------(4)}$$

2. Calculate the available bandwidth for a node i

ISSN: 0976-3104

$$Bdw_i = Bdw_{ch} * \left(\frac{t_i}{tot\_time} \right) * 0.8 \text{ ---------- (5)}$$

Where Bdw_i is the bandwidth of node i, Bdw_{ch} is the channel bandwidth and 0.8 is the weight factor.

3. Apply cross layer design over network and Mac layer parameters.

$$clfactor = \left(Max(Bdw_i, nodeload_i) \right) \text{ ---------- (6)}$$

Recovery of intermediate node failure

On identification of any route-disruption resulting out of the mobility of the subsequent node along that specific route, the network layer at an intermediate node (referred from now on as FP or failure point) dispatches a packet, indicating Path Failure Notification (PFN) to the source. When the PFN packet is received, every intermediate node invalidates a particular route and blocks further traffic of packets through that route to that specific destination. However, the PFN can be discarded, if the intermediate node has knowledge of any alternate route, which can be utilized for further support in communication. Or else, the node just disseminates the RFN towards the source. The source shifts into snooze mode after the receipt of PFN and carries out subsequent recovery mechanism. The source continues to be in the snooze mode till it receives the restoration notification of the route through the packet of a Route Re-establishment Notification (RRN).

Use of Path Failure Timer checks the source from remaining in the snooze mode indefinitely expecting the arrival of an RRN that could either have been delayed or lost. This timer starts on the source receiving the first PFN. On expiry of the Path Failure Timer, the frozen timers start (as if they received an RRN) and permit the congestion control mechanism of the TCP to take care of the failure.

RESULTS AND DISCUSSION

Radiographic For the purpose of simulation, we utilized NS-2 [22] network simulator. A network model of 50 randomly placed nodes covering an area of 1000m x 1000m figure in our simulation. The simulator functions with a range of 250 meters radio propagation with channel capacity of 2 Mbits/Sec. The size of the multicast group varies with one source in each group sending at the rate of 20 Packets/Sec. 300 seconds of simulation is executed in each simulation. We collected data and averaged over the results arrived at by changing the send numbers in different multiple runs for each changing scenario.

The proposed model utilized the following metrics in order to estimate the performance of the proposed CLAMR mechanism:

•**Packet Delivery Throughput**: This can be described as a corresponding quantification of data packets received at the destination and the dispatched data packets by the CBR sources.

•**End-to-End Delay of Data Packets**: It is the delay in time between the times of packet origin at the source and packet reaching at the destination. We do not consider the enrooted lost data packets here. But the delay metric certainly considers the delays cropping up due to route discovery, queuing and transmission. Performance of our approach is evaluated by a comparison against the approaches of ODMRP and CL-ODMRP. The performance of the proposed system in comparison to the methods already in existence explicitly appears to be much enhanced as can be perceived from **Figure–1** and **Figure–2**.

The performance of the protocol is further evaluated in varying network scenarios such as the speed of node moving, size of the multicast group and multicast group number.

Node moving speed: 20 multicast destinations are set in this simulation with only 1 source sending the data packets. The rate of traffic generation is 10 packets per second.

Figure-1 illustrates the performance of packet delivery ratio of CLAMR with CL-ODMRP and ODMRP, respectively, in varying moving speeds. A similarity in performance can be perceived with CL-ODMRP and ODMRP while the speed of the node moving is low. High mobility of the node, generating high dynamics in a built forwarding structure is the reason for this.

THE IIOAB JOURNAL
ISSN: 0976-3104

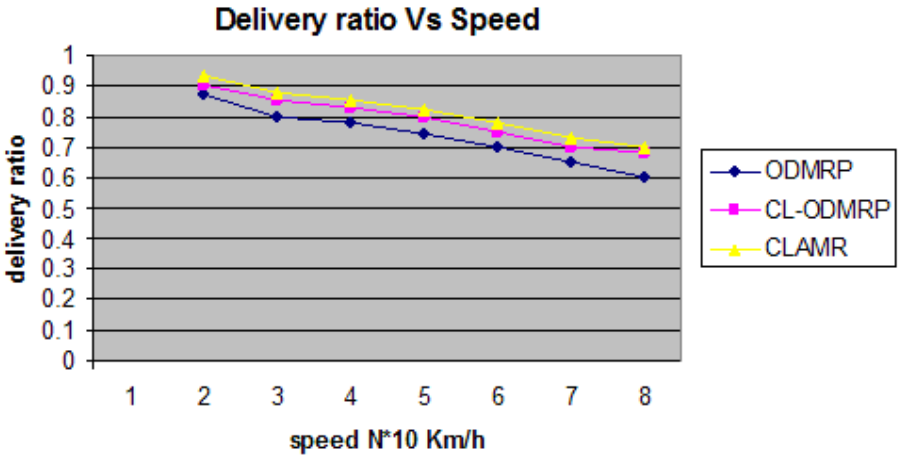


Fig: 1. Packet delivery rate vs speed

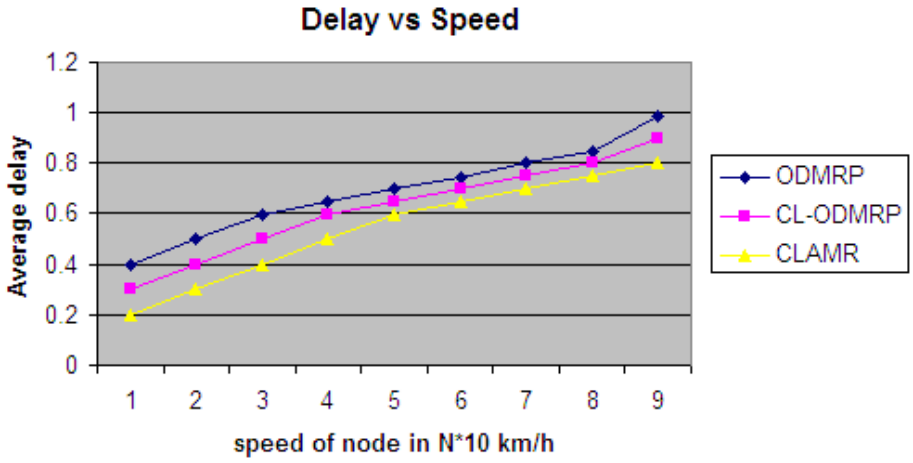


Fig: 2. Comparison of Average delay with CLAMR

Figure-2 displays a comparison of the average delay of CLAMR against CL-ODMRP and ODMRP, respectively in the context of different moving speeds. A reduction of delay upto 15% as against ODMRP can be achieved through the introduction of cross layer- based ant colony optimization strategy CLAMR.

CONCLUTION

In this paper, the proposed model Cross-Layered Ant Colony Optimization Multicast Routing Protocol (CLAMR) utilizes the nature inspired ant colony optimization algorithm to find multiple paths during the route discovery phase. A QoS factor has been included to find goodness of the path. The QoS factor is obtained from the cross layer technique. The proposed model yields effective throughput and less delay with the optimizations.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

THE IIOAB JOURNAL ISSN: 0976-3104

REFERENCES

[1] Ade SA, Tijare PA. [2010] Performance Comparison of AODV, DSDV, OLSR and DSR Routing Protocols in Mobile Ad Hoc Networks, *International Journal of Information Technology and Knowledge Management*. July-December 2010. 2 :545–548

[2] SS Tyagi. RK Chauhan Performance Analysis of Proactive and Reactive Routing Protocols for Ad hoc Networks International *Journal of Computer Applications* (0975 - 8887) Volume 1 - No. 14.2010.

[3] C Siva Ram Murthy and BS Manoj.[2004] Ad Hoc Wireless Networks Architectures and Protocols, PRENTICE HALL, 2004.

[4] Abolhasan M, Wysocki TA, Dutkiewicz E. [2004]'A review of routing protocols for mobile ad hoc networks', *Elsevier J Ad hoc Netw*, 2: 1–22

[5] Luo J, Ye D, Xue L, Fan M. [2009]'A survey of multicast routing protocols for mobile ad-hoc networks', *IEEE Commun Surv Tuto,* 11 (1): 78–91

[6] Nasipuri A, Das SR. [1999] 'On-demand multipath routing for mobile ad hoc networks'. *Proc. Eighth Int. Conf. on Computer Communications and Networks*, 64–70.

[7] Gerla M, Lee Y, Park J, Yi Y. [2005]'On demand multicast routing with unidirectional links'. Proc. IEEE WCNC'05, 2162–2167

[8] Zhang B, Mouftah HT.[2006] 'Destination-driven shortest path tree algorithms', *J High Speed Network.* 15(2):123– 130

[9] Park J, Moh S, Chung I.[2008] 'A multipath AODV routing protocol in mobile ad hoc networks with SINR-based route selection'. *IEEE* Int. Symp. Wireless Communication Systems, ISWCS'08, 21–24 October 2008, 682–686.

[10] Lee SJ, Gerla M. [2001]'Split multipath routing with maximally disjoint paths in ad hoc networks'. *IEEE* Int. Conf. on Communications, 10:3201–3205.

[11] Misra S, Obaidat MS, Dhurandher SK, Verma K, Gupta P. [2009] Using ant-like agents for fault-tolerant routing in mobile ad-hoc networks. *ICC*, 1–5.

[12] Arun Kumar BR, Lokanatha C Reddy, Prakash.S.Hiremath. [2008] Performance Comparison of Wireless Mobile Ad-Hoc Network Routing Protocols *[JCSNS International Journal of Computer Science and Network Security*.8 (6):

[13] Misra S, Venkata Krishna P, Bhiwal A, Singh Chawla A, Wolfinger BE, Lee C.[2011] 'A learning automata-based fault-tolerant routing algorithm for mobile ad-hoc networks', *J Supercomput.*. 1–20, doi: 10.1007/s11227-011-0639-8

[14] Mehfuz S, Doja MN. [2008]. Swarm intelligent power-aware detection of unauthorized and compromised nodes in MANETs', *J Artif Evol Appl*, vol. 2008, article 3 (January 2008), Article ID 236803, 16 pages, doi: 10.1155/2008/236803

[15] P Venkata Krishna, V Saritha, G Vedha, A Bhiwal, AS Chawla.[2012] Quality-of-service-enabled ant colony-based multipathrouting for mobile ad hoc networks, *IET Commun* 6(1): 76– 83

[16] P Venkata Krishna and Iyengar.[2007] A Cross Layer Based Qos Model For Wireless And Mobile Networks *Journal of Mobile Communication* 1 (4): 114–120.

[17] P Venkata Krishna. [2008]A study of quality of service metrics for wireless systems, Ph. D thesis, VIT University,

[18] Zou Yuanyuan, Tao Yang.[2012] A method of selecting path based on neighbor stability in Ad Hoc network, Computer Science and Automation Engineering (CSAE), 2012 *IEEE International Conference* on , .2:675–678, 25–27 ay 2012

[19] Reddy CP, Krishna PV. An efficient bandwidth management framework for wireless mesh networks, Innovative Computing Technology (INTECH), 2012 Second International Conference on, 102,106, 18–20 Sept. 2012 doi: 10.1109/INTECH.2012.6457765

[20] Reddy Ch P, Venkata Krishna P. [2014]Cross layer based congestion control in wireless mesh networks (2014) Bulgarian Academy of Sciences, *Cybernetics And Information Technologies* , 14(2): July, 2014. DOI: 10.2478/cait-2014–0020.

[21] Reddy Ch P, Venkata Krishna P. [2014] Ant-inspired level-based congestion control for wireless mesh networks. *Int J Commun. Syst doi*: 10.1002/dac.2729

[22] NS-2, the NS Manual (fonnally known as NS Documentation).

[23] Available at http://www. isi.edu/nsnam/ns/doc.

ABOUT AUTHORS

Dr. Pradeep Reddy.CH *is currently working as Associate Professor with School of Information Technology and Engineering, VIT University, Vellore, India. He has a total of 8 years of experience in both teaching and research. He received his B. Tech in Computer Science and Engineering from PBR VITS, JNTU in 2004, Andhra Pradesh, India and M. Tech in Computer Science & Engineering from VIT University, Vellore, India. He did his Ph.D. in Computer Science and Engineering from VIT University, Vellore in 2014. His research interests include Mobile Technologies, Wireless Networks and Open Source Technologies. He has published several papers in national and international refereed journals and conferences. He is a member of various professional organizations such as CSI, ACM and IAENG.*

Dr. Jagadeesh Gopal *is currently working as Division Chair, Software Engineering with School of Information Technology and Engineering, VIT University, Vellore, India. He has a total of 15 years of experience in both teaching and research. He completed his B. Tech in Computer Science and Engineering and M. Tech in Information Technology. He is pursuing his Ph.D from VIT University, Vellore. His research interests include Software Engineering, Soft Computing, Wireless Networks. He has published several papers in national and international refereed journals and conferences. He is a member of various professional organizations such as CSI, ACM, ISTE and IAENG.*

Dr. Arun Kumar Sangaiah *obtained his Doctor of Philosophy (Ph.D.) in Computer Science and Engineering from VIT University, Vellore, Tamil Nadu, India. He is presently working as a Associate Professor in School of Computing Science and Engineering, VIT University, India. His experience and areas of interest focus on Software Engineering, Soft Computing, Wireless Networks, Bio-Informatics, and Embedded Systems. He has author of more than 100 publications in different journals and conference of National and International repute. Dr.Arun Kumar is member of international advisory board of IJIIT (IGI publisher) and editorial board member of IJIES, Japan. Also, he has served a special guest editor of various international journals (Inderscience, Hindawi, IGI publishers). He is active member for Compute Society of India. He has guided many research students and post-graduate students in the field of software engineering, communication networks, ad hoc networks, database, and soft computing techniques.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Rajesh et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

DATA AGGREGATION IN NOISY WIRELESS SENSOR NETWORKS USING CHAOS THEORY

G. Rajesh*, S Mathivanan, B. Vinayagasundaram

Dept. of Computer Science, Anna University, Tamil Nadu, INDIA

ABSTRACT

In environmental monitoring applications, the data periodically sensed by Wireless sensor networks (WSNs) have a strong redundancy. The Network consumes more power to transmit the redundant data packets to sink through intermediate nodes which reduces network lifetime. To overcome the redundant data transmission, Chaos Theory based Data Aggregation (CTAg) prediction method is proposed here. The proposed method minimizes the number of packets forwarded to sink node by eliminating redundant packets by using chaos theory based prediction technique. The proposed approach is evaluated using temperature monitoring application dataset collected from Intel Berkley Lab. The CTAg method significantly reduces communication redundancy, data redundancy, and mean square deviation error and also increase prediction accuracy, in turn evidently proved that the lifetime of the network is improved.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 8th –Aug-2015

KEY WORDS

Wireless sensor network; Redundancy; Data aggregation; Time series prediction; Chaos theory;

***Corresponding author: Email:** raajiimegce@gmail.com, gr@annauniv.edu **Tel:** +91-044-22516023

INTRODUCTION

Wireless sensor networks (WSNs) are groups of sensor nodes that are tiny with low power and low cost. Each node consists of a sensing unit, processing unit and communication subsystem [1]. These sensor nodes collect real-time data from the physical environment and transfer the collected data to base station. They are widely used in military, health science, and commercial applications. For example, in health science, a doctor can monitor the physiological information about the patients remotely. In the commercial application, it is used for managing inventory and monitoring product quality [2].

The energy consumption in a wireless sensor network is due to sensing, computing, communication and mobility of nodes. The energy for communication is more expensive, which affects the energy of intermediate nodes during multi hop transmissions of data packets compared to sensing and computation. Energy dissipation during communication that is caused by transmitter or receiver is considered as E_{elec} and the power dissipation of the transmitter amplifier is taken as ε_{amp}. The power consumed by the transmitter for k-bit packet transmission [3, 6] to a distance'd' and assume d^m as path loss.

$$E_{Tx}(k,d) = E_{elec} * k + \varepsilon_{amp} * d^m * k \qquad (1)$$

Let $E_{elec} = E_T(k,d) = E_R(k) = 50nJ/bit$, ε_{amp}=100pJ/bit/m^2 and $m = 2$.

$$E_{Rx}(k) = E_{elec} * k \qquad (2)$$

Where $E_{Rx}(k)$ is the k–bit message in receiving of energy.

In an environmental monitoring application, information collected by sensor nodes tends to exhibit strong temporal correlation. This redundancy leads to communication overhead, consumes node energy and reduces network efficiency. To overcome this issue, the data aggregation techniques [5] were introduced. Wireless sensor networks are partitioned into source node, intermediate node

THE IIOAB JOURNAL ISSN: 0976-3104

(aggregator node) and sink node. The aggregator node performs the data aggregation [4]. It collects data from multiple sensor nodes, which are fused and transferred to the base station. Thus the data aggregation technique eliminates redundancy and minimizes numerous transmissions and thus preserves the network energy.

Chaos theory is the study of complex, nonlinear, dynamic systems [16]. It is a branch of mathematics that deals with systems that appear to be orderly (deterministic) but, in fact, harbor chaotic behaviors. It also deals with systems that appear to be chaotic, but, in fact, have underlying order. Chaos theory studies the behavior of dynamical systems that are highly sensitive to initial conditions, an effect which is popularly referred to as the butterfly effect. The deterministic nature of these systems does not make them predictable. This behavior is known as deterministic chaos, or simply chaos.

The proposed Chaos Theory based Data Aggregation (CTAg) approach predicts the next period data based on the earlier sensed data. The data are collected from ordinary sensor nodes. The aggregator node made a decision between prediction error and prediction threshold, in order to make a conclusion whether to transmit current data to sink node. This technique eliminates the redundancy and minimizes the communication density. The proposed method can also provide better prediction accuracy and prolong sensor node lifetime when compared to other traditional prediction based data fusion techniques. This paper is organized as follows: Section II discusses the prediction base data aggregation and chaos theory based data aggregation (CTAg). Moreover, implementation details and results are given in Section III. Finally, conclude and define our future work in Section IV.

MATERIALS AND METHODS

Prediction based data aggregation- a survey

The energy management is one of the major issues in wireless sensor networks. A sensor utilizes high energy for communication rather than sensing and processing. The redundant communication in noisy channels causes the depletion of network energy. The prediction based data aggregation approach reduced unnecessary data transmission and so energy expenditure in communication subsystem was minimized. Hyuntea Kim et al., [8] exploited linear data prediction method to improve communication efficiency and to minimize energy consumption with data correlation. As the model is designed considering some factors such as the selective transmission, it reduced data accuracy and adjustments in aggregation period caused the network to meet the additional delay. Guiyi Wei et al., [7] proposed a method that saves network energy and eliminates redundant communication by exploiting prediction based data aggregation protocol. However, in this method synchronization time increased due to synchronization has to be done prior to each transmission. Guorui Li et al., [9] proposed an Auto Regressive Integrated Moving Average Model (ARIMA) that predicts the next time value based on the previous observed values. When the prediction error is less than the preconfigured threshold value the aggregator would not transmit the data sensed by the source node. Otherwise, it transmits the data to sink node. Therefore ARIMA model reduced the amount of data transmitted between the ordinary sensor node and aggregator node. Since this method performed aggregation on the ordinary sensor node and aggregator node it increased the computational complexity and reduced accuracy. Rajesh G et al., [10] proposed the data fusion method using Simpson's 3/8 rule to forecast next time data based on the early sensed information. When prediction error is greater than the prediction threshold the cluster head transmits the actual sensed value to the base station. Otherwise, it would not transmit data to the base station. This method reduced unnecessary transmission between cluster head and base station. However, this method provides less prediction accuracy since the deviation error is increased between subsequent values. There are several data fusion techniques in Wireless sensor networks. The main features of the proposed work are that it, Improves the performance of the forecast and Performs less computation to obtain the forecasted data.

Chaos theory based data aggregation (CTAg) technique

The typical features of chaos include: 1) Nonlinearity. If it is linear, it cannot be chaotic. 2) Determinism. It has deterministic underlying rules every future state of the system must follow. 3) Sensitivity to initial conditions. Small changes in its initial state can lead to radically different behavior in its final state. Long-term prediction is mostly impossible due to sensitivity to initial conditions. A dynamic system is a simplified model for the time-varying behavior of an actual system [17]. These systems are described using differential equations specifying the rates of change for each variable. A dynamical system of dimension N system first–order differential equations for N variables $x_1(t), x_1(t) \ldots x_N(t)$ evolve with time t according to,

$$\hat{x}_1 = f_1(x_1, x_2, \ldots x_N, t) \quad (3)$$

$$\hat{x}_2 = f_2(x_1, x_2, \ldots x_N, t) \quad (4)$$

$$\hat{x}_N = f_N(x_1, x_2, \ldots x_N, t) \quad (5)$$

Where f_1, f_2 are assigned functions and a dot is a derivative with respect to time.

ISSN: 0976-3104

The system following Characteristics of a Chaotic System:

- Sensitivity to initial conditions
- Non-linear
- Dynamic and mixed topology system and Continuous or periodic time.

So that the Chaos is the aperiodic long–term behavior in a deterministic system that exhibits sensitive dependence on the initial condition. These characteristics enables chaos theory based data aggregation (CTAg) prediction method is suitable for eliminating data redundancy in WSNs.

Considered a hierarchical wireless sensor network G (SN, E) where, SN represents the sensor nodes and E represents links connecting the nodes. These sensor nodes collect weather monitoring data (Temperature, Humidity) periodically. Each node transmits data to sink node through the intermediate node or aggregator node (A). The aggregator (A) will perform data fusion by eliminating redundant data using chaos theory before transmitting the gathered data towards the base station. This will minimize the amount of data transmitted between aggregator node and sink node.

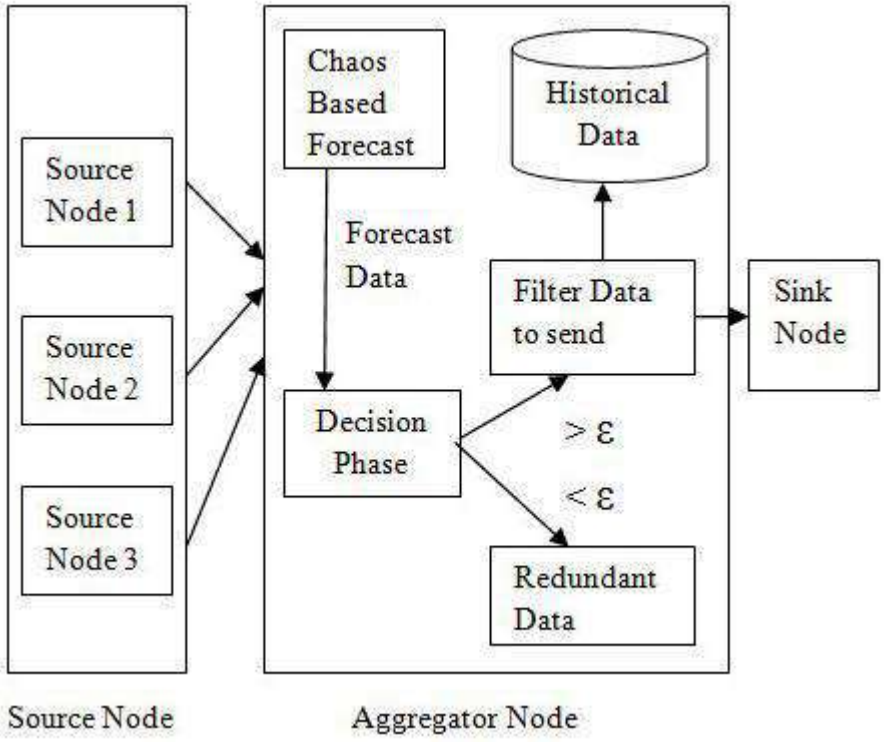


Fig: 1. Architecture for chaos theory based aggregation.

Figure–1 illustrates, the CTAg architecture that consist of two distinct phases, chaos based forecasting and decision phase. The chaos based forecasting phase uses previous time series measurements representing (6) to compute next time period data using (7) and (8). The decision phase will accomplish the comparison between forecast errors e_t and prediction threshold ε to make a decision whether to send actual sensed data to the sink or not. The prediction error was calculated using (10). The prediction error is greater than the prediction threshold the aggregator node transmits sensed data to base station. The aggregator does not send sensed data to sink node when forecast error is less than the prediction threshold.

$$x_t = \left(x(t_n), x(t_{n+\tau}), x(t_{n+2\tau}), \ldots, x(t_{n+(m-1)\tau})\right) \quad (6)$$

Where x_t is the chaotic time series, m is the embedding dimension and τ is the delay time.

$$x'_{t+n} = f(x) \quad (7)$$

Where x'_{t+n} the prediction data of next time and n is step length of time series forecast.

$$f(x) = x_t + \left(\tau - (\tau - n)\right) * FE \quad (8)$$

ISSN: 0976-3104

Where, $\tau = \frac{x - x_{n-t}}{h}$ x is the recent data, x_{n-t} is the previous time data and h = t-t_n. Let be t is the time for recent data sensed and t_n is the time for earlier sensed data value.

The Forward Error (FE) Δ is the product of Minimum Forward Error Value (MFEV) with current sensed data and divide to embedded dimension is the order of MFEV.

$$FE(\Delta) = \frac{\Delta^m x_t}{m!} \quad (9)$$

$$e_t = x_{t+1} - x'_{t+1} \quad (10)$$

Where e_t is the prediction error at period t.

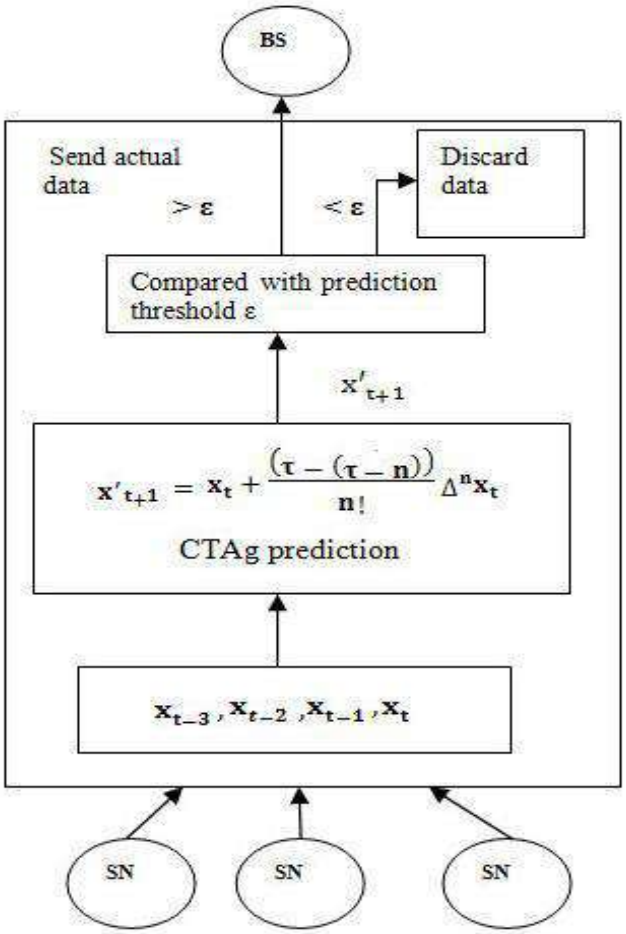


Fig: 2. Chaos time series forecasting model.

Figure–2 illustrates, when an intermediate node performs chaotic time series forecasting is evaluated future value x'_{t+1} based on the Sensor Nodes (SN) gathered previous time temperature data such as $x_{t-3}, x_{t-2}, x_{t-1}, x_t$... after computing predicted value, to measure the forecast error $x_{t+1} - x'_{t+1}$ mean that the deviation between actual sensed value and prediction value, then it is compared with prediction threshold value ε. If the prediction error is higher than the threshold, the actual sensed data transmitted to sink node. Otherwise, the aggregator is not forwarding the packet to the Base Station (BS). Here the aggregator node, predicted threshold based will perform filtering the redundant data. The predicted threshold set as the mean value of the prediction error. The predicted threshold is adaptive. The less deviated data is discarded from gathered continuous time series data set. The more relevant data or non-redundant data is towards base station (BS). But the major and most significant limitation of

THE IIOAB JOURNAL ISSN: 0976-3104

chaos theory is the feature that defines it, sensitive dependence on initial conditions that may affect prediction accuracy.

Algorithm 2, illustrates the aggregator node receives predicted data at the point of time $t+1$. The aggregator decides to transmit data towards the sink or not by doing a comparison with predicted error and prediction threshold.

Algorithm 1: Chaos Forecasting Phase

Input:

x_t → Time series sensor data

Output:

x'_{t+1} → Forecasted Data

Process:

1. Get earlier time ser $\Delta^m x_t$

x_t = {x(0), x(1), m2),x(3),..,x(t)}

2. If History_size > 0

2.1. Compute $\Delta x_t = x_t - x_{t-1}$

2.2. Compute the next time series of the data

$$x'_{t+1} = x_t + \frac{(\tau-(\tau-n))}{n!}\Delta^n x_t$$

2.3. Return x'_{t+1}

3. End if

Algorithm 2: Decision Phase

Input:

x'_{t+1} Forecasted Data

x_{t+1} Actual Sensed Data

Output:

Returned Actual Sensed Data

Process:

1. At the period t aggregator node computes x'_{t+1}

2. At period t+1, aggregator node receiving actual data x_{t+1}

3. If $(x_{t+1} - x'_{t+1} < \varepsilon)$

3.1. Redundant Data

3.2. [History] = Sensed data

4. Else

4.1. Transmit Non-Redundant Data

4.2. [History] = Forecast data

5. End If

THE IIOAB JOURNAL ISSN: 0976-3104

If the prediction error variation is less than a threshold value the aggregator assumes that the sensed data is redundant data. So it does not forward to sink node. Otherwise, assume that the non-redundant data and it conveyed to sink node.

RESULTS

The CTAg algorithm simulated using OMNET++ simulator with MIXIM package. The actual time series data are collected from wireless sensor networks. Here N=100 nodes are uniformly used to collect data from a sensing area of $A = 500 \times 500\ m^2$. The data are traversed using AODV protocol. In this network, the sensor node spends E_{Tx} energy to carry the data 'd' distance to next node and utilizes energy E_{Tr} for receiving data.The total communication energy is e_{ij}. The CTAg is intended to reduce redundant communication and conserve the network energy e_{ij}. In this section, the CTAg algorithm is applied to the temperature monitoring application. The dataset is collected from Intel Berkley Lab [17].

Table–1 represents the table contains date, time, and node id of the sensed data. The Epoch is a monotonically increasing sequence number from each node. In this dataset we consider the Time (t), Mote id (MID), Epoch (E) and Temperature (T) for evaluation. The CTAg algorithm performance is compared with conventional method of Kalman Filter [KF] prediction based data aggregation [15]. It is observed that the proposed algorithm is better among the data aggregation technique.

Table: 1. Example temperature application dataset

Date	Time	Epoch	Mote id	Temperature
03/03/04	09:45	22	1	20.145
03/03/04	10:22	23	1	20.165
03/03/04	10:46	24	1	20.165
03/03/04	11:16	25	1	20.169
03/03/04	11:20	26	1	20.244

Performance analysis

Table–2 represents the comparison of quantitative performance of proposed CTAg algorithm and KF. This table contains the actual data, predicted data, prediction error, and data transmission status and prediction threshold. Here, the prediction threshold as 0.28 is used to evaluate the performance of prediction. In Data Transmission Status, 0 indicates that the data is similar to the previously sent data and 1 indicates that the data is different from previous data. The aggregator will transmit to the sink node data marked as 1 and discard data marked as 0. Hence the communication overhead is reduced. The communication overhead is the ratio of the sensor node actual power consumes the number of packet forwarded to base station (BS) to power consumes after aggregated packet transmitted to base station (BS). Thus the CTAg aggregation method reduces the redundant packet transmission and power consumption.

Figure–3 represents comparison of actual sensed temperature with Kalman Filter [KF] and CTAg method forecasted the temperature (T) in different time interval. The temperature predicted using CTAg method is close to actual sensed value due to small prediction error and KF method has high variability with sensed data and CTAg due to high deviation error between the sensed and predicted values. It can also be seen that the predicted value of CTAg varies consistently with the observed value and forms a steady growth curve with almost constant prediction error. This stable prediction error rate eliminates the need for the fixing prediction threshold using soft computing technique.

THE IIOAB JOURNAL
ISSN: 0976-3104

Figure–4 shows the amount of deviating data that has been transferred. The CTAg method performs better for the forecast threshold up to 0.28 when compared to KF. The CTAg restricts redundant packet transmission from aggregator to sink node compared to the Kalman filter.

Table: 2. Sample resultant data

Method	Actual Data	Predicted Data	Prediction Error (%)	Data Transmission Status
CTAg	20.51	20.77	0.26	0
	20.54	20.79	0.25	0
	20.61	20.92	0.32	1
	20.86	21.04	0.18	0
KF	20.51	20.81	0.30	1
	20.54	20.78	0.24	0
	20.61	20.92	0.31	1
	20.86	21.13	0.27	0

Figure–5 based on the forecast threshold 0.03, 0.08, 0.13, 0.18, 0.23, and 0.28. When the predicted threshold is 0.28 the difference between the observed and forecasted value is minimum. Hence it can be inferred that considerable accuracy is achieved at prediction threshold 0.28.

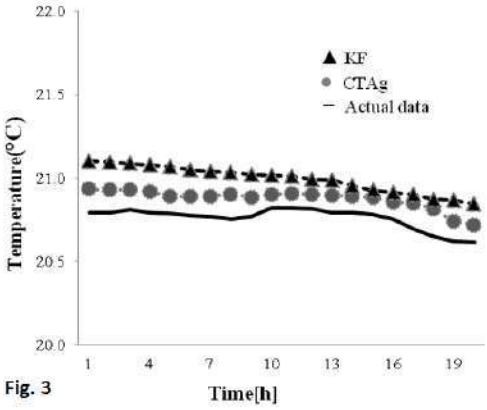


Fig. 3

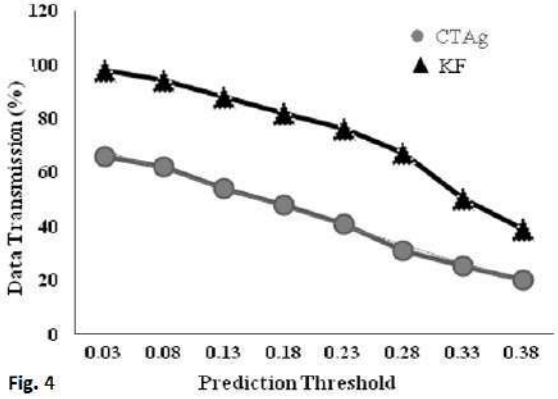


Fig. 4

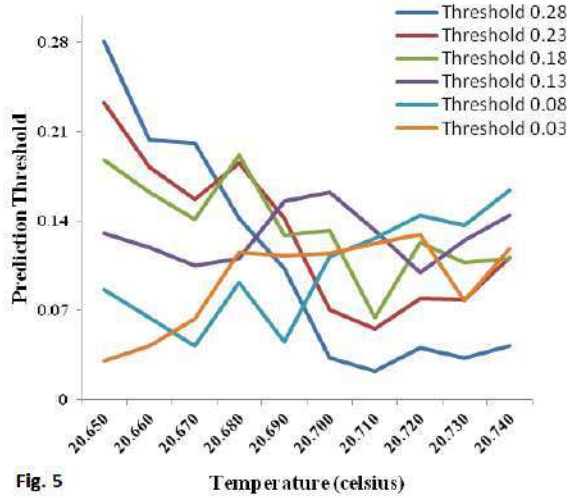


Fig. 5

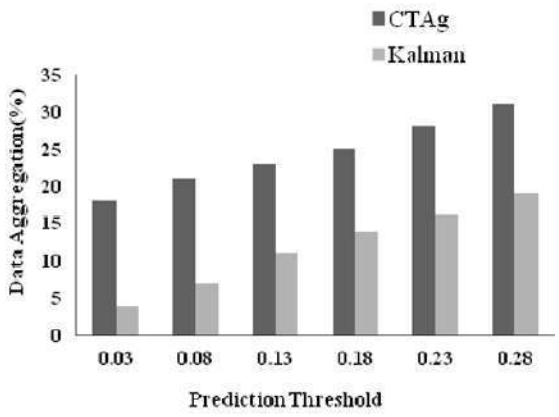


Fig. 6

Fig: 3. Comparison of prediction and sensed temperature in time. Fig: 4. Comparison of data transmission (%) in variable temperature. Fig: 5. Comparison of adaptive threshold. Fig: 6. Comparison of data aggregation (%) in different threshold.

ISSN: 0976-3104

In **[Figure–6]**, based on the forecast threshold, the number of data aggregation is larger when compared to the Kalman filter. Due to this data aggregation is reduced redundant transmission and communication consumption power.

CONCLUTION

The energy efficiency is the important key Wireless sensor networks. With data transmission is the major part of energy consumption, chaos theory based time series prediction method to enhance energy efficiency. The proposed Chaos Theory based Data Aggregation (CTAg) based approach reduces redundant data, communication overhead and number of packet transmission between aggregator and sink node by using adaptive thresholds. The time series prediction using CTAg method was energy efficient and performed less computation to obtain the forecasted data. The experiments also show CTAg achieves better performance compared to other prediction approaches like Kalman Filter [KF] based prediction.

.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Akyildiz IF, Su W, Sankarasubramaniam Y, Cyirci E. [2002] Wireless sensor networks: A survey. *IEEE Communications Magazine*: *August*.

[2] Yick J, Mukherjee B, Ghosal D. [2008] Wireless sensor network survey. *The International Journal of Computer and Telecommunications Networking*. dl.acm.org/citation.cfm?id=1389832

[3] Heinzelman WR, Chandrakasan A, Balakrishnan H. [2000] Energy-Efficient Communication Protocol for Wireless Microsensor Networks. *Proceedings of the 33rd Hawaii International Conference on System Sciences*

[4] Krishnamachari B, Estrn D, Wicker S. [2002] The impact of data aggregation in wireless sensor networks. *Proceedings of International conference on Distributed Computing Systems Workshops (ICDCSW)*.

[5] Rajagopalan R, Varshney PK. [2006] Data-aggregation techniques in sensor networks survey. *IEEE Communications Surveys*.

[6] Yang G, Zheng J, Shi JH, Chen H. [2009] Energy Balance Hierarchical Data Aggregation Mechanism for Wireless Sensor Network. *WASE International Conference on Information Engineering*.

[7] Guiyi Wei, Yun Ling, Binfeng Guo, Bin Xiao, Athanasios V. [2011] Vasilakos Prediction-based data aggregation in wireless sensor networks: Combining grey model and Kalman Filter. *Computer communication* 34: 793–802.

[8] Hyuntea Kim, Jaebok Park, Giwhan Cho. [2007] Statistical Data Aggregation Protocol based on Data Correlation in Wireless Sensor Networks. *International Symposium on Information Technology Convergence*.

[9] Li G, Wang Y. [2011] An Efficient Data Aggregation Scheme Leveraging Time Series Prediction in Wireless Sensor Networks. *International Journal of Machine Learning and Computing*, www.ijmlc.org/papers/55-A836.pdf

[10] Rajesh G, Vinayagasundram B, Saravana Moorthy G. [2014] Data Fusion in wireless sensor network using Simpson's 3/8 rule, *International Conference on Recent Trends in Information Technology*.

[11] Liu L, Yang H, Lai M. [2005] Electricity price forecasting model based on chaos theory. *Power Engineering Conference (IPEC), The 7th International Conference on*,

[12] Liu H, Huang D, Wang Y. [2011] Chaotic Dynamics Analysis and Forecast of Stock Time Series. *International Symposium on Computer Science and Society* (ISCCS),

[13] Ci S, Sharif H. [2012] Performance comparison of kalman filter based approaches for energy efficiency in wireless sensor networks, *In proceedings of the 3rd ACS/IEEE International Conference on computer systems and Applications* (AICCSA'05), *IEEE, Cairo, Egypt, 58–65*.

[14] Giona M, Cimagalli V, Morgavi G, Perroiie A. [1990] Local prediction of chaotic time series. *Proceedings of the 33rd Midwest Symposium on Circuits and Systems*.

[15] http://db.csail.mit.edu/labdata/labdata.html.

[16] http://www.abarim-publications.com/ChaosTheoryIntroduction.html#.VS94WdyUeNM

[17] http://en.wikipedia.org/wiki/Chaos_theory

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Arumugam and Sangaiah

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

DETECTION OF CARDIAC ARRHYTHMIA FROM ECG SIGNALS

Maheswari Arumugam[1*] and Arun Kumar Sangaiah[2]

[1]*Department of ECE, Sambhram Institute of Technology, Bangalore 560 097, INDIA*
[2]*School of Computing Science and Engineering, VIT University, Vellore-632014, INDIA*

ABSTRACT

The first and foremost step in the analysis of ECG for various cardiac disease is the extraction of ECG and classification of ECG based on feature extraction. The prediction of any cardiac disease will come correctly if the feature extraction and classification of ECG are correct. It is very difficult for any expert physicians to predict exactly the cardiac arrhythmia as the volume of data is huge. The monitoring of ECG signals for a long period of time is needed as the changes in ECG may occur at any fraction of time. The computer aided approach helps significantly for monitoring ECG signals for long period of time. Many methods like wavelet transform, Independent Component analysis (ICA), Permanent Component analysis (PCA), Time Domain, Wavelet Transform, Power Spectral Density and Fuzzy logic with Neural network techniques are used either separately or in combination by the Researchers for extracting features from ECG signal. It is found from this survey that among many classifiers, most of the research works uses SVM classifier as it gives high classification accuracy. It is also found from this survey that most of research work uses the ECG data from MIT-BIH database or from Physionet. The parameters which determines the suitable technique for feature extraction and classification are found to be classification accuracy, sensitivity and specificity. This paper presents a survey of various approaches used in the feature extraction and classification of ECG signals.

Received on: 30th-March-2015
Revised on: 01st-May-2015
Accepted on: 20th– May-2015
Published on: 8th–Aug-2015

KEY WORDS

Feature Extraction; Classification; cardiac arrhythmia; classification accuracy; testing data; training data

***Corresponding author: Email:** maheswari.a@gmail.com, **Tel.:** +91 8884953957; **Fax:** + 91 80 23641701

INTRODUCTION

The main tool used in clinical practice to record the electrical activities of heart is ECG. The amplitude and interval features of ECG contains useful information for analysis. The exact identification of the heart disease depends on choosing the correct approach for extracting the features in ECG signals. ECG recording contains volumes of data which is very difficult for manual analysis by the physician. The computer aided analysis of ECG data has drawn the attention of everyone. Many researches are focused in this direction. Hanlin et.al [1] used wavelet transformation for feature extraction from 12 lead ECG signal. They used semi- supervised approach for clustering unlabeled data. Finally the SVM classifier was involved for ECG classification. Ge Dingfei [2] concentrates on the feature extraction of a cardiac abnormal condition called premature ventricular contraction (PVC) and normal sinus rhythm (NSR) and discusses in detail the discrimination between them by conducting data analysis on the data selected from MIT-BIH database. A new method known as Random Projection [7] is used for feature extraction on multi lead ECG signal which mainly concentrates on dimensionality reduction. This method is energy efficient and it is feasible to implement in portable systems for health monitoring. The structure of this paper is as follows: Section 2 discusses the preprocessing methods used to make ECG signal ready for feature extraction along with the techniques for feature extraction by various authors. Section 3 is a discussion on performance of several methods used by the researchers to achieve their goal. Section 4 reflects the research gaps in each of these research works after making a comparon of these techniques with respect to classification accuracy and data size graphically.

RESOURCES AND METHODS

Data preprocessing methods

The ECG data contains unwanted data which is due to muscle contractions, power line interference and baseline drift. The unwanted signal must be removed to get the original ECG data.

THE IIOAB JOURNAL
ISSN: 0976-3104

In paper [1], the preprocessing phase comprises of elimination of noise, detection of R peaks and the segmentation of heartbeats. Here Pan Tompkins algorithm is used to find the R peaks. The research work [2] uses the data from MIT-BIH database. The baseline drift and power line interference are removed from ECG signals by a band pass filter with lower frequency pass band of 1Hz and upper frequency pass band of 35 Hz. The ECG data [3] is preprocessed to remove baseline wander and filtered with a band pass filter to remove high and low frequency artifacts. The images for five cardiac conditions namely Normal, Ventricular Flutter, Ventricular Tachycardia, Nodal and Left Bundle Branch Block were collected from physionet database in the work [4]. Totally 100 images were collected from the database. The preprocessing of ECG includes binarization, 1D signal extraction and base line drift removal with median filter. Two kinds of data is used for feature extraction and classification of ECG signals in paper [5]. In this paper the feature points positioning is also done in the preprocessing stage. The data [6] is very large and it is taken from Physiobank archive. Nearly 1200 feature vectors for each class of disease is taken. Here three sets of features are taken for time domain, wavelet transform and for power spectral density. In the preprocessing stage [7] every record consists of 15 simultaneously measured signals. Each of the signals are digitized at 1000 samples / sec and has a 16 bit resolution. The database taken from 60 subjects includes 20 records for normal sinus rhythm, another 20 records for myocardial infarction and 20 cases of cardiomyopathy. Among these 30 subjects are used for training sets and the remaining as testing sets. At any time the selected beat has 700 samples. The analysis in [8] is conducted on ECG signals from MIT-BIH Arrhythmia Database on three kinds of signals namely, the normal ECG, Premature Ventricular Contraction (PVC) and Left Bundle Branch Block (LBBB). The preprocessing phase [9] is applied to whole of the chosen database and the principle characteristics are extracted using Principal Component Analysis (PCA) technique. The tests are conducted on 4 databases randomly generated which comes around 4002 samples. Out of this, 2802 samples of each database is used for training set and the remaining 1200 samples are used as testing set. The raw ECG signal [10] is filtered in frequency domain to remove dc offset. Then fast Fourier transform (FFT) is applied to get low frequency offset removed and baseline drift at zero reference line. Again inverse FFT is applied to get preprocessed signal.

Methods of feature extraction

In paper [1], a most popular time-frequency transformation called discrete wavelet transform technique is used to map the 12 lead ECG segments into the WT space. The new method introduced in this paper called Semi-supervised discriminant analysis distinctly distinguishes the features for various diseases in the ECG segment. A new beat detection algorithm which does not affect beat shape is used in the research work [2]. This algorithm uses two feature sets of data. The first feature set includes energy of wavelet coefficients and RR ratio for the consecutive beats. The second feature set are selected by pre-selection of the coefficients and further by forward selection process. Further a 6-level discrete wavelet transform decomposition is performed on each beat using a 10th order Daubechies. The morphological features are extracted using wavelet transform and independent component analysis together [3]. The dynamic features are obtained from RR interval information. The wavelet analysis is performed with Daubechies wavelet of order 8. FastICA algorithm is used here to extract 18 ICA coefficients for every heartbeat. The dimensionality of feature is reduced to 26 by principal component analysis and this results in 9932% variance. The statistical and morphological feature extraction is done in paper [4] through Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) techniques.

The technique employed for extracting the features from ECG in the work [5] is Independent Component analysis (ICA). The authors' claim that by taking 10 independent components for each segment of multi lead data more effective results can be obtained. The feature extraction [6] uses the combination of time domain, wavelet domain and power spectral density features. Finally three divergent feature sets consisting of 8 time domain, 32 wavelet transform and 11 power spectral density features are obtained. Another important point here is that all these features are extracted without using QRS detection. The Random Projection [7] technique is used for generating feature vector for all records in the database. This technique reduces the dimensionality of a data set but preserves the geometrical structure. Accordingly Random Projection has reduced the dimensionality of beat from to 50 samples from 700.

In [8] Kernel Principal Component Analysis (KPCA) is used for extracting the non-linearly related structures to the input space through the solution of an eigenvalue problem in a feature space of higher dimension. Two approaches are used here. The first one uses the combination of Principal Component Analysis (PCA) and modified fuzzy one-against-one (MFOAO) for categorizing multiclass signals. The second approach uses PCA, unbalanced clustering (UC) and fuzzy one-against-one (FOAO) algorithms. The UC algorithm is used for discarding the outliers. The feature extraction [9] is done with Principal Component Analysis (PCA) technique and it is combined with unbalanced clustering algorithm (UC) to reduce the complexity in feature extraction. In [10] Sub-band analysis is used to divide the signal into various frequency band components. Every beat in every kind of ECG consists of 250 data points after employing a 6 level decomposition on ECG signal. Three kinds of sub band energy calculations namely normal, mean and Relative mean are performed on the extracted energy coefficients of ECG.

ECG classification methods

In paper [1], the data in the semi supervised discriminant analysis feature space are classifies using support vector machine (SVM). Here the dataset is split into two halves randomly for training and testing. The training set includes 14301 heartbeats and the testing set has 14300 heartbeats. The dataset includes data for six diseases such as Normal, Sinus bradycardia, Left bundle branch beat, Electrical axis left side, Right bundle branch block and Left ventricular hypertrophy.

The ECG classification technique used in paper [2] is Support Vector Machine [SVM]. Here SVM separates all the samples into 2 classes by an optimal hyperplane. The total data in this work are divided into 10 equal size folds. Among these, 9 folds are used for training data. The last fold is trained with 10 classifiers. Finally the ultimate classification performance is determined by averaging the results from all the 10 test sets. The classification tool used in paper [3] is SVM classifier with a Gaussian radial basis function kernel. Based on the training data the SVM classifier is trained. The trained SVM classifier is used to evaluate the data. The implementation of SVM algorithms is through LIBSVM package. The classification of ECG in paper [4] is done using Euclidian and Manhalanobis classifiers. Here 100 segments are taken for training and 50 segments for test samples. These 50 segments are obtained from the extracted 198 segments.

THE IIOAB JOURNAL
ISSN: 0976-3104

The classification of beats through experiments based on multi-lead data in [5] shows that 11 types of arrhythmia beats can be classified. This work also includes the classification of experimental data from 500 persons. Here the total number of beats collected is 6366 and the abnormal beats in this is 2611. The classification accuracy here is 90.47%, sensitivity is 0.9001 and specificity is 0.9074. The authors indicate that this method can be applied to practical environment. This work also indicates that the training data proportion also affect the result.

The classification algorithms [6] used are decision tree and neural network. In this paper a fivefold cross validation technique is used for the evaluation of the entire ECG database. The experiments also indicate that the classification with decision tree algorithm is better than with neural networks. The classification performance success score, called F- score [6] with entire 51-dimensional features of time domain, wavelet transform and power spectral density feature extractions is 0.947 and it is the highest.

A Neuro-Fuzzy classifier [7] is used for feature classification in ECG signals. This method of classification results in 100% recognition rate for 25 Random Projection coefficient. The method used for classification in [8] is Multi SVM classification on 10 normal signals, 10 Premature Ventricular Contraction (PVC) signals and 14 Left Bundle Branch Block (LBBB) signals. The accuracy of separation of normal beats to abnormal beats is 100% while it is 70% in PVC and 55.71% in LBBB with building of separate SVMs for each pair of classes. In [9] a study is conducted to differentiate five types of heart beat namely normal beat, left bundle branch beat, right bundle branch beat, premature ventricular contraction beat and atrial premature contraction beat. The fuzzy support vector machine (FSVM) is used here for the binary classification of ECG signals. The principal component analysis (PCA), unbalanced clustering (UC) and Fuzzy one-against-one (FOAO) combination is used here for the classification of long term ECG records. The classification of the extracted features [10] is done by using Artificial Neural Network (ANN). The feed forward network uses 1 input layer, 1 hidden layer and 1 output layer. The Multilayer Perceptron (MLP) uses backpropagation algorithm to find the error signal.

RESULTS AND DISCUSSIONS

Discussion on performance

In paper [1], the semi-supervised discriminant analysis performance accuracy rate is 97.64%. When compared with other algorithms. The SVM classifier used for Multi-category classification here uses the cross validation of optimal parameter of SVM for better classification performance. The classification accuracy depends on the features and classifier used. In paper [2] for 770 ECG records, the overall accuracy of first feature set is 88.07% and the overall accuracy for the second feature set is 77.13%. The overall classification accuracy with both feature sets combined is 93.17%. This work shows that ECG features with lower dimensions can be extracted which improves the classification results.

The feature extraction method used in paper [3] separates 15 types of heartbeats depending on ICA features, features in wavelet and RR interval. The average classification accuracy here is 99.66%. The authors in paper [4] claim that the best classifier found for their work is Manhalanobis Distance Classifier which has a sensitivity (Se) of 96% and specificity (Sp) of 92%. In multi-lead ECG classification system based on improved ICA and SVM [5], 10 independent components from each segment are used for extracting ECG features and this is found to give more effective results. The classifier used here is SVM classifier and it classified 11 types of arrhythmia beats. The accuracy of the technique in comparison to other works is 98.18%. The feature sets [6] of time domain, wavelet transform and power spectral density are compared for processing time on a computer with Intel core i7 processor, 2.8 GHZ processor with 4 GB RAM and it is found that time domain features has less processing time with medium classification performance. In ECG analysis the preprocessing step which is critical is QRS detection. The ECG signals are classified successfully without QRS detection.

The computational complexity of Random Projection based feature extraction [7] technique is low. This helps to implement this technique in real time on a wireless wearable sensor platform. The combined accuracy of KPCA and SVM in normal beats [8] is 100% for normal beats while the accuracy is 70% in Premature Ventricular Contraction (PVC) and 85.71% in Left Bundle Branch Block (LBBB). The analysis shows that Kernel Principal Component Analysis (KPCA) with One against all (OAA) is superior to deal with Multi SVM classification.

When used with a large database, the Modified Fuzzy One –Against-One [9] consumes more time but gives good result. In the research work [10] 5 types of arrhythmias are taken for feature extraction and classification. They are normal beat (N), premature ventricular contraction (PVC), paced beat, right bundle branch block (RBBB) and left bundle branch block (LBBB). A total of 25 heartbeats are used for testing which includes 20 samples for each type of arrhythmia. Nearly 100 heartbeats are taken for training. The classification result shows that Normal subband energy (NSE) gives a high classification accuracy of 100% for normal beats, Mean subband energy (MSE) gives the highest classification accuracy of 91.6% for PVC and Relative mean subband energy (RMSE) gives a higher classification accuracy of 95.8% for normal beats, LBBB and RBBB.

THE IIOAB JOURNAL
ISSN: 0976-3104

Evaluation of various feature extraction and classification methods

This section of the paper gives an assessment of various methods used for feature abstraction and classification of bio signals. A clear view of several techniques used by various authors for the assessment of bio signals is depicted in **Table– 1.** The data from MIT-BIH database is used most commonly by many researchers.The comparison table also indicates that SVM classifier is popularly used. The review study also shows that the size of the data set has a definite relation with the method used for feature extraction and classification of bio signals.

The metrics used for comparison of works of various researchers are classification accuracy, sensitivity and specificity. In certain [6] works the processing time for feature extraction and classification of bio signals are also considered. The study indicates that as the size of the data set increases, the complexity also increases and the process becomes tedious. Therefore an intelligent combination of different extraction methods and classifiers is the solution in this scenario.

Table: I. Comparison of various feature extraction and classification methods

Author and year	ECG data collection	Feature extraction method	Classification method	Efficient features	Performance metrics
Hanlin Zhang, 2013	12 lead ECG data	Wavelet transform	SVM classifier	High classification accuracy, good generalization ability	Total classification accuracy = 97.64%
Ge Dingfei, 2012	MIT-BIH database	Wavelet transform	SVM classifier	possible and feasible to extract ECG features with lower dimensions from wavelet coefficients improvement in the classification results	193 ECG records considered and heartbeat classification accuracy is 88.07%
Can Ye, 2010	MIT-BIH Arrhythmias database	Wavelet Transform (WT) Independent Component Analysis (ICA)	SVM classifier	Single-lead performance is superior than any other previous works carried so far	Heartbeat classification accuracy is 99.91% with 84630 records out of 84707 ECG records correctly classified, 15 classes of heartbeats are classified final arrhythmias detection accuracy is 99.93%
Rizwan R. Sheikh, 2009	Physionet	PCA LDA	Two classifiers namely, Euclidean and Manhalanobis used	Classification of cardiac segments based on statistical and morphological features extracted from ECG	Best classifier found is Manhalanobis Distance Classifier with sensitivity Se=96% and specificity Sp=92%
Mi Shen, 2010	MIT-BIH Arrhythmia Database and 2500 practical data gathered from 500 persons	An improved Independent Component Analysis (ICA)	SVM classifier used for multi classification	Classification accuracy compared to other works is 98.18%	For multi classification average accuracy of testing data is 98.18% and the average Sensitivity is 98.68% For practical data 2-classification Experiment accuracy of testing data is 90.47% and the Sensitivity is 90.01%
Serkan Gunal, 2013	Physiobank archive	Time domain (TD), Wavelet transform (WT) and power spectral density (PSD)	Decision tree, Neural network	High classification performance if TD, WT and PSD combined in a large database Less processing time for TD with medium classification performance	For combined TD+WT+PSD features F-score = 94.7% for decision tree F-score = 90.7% for Neural network

THE IIOAB JOURNAL
ISSN: 0976-3104

Iva Bogdanova, 2012	Physikalisch-Technische Bundesanstalt (PTB) Diagnostic ECG database	Random Projection	Neuro-Fuzzy classifier	Feasible and energy efficient Can be implemented wireless sensor platforms due to its low complexity	30 subjects ECG data constitutes training set and 30 subjects ECG data constitutes testing set Jaccard index is 90% for 20 coefficients
Maya Kallas, 2012	MIT-BIH Arrhythmia Database	Kernel Principal Component Analysis (KPCA)	Two multi-SVM classification schemes used are 1.One-Against-One (OAO) 2. One-Against-All (OAA)	Higher average classification accuracy	Very high average classification accuracy of 97.39% for OAA with KPCA
Mohamed cherif Nait-Hamoud, 2010	Real database from MIT-BIH arrhythmia database of ECG arrhythmias	Principal Component Analysis (PCA)	Fuzzy Support Vector Machine (FSVM)	The complexity of classification is reduced with Unbalanced Clustering algorithm	The average classification accuracy when PCA+UC+FOAO algorithms used and R=2 is 97.895%
Pratiksha Sarma, 2014	MIT-BIH arrhythmia database	Wavelet subband energy based	Multilayer Perceptron Neural network	Performance optimization of classifier by using statistical properties of subband energy features	Normal subband energy classification accuracy = 90.78% Mean subband energy classification accuracy = 84.12% Relative mean subband energy classification accuracy = 91.62%

The relation between the size of the dataset in terms of heartbeats and the classification accuracy is shown separately in **Table– 2**. This table indicates that as the total number of samples increases, the complexity increases for feature extraction and classification. A suitable method has to be carefully selected in this case. The comparison also shows that wavelet transform for feature extraction and SVM classifier for classification gives better results.

Table: 2. Relation between data size and classification accuracy

Ref. No.	Size of dataset	Classification accuracy
[1]	Training set data size – 14301 Testing set data size – 14300	97.54%
[2]	770 classes	88.07%
[3]	85300 beats	99.25%
[4]	Training – 100 segments Testing – 50 segments	Se=96% and specificity Sp=92%
[5]	Training data – 7016 heartbeats Testing data – 91.509 heartbeats	90.74% Se=90.019%, Sp=90.74%
[6]	1200 features per class	F-score = 94.7% for decision tree F-score = 90.7% for Neural network
[7]	Training data – 30 real subjects ECG Testing data – 30 real subjects ECG	Jaccard index is 90% for 20 coefficients
[9]	Training data – 2802 samples Testing data – 1200 samples	97.895%
[10]	Training data – 25 heartbeats Testing data – 100 heartbeats	Average classification accuracy = 88.84%

THE IIOAB JOURNAL
ISSN: 0976-3104

DISCUSSION

The graph in **figure– 1** gives an idea of which technique to use when the data size is small and big. It is found that good results can be obtained by choosing Independent Component Analysis technique for feature extraction from ECG signals and SVM classifier method for classification of ECG signals.

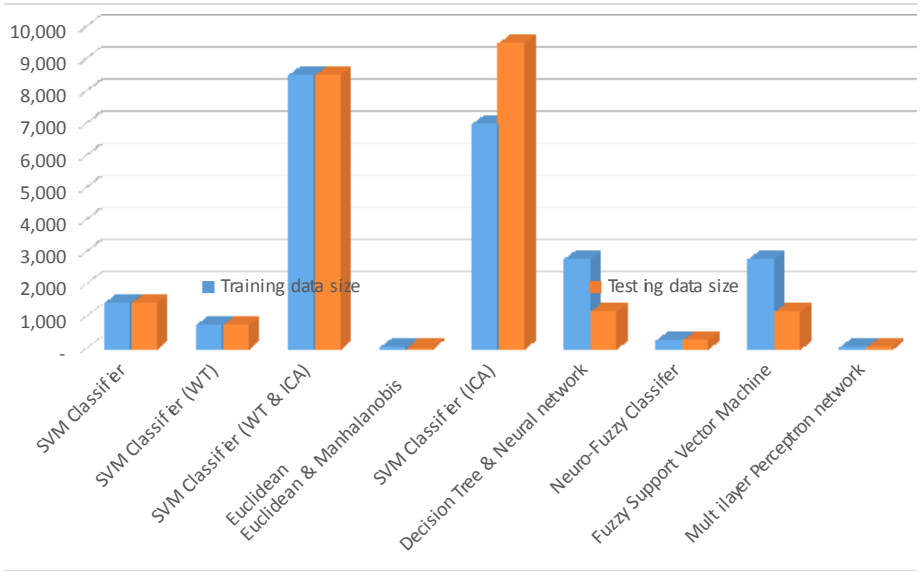


Fig: 1. Dependency between data size and techniques used

The success of the selected method for classification of ECG signals depends on the classification accuracy. This is depicted graphically in **figure–2**. It is found that among many classifiers, SVM classifier in combination with Fuzzy network gives good results.

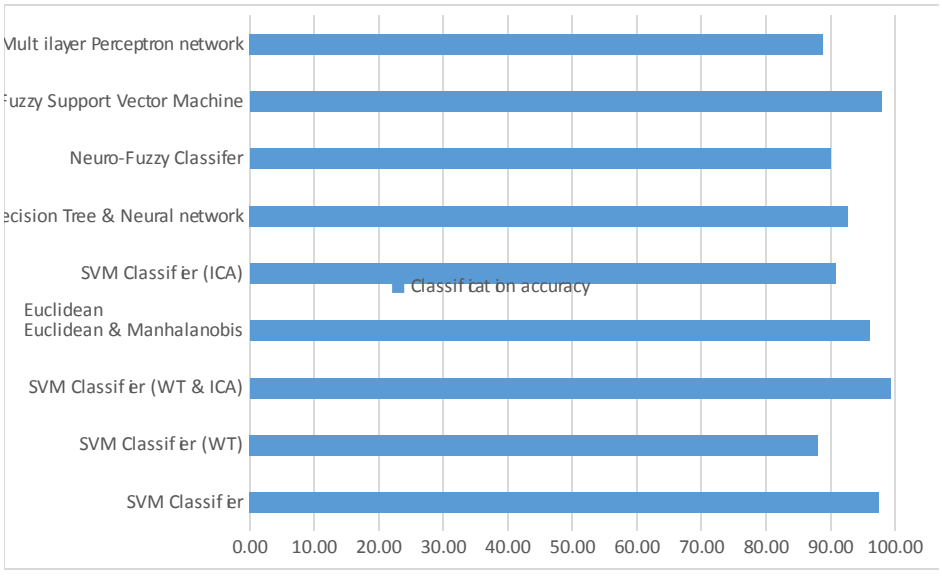


Fig: 2. Classification techniques and Classification accuracy

THE IIOAB JOURNAL
ISSN: 0976-3104

Research gaps

The technique of semi supervised learning ability [1] gives good result only for small training set and when the labeled data is insufficient. Experiment and analysis have shown that better classification results can be obtained for lower dimension features that that of features with higher dimensions. The study on feature extraction [2] includes only wavelet coefficients with reduced dimension of 34 only. The work in paper [4] is useful only for off-line ECG signal analysis. The authors indicate that their work is helpful for medical students for learning ECG off-line analysis. The research gap in this work is that this is not suitable for online ECG analysis on real time bio signals. The results in [5] shows that the structure of the training data should be in proportion to the overall data to get better results. When the amount of data increases, the computational complexity also increases. Efficient methods of feature selection is mandatory here to get effective results. Focus on fusion of classifiers is necessary as there exists a relationship between the classifiers and the diseases they classify. The evaluation of feature sets [6] and classification is done on four types of heart conditions namely, normal, congestive heart failure, ventricular tachyarrhythmia and atrial fibrillation. No mention is made about the application of this method to other kinds of heart diseases. Also, this technique is applicable for off-line ECG analysis only. For a bigger data set [7] it is necessary to examine the average recognition rate for several executions of the proposed hypothesis. Here it is concluded that due to low computational complexity the Random projection concept is applicable for wireless sensor platforms. But how exactly this technique can be implemented is not highlighted. The research experiments [8] are conducted on MIT-BIH arrhythmia database on only three kinds of cardiac diseases. There is no mention about the application of this technique in real time ECG signal analysis. The results [10] shows that for a particular kind of cardiac disease only a particular type of wavelet subband method must be applied to get high classification accuracy. The decision on which type of wavelet subband energy method is applicable for what kind of cardiac disease comes out after conducting many experiments on trial and error basis.

CONCLUSION

This paper gives a brief review on various methods applicable for feature extraction and classification of bio signals used by different researchers. The research works of various researchers are studied and compared on the basis of techniques used for feature extraction and classification and the resulting accuracy. It is noted that many research works uses wavelet transform for feature extraction and SVM classifier for classification of bio signals. Another interesting feature noted is that the selection of data size has a great impact on the selection of technique to be used for feature extraction and classification of bio signals. Further the study indicates that many research works are being carried out in this area as a feasible solution suitable for clinical application is not achieved so far.

ACKNOWLEDGEMENT

The author is grateful to Dr. Arun Kumar Sangaiah, Associate Professor, School of Computer Science and Engineering, VIT University, Vellore whose constant support has helped to bring out this work successfully.

CONFLICT OF INTERESTS

The authors Maheswari Arumugam and Dr. Arun Kumar Sangaiah of The IIOAB Journal are the Research Scholar and Research Supervisor of VIT University, Vellore. The terms of this research work is reviewed and approved by the VIT University in accordance with its policy on objectivity in research.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Hanlin Zhang, Kai Huang, Dong Li and Liqing Zhang [2013]A 12-lead Clinical ECG Classification Method Based On Semi-supervised Discriminant Analysis, 2013 6th International Conference on Biomedical Engineering and Informatics (BMEI 2013).

[2] Ge Dingfei. [2012] Study of ECG Feature Extraction for Automatic Classification Based on Wavelet Transform, The 7th International Conference on Computer Science & Education (ICCSE 2012) July 14-17, 2012. *Melbourne, Australia*, 500–503.

[3] Can Ye, Miguel Tavares Coimbra. [2010] Arrhythmia Detection and Classification using Morphological and Dynamic Features of ECG Signals, 32nd Annual International Conference of the *IEEE* EMBS Buenos Aires, Argentina, August 31 - September 4, 2010,1918 – 1921.

[4] Rizwan R Sheikh, Imtiaz A Taj. [2009] Cardiac Disorder Diagnosis Based on ECG Segments Analysis and Classification, 978-1-4244-4361-1/09 ©2009 *IEEE*

[5] Mi Shen, Liping Wang, Kanjie Zhu, Jiangchao Zhu. [2010] Multi-lead ECG Classification based on Independent Component Analysis and Support Vector Machine, 2010 3rd International Conference on

THE IIOAB JOURNAL
ISSN: 0976-3104

Biomedical Engineering and Informatics (BMEI 2010) 960 – 964.

[6] Serkan Gunal, Semih Ergin, Efnan Sora Gunal, and Alper Kursat Uysal. [2013] ECG Classification using Ensemble of Features, 978-1-4673-5239-0/13 *IEEE*.

[7] Iva Bogdanova, Francisco Rincon and David Atienza. [2012] A Multi Lead ECG Classification based on Random Projection Features, 978-1-4673-0046-9/12 ©2012 *IEEE*, ICASSP 2012, 625 – 628.

[8] Maya Kallas, Clovis Francis, Lara Kanaan, Dalia Merheb, Paul Honeine and Hassan Amoud. [2012] Multi-Class SVM Classification Combined with Kernel PCA Feature Extraction of ECG Signals, 19th International Conference on Telecommunications 2012 (ICT 2012), 978-1-4673-0747-5/12, *IEEE*.

[9] Mohamed cherif Nait-Hamoud, Abdelouaheb Moussaoui. [2010] Two Novel Methods for Multiclass ECG Arrhythmias Classification Based on PCA, Fuzzy Support Vector Machine and Unbalanced Clustering, 978-1-4244-8611-3/10 ©2010 *IEEE*, 140 –145.

[10] Pratiksha Sarma, SR Nirmala, Kandarpa Kumar Sarma. [2014] ECG Classification using Wavelet Subband Energy based Features, 2014 International Conference on Signal Processing and Integrated Networks (SPIN), 978-1-4799-2866-8/14 *IEEE*, 785 – 790.

ABOUT AUTHORS

Prof. A. Maheswari *was born in Coimbatore, TamilNadu, India. She received her B.E degree from Government College of Technology and M.S in Embedded System Technologies from Anna University, Chennai. She is currently pursuing PhD from Vellore Institute of Technology, Vellore, India. Presently she is working as an Assistant Professor in the Department of Electronics and Communication Engineering, Sambhram Institute of Technology, Bangalore, India. Her research interests include Medical Sciences, broad area of Embedded Systems and Wireless Technologies. Moreover she is very much interested in applying the technology to uplift the life of common man. She is a life member of Indian Society of Technical Education.*

Dr. S. Arun Kumar *received Master of Engineering (M.E) Degree in Computer Science and Engineering from Government College of Engineering, Tirunelveli, Anna University Chennai. He had received Doctor of Philosophy (Ph.D.) Degree in Computer Science and Engineering from VIT University, Vellore, Tamil Nadu, India. He is presently working as Associate Professor in School of Computing Science and Engineering, VIT University, India. His areas of interest include Software Engineering, Soft Computing, Wireless Networks, Bio-Informatics, and Embedded Systems. He has author of more than 50 publications in different journals and conference of National and International repute. His teaching areas include Software Engineering, Wireless Networks, Data and Computer Communications, Soft Computing, Programming Languages, and etc. His current research work include Global Software Development, Wireless Ad hoc and Sensor Networks, Machine learning, Cognitive Networks and Advances in Mobile Computing and Communications. He is active member for Compute Society of India. He has guided many research students and post-graduate students in the field of communication networks, ad hoc networks, database, and soft computing techniques.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Anusha et al.

THE IIOAB JOURNAL

REVIEW ARTICLE OPEN ACCESS

A NEW MCDM APPROACH INTEGRATING QFD, DEMATEL WITH TOPSIS FOR EXPLORING THE EFFECT OF SOCIAL NETWORK USAGE ON ACADEMIC PERFORMANCE

Challa Anusha[1], Jesna James Parappilly[1*], Arun Kumar Sangaiah[2]

[1]*School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, INDIA*
[2]*School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, INDIA*

ABSTRACT

The rapid advances in internet applications Social Networking Sites (SNS) have become worldwide used internet applications and play a prominent role in student academic performance. In order to determine the evaluation criteria of SNS usage from the perspective of student academic performance, we have integrated multi-criteria decision making (MCDM) approaches: (a) Quality Function Deployment (QFD), (b) Decision Making Trial and Evaluation Laboratory (DEMATEL) and (c) Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) under fuzzy environment is proposed. In this method, the personality, cultural and technology criteria for determining SNS usage on academic performance is established. These three kinds of criteria are assessed from the student and the faculty perspective respectively. Students give their subjective responses about the importance of SNS usage and rating the alternatives with respective to academic performance. Similarly, faculty members give their subjective preferences about the relationship between personality, cultural and technology criteria and the correlation between SNS usage on academic performance of students. Further, the hybridization of QFD, DEMATEL, and TOPSIS has not been available in the literature. Based on this context, we have combined QFD, DEMATEL, and TOPSIS approach for evaluating SNS usage on academic performance from the perspective of university students in India. The proposed approach has been tested in a real case study among VIT university students in India. A numerical illustration for SNS usage on academic performance is also given to demonstrate the application of hybrid MCDM approach.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 8th –Aug-2015

KEY WORDS

Social Networking Sites (SNS); Academic Performance Criteria (AC); Fuzzy Multi-Criteria Decision Making (FMCDM); Quality Function Deployment (QFD); Decision Making Trial and Evaluation Laboratory (DEMATEL); Technique for Order Performance by Similarity to Ideal Solution (TOPSIS

***Corresponding author: Email:** jesnajames.parappilly2011@vit.ac.in, **Tel.:** +40-9001010010; **Fax:** +40-9001010012

INTRODUCTION

Social Networking Sites (SNS) includes Facebook and Twitter is gaining popularity and widely used applications among Asian university students. In India, SNS sites are currently utilized by 120.5 million of users on (2014) the demand is estimated at 224.2 million of social network users in 2018, up from 63.1 million in 2012 [19]. Moreover, an earlier study revealed that more than 80% of university students have used Facebook, which is a vital element in university social culture [10]. The earlier studies [1, 3, 4, 6-8, 11-13, 16-18, 20, 21] have addressed the relationship between academic performance and SNS usage through various contexts such as personality (big-five personality factors), technology context (ICT tools) and cultural context (personal importance of SNS and motives of SNS usage). Consistent with the earlier research works, this study has integrated the SNS usage and student academic performance relationship in three contexts: *personality context* (big-five personality factors), *technology context* (ICT tools), and *cultural context* (personal importance of SNS and motives of SNS usage). However, a comprehensive approach for measuring personality, technology, and cultural factors for evaluating SNS usage with relate to student academic performance under fuzzy environment has not been adequately available in the literature. Subsequently, fuzzy multi-criteria decision making for integrating QFD, DEMATEL, and TOPSIS for validating SNS usage on academic performance was not reported in the literature. Thus, to address this research gaps motivated us to develop a combined approach based on QFD-DEMATEL-TOPSIS under fuzzy environment is presented in this study.

THE IIOAB JOURNAL
ISSN: 0976-3104

To address these research gaps, an empirical study has been carried out in VIT University, India to evaluate SNS usage effectiveness from the perspective of student academic performance. The rest of this paper is organized as follows: Section 2 presents the literature reviews on SNS criteria and it measurements have been used in this study. Sections 3 present the QFD-DEMATEL-TOPSIS approach and assessment framework used in this research. Section 4 and 5 presents the empirical study results and a discussion of the study respectively. A conclusion of the study is presented finally to address the significance of QFD-DEMATEL-TOPSIS to address the SNS effectiveness.

LITERATURE OF PAST RESEARCH WORKS

This section presents the earlier researches and constructs an assessment methodology that is used in this study.

Personality, cultural, and technology context for SNS usage on academic performance

Earlier research studies have addressed the personality context through big-five personality factors (extraversion, neuroticism, agreeableness, openness to experience, and conscientiousness) to measuring SNS usage on academic performance [7, 11-13,16-18]. Consistent with previous research works, this study has addressed personal characteristics from the perspective of SNS effectiveness on student academic performance through big-five indictors. In technology context, earlier studies [4, 14, 15] have investigated the use of ICT in SNS and its positive outcome in educational settings. Further, the research relating SNS and integrate the use of ICT tools that support for SNS has been addressed by few studies. Besides, in technology context the focus of this study has revealed the SNS and integrates use of ICT towards academic performance. In cultural context, many of the researchers [6, 8, 15, 17, 20] have addressed the cultural difference and the motivations behind the usage of SNS towards the academic performance. Based on this context, the cultural aspects have been evaluated for SNS effectiveness on student academic performance addressed in this study via two dimensions: personal importance of SNS and motives for SNS usage. The source of measurements and possible evaluation criteria's of SNS effectiveness on academic performance are collectively determined through earlier literature [1, 3, 4, 6-8, 11-18, 20, 21, 27] as follows:

Table: 1. Summary of source of measurements and possible evaluation criteria's of SNS effectiveness on academic performance

Personality Context*(Extraversion, Neuroticism, Agreeableness, Openness to new Experience, Conscientiousness)*
Does extraversion motivate you to form new connection that improves your academic performance through SNS?
Do you think extraverts perform better in academics than introverts?
Does being an extrovert help you in gaining more knowledge through SNS?
An extrovert joins more online groups to gain knowledge?
Does Neurotic people find SNS environment to be less anxiety provoking than everyday classroom studies?
Does SNS helps introvert and neurotics students to reach their full potential regarding interaction with their peers and professors?
Does SNS helps neurotic student to share their idea more freely on SNS?(Yes/No)
Does SNS helps in collecting more people who agrees for same ideology in studies?
To what extent you have found users which engage and endorse interpersonal cooperation easily in studies?
Do you find agreeable individual more cooperative, trusting, and helpful?
Do you think SNS to be innovative in field of education?
Do you think SNS to be better means of education than others?
Are you willing to accept SNS as an educational tool?
Does less conscientiousness student spent more time using SNS and achieve better results in education?
Does conscientious individuals are likely to be high achievers as they have a strong work ethic?
Cultural Context *(Personal Importance of SNS and motives for SNS use)*
Do you find any information regarding your career or academic interests on social networking sites?
Do you think social networking sites can be an effective tool for e-learning?
Do you think social networking sites are more effective in communicating with your teachers than in actual class?
Do you think social media sites improve your academic performance?
Do you prefer to express your ideas and feelings on social networking sites?
Technology Context *(ICT tools)*
Does student feel more comfortable in adapting new ICT technology?
Do you find ICT tools helpful for easily access to SNS?
Academic Performance
Are Reflective learning styles include synthesis-analysis and elaborative processing more effective than agented learning styles include methodical study and fact retention?
Do reflective learning styles (synthesis-analysis and elaborative processing) facilitate deeper understanding?
Do you think openness is most beneficial to learning when students adopt reflective learning styles?
Do you think both personality traits and learning styles are influencing with academic achievement?

THE IIOAB JOURNAL ISSN: 0976-3104

The panel of experts has validated the questionnaires and finalized these factors as evaluation criteria to validate SNS usage effectiveness on student academic performance under this study as shown in **Table– 1**.

Table: 2. Linguistic Terms

Linguistic Variable	Corresponding TFN Multiplicative/Fuzzy	Crisp Value
Strongly agree	(3,4,5)/(0.9,1.0,1.0)	5/1.0
Agree	(2,3,4)/(0.7,0.9,1.0)	4/0.9
Neither Disagree Nor Agree	(1,2,3)/(0.3,0.5,0.7)	3/0.7
Disagree	(0,1,2)/(0.0,0.1,0.3)	2/0.3
Strongly Disagree	(0,0,1)/(0.0,0.0,0.1)	1/0.1

QFD, DEMATEL and TOPSIS

Currently, there is an interest to use fuzzy QFD or House of Quality (HOQ) in multi-criteria decision making approaches. In this paper SNS usage on student academic performance has been categorized in to QFD matrixes (often HOQ) has been applied to determine the importance of parameters. Similarly, the earlier studies [5, 9] have been given a combined methodology of group decision making based on fuzzy linguistic variables for QFD applications. Based on this context, this study has employed QFD and MCDM approaches for measuring SNS usage on academic performance of university students.

Recently, fuzzy DEMATEL approach has been used for evaluation of attributes, interrelationship among the criteria and especially dealing with human uncertainty and subjective vagueness within the decision making process by the use of fuzzy set theory. In the recent studies [22, 24, 25] DEMATEL approach has been investigated in different areas of application in the context of MCDM problems. Likewise this research focuses on DEMATEL approach on designing hybrid methodology for the real data set obtained from VIT University for evaluating SNS usage effectiveness on student academic performance is presented.

TOPSIS, one of the conventional MCDM methods, has been widely used to compute the relative importance of alternatives and solving practical decision making problems with its high computational efficiency and comprehensibility. Moreover, current studies have adopted TOPSIS to solve MCDM problems [2, 24, 26]. Similarly, the basic idea of using TOPSIS in this paper is to compute the ideal solution (best values realistic of criteria) and negative ideal solution (worst values realistic of criteria) for ranking the SNS usage on academic performance factors perceived by university students.

To the best of our knowledge, up to date research on evaluation of SNS usage on academic performance case study under fuzzy environment is very limited. Moreover, assessment framework for the integration of personality, cultural and technology factors for the effectiveness of SNS usage on academic performance has not been adequately presented in the available literature. Further, the hybridization of QFD, fuzzy DEMATEL, and TOPSIS has addressed only in very few studies. Based on this context we have integrated QFD-DEMATEL-TOPSIS approach for evaluating SNS usage on academic performance from the perspective of university students in India.

FRAMEWORK FOR EVALUATING SNS USAGE ON ACADEMIC PERFORMANCE IN FUZZY ENVIRONMENT

In this study, the QFD approach has been integrated with the fuzzy DEMATEL and TOPSIS approach for the evaluation of SNS usage effectiveness on students' academic performance from the perspective of personality, cultural and technology contexts under a fuzzy environment is proposed.

The proposed QFD-DEMATEL-TOPSIS methodologies for SNS usage on academic performance evaluation framework consists of three parts. First, we have used QFD approach for determine the importance of parameters in the respective contexts. Second, we have applied the DEMATEL approach for determining the weights of the SNS criteria. Finally, we have used fuzzy TOPSIS to identify the rank and significance of the SNS attributes from the perspective of student academic performance.

The construction of proposed framework and computation procedure of hybridization of QFD-DEMATEL-TOPSIS approach under a fuzzy environment is depicted in **Figure –1**.

THE IIOAB JOURNAL
ISSN: 0976-3104

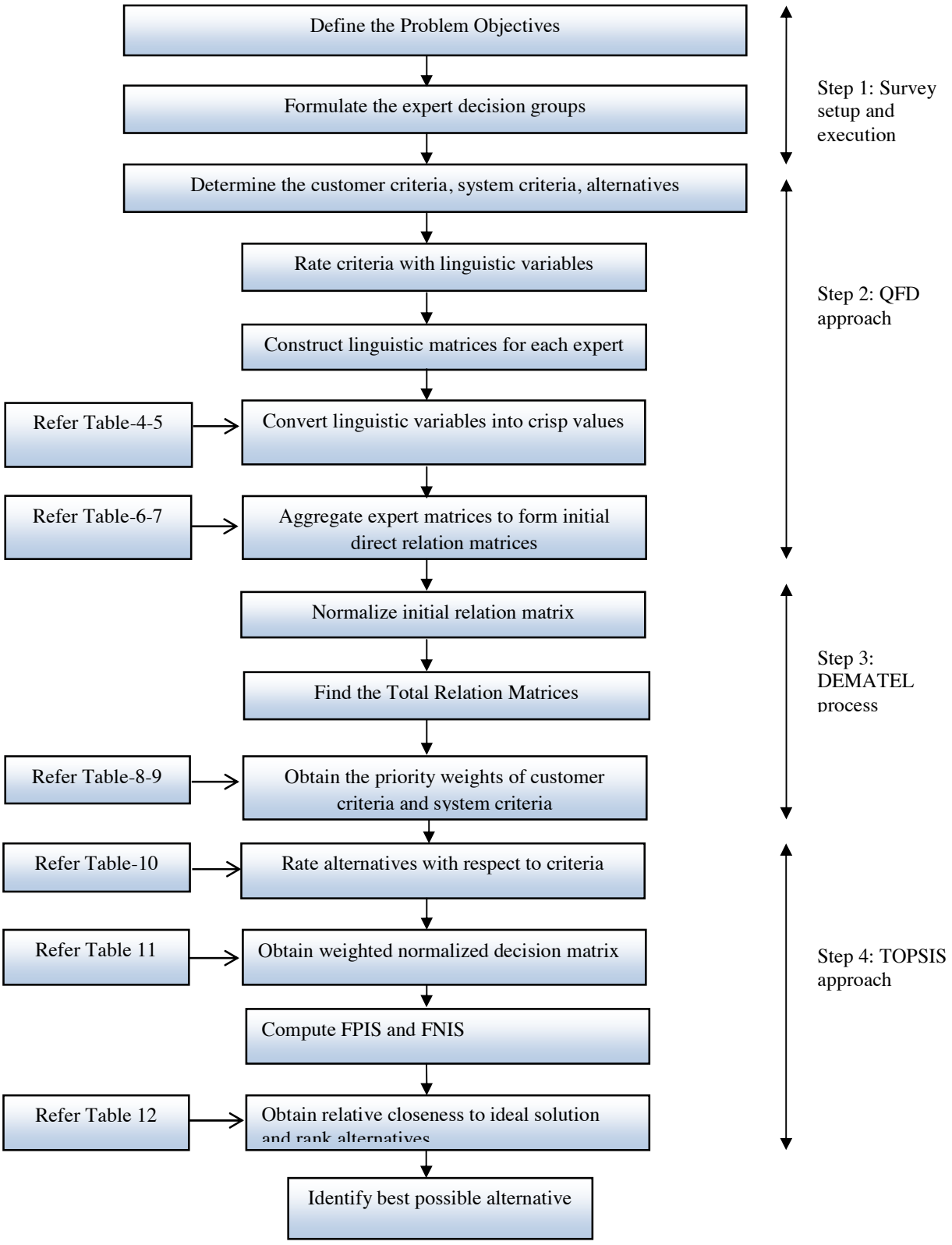


Fig: 1. Proposed technique for evaluating SNS effectiveness on academic performance

THE IIOAB JOURNAL
ISSN: 0976-3104

EMPIRICAL CASE STUDY FOR EVALUATING THE SNS EFFECTIVENESS ON ACADEMIC PERFORMANCE

The main focus of this study is to determine the SNS usage effectiveness at individual, cultural and technology levels in academic performance research phenomenon of university students. To achieve this goal, an empirical study has been carried out in VIT University students located in India. The demographic details of the respondents of this study were undergraduate science, engineering students, and faculty members of VIT University. Two groups of student, S1 and S2, and two groups' faculty, F1 and F2 were approached for representation of their fuzzy preferences. Consequently, the empirical study has been tested among 40 DMs' (20 students and 20 faculties) of this organization to validate the effectiveness of SNS usage on academic performance. According to earlier study [2], we have followed fuzzy preference and multiplicative preference relations for DM's judgments over set of alternatives/criteria. The hybrid fuzzy QFD-DEMATEL-TOPSIS approaches were applied in this case study, as illustrated in the following sections.

Quality function deployment application

The basic steps of QFD approach used in this study are as follows:

Step 1: Identify the alternatives, student personality criteria, and academic performance criteria. The identified criteria are as follows:

Table: 1. Criteria for evaluation of SNS usage effectiveness

Criteria	Code	Description
		Personality criteria
Extraversion	**SC1**	Effectiveness of extraversion on SNS usage on academic performance
Neuroticism	**SC2**	Effectiveness of neuroticism on SNS usage on academic performance
Conscientiousness	**SC3**	Direct and positive effect of conscientiousness on SNS usage on academic performance
Agreeableness	**SC4**	Positive effect and agreeableness on SNS usage on academic performance
Openness to experience	**SC5**	Previous and openness to experience on SNS usage on academic performance
Academic Performance criteria		
ICT Tools	**AC1**	Effectiveness of ICT tools usage of SNS on academic performance
Personal Importance of SNS	**AC2**	Effectiveness of personal importance on SNS usage on academic performance
Motives of SNS Usage	**AC3**	Direct positive effect and motives of SNS usage on academic performance

Step 2: Construct the relational matrices and linguistic weight matrices based on the DM's ratings. The matrices, after converting to crisp values have been represented **[Table–4 and -5]**

Step 3: The rating matrices are aggregated to determine corresponding weights.

Fuzzy DEMATEL application

Step1: The initial direct-relation matrices are obtained from the QFD application and are represented **[Table–6 and -7]**.

Step2: The initial direct relation matrices D1 and D2 are normalized by equations (1)-(2) to form normalized matrices N1 and N2.

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

$$N = \frac{1}{m} D \quad (2)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

Step3: The total relation matrices R1 and R2 are computed by the following equation.

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

Step4: Weights of customer criteria and system criteria are computed using the equations (4)-(6)

$$r_i = \sum_{1 \le j \le n} R_{ij} \quad (4)$$

$$c_j = \sum_{1 \le i \le n} R_{ij} \quad (5)$$

$$W_j = \sum_{j=1}^{n}(r_i + c_i) / \sum_{i=1}^{n}\sum_{j=1}^{n}(r_i + c_i) \quad (6)$$

The priority weights for student criteria (SC) and academic performance criteria (AC) are tabulated in **Table–8 and Table–9** respectively.

Fuzzy TOPSIS Application

The basic steps of Fuzzy-TOPSIS approach used in this study are as follows:
Step 1: Construct fuzzy assessment decision matrix, determine the alternatives, and normalize the scores in order to find the best alternative **[Table–10].**

Step 2: Input the weights which is obtained from the DEMATEL method to calculate the weighted normalized decision matrix as given in **Table–11.**

Step 3: The best evaluation and worst evaluation value with respect to each criterion is determined through FPIS and FNIS.

Step 4: Obtain relative closeness coefficient to the ideal solution and rank the alternatives **[Table–12].**

RESULTS

The numerical results of the empirical study have been illustrated as follows:

Table: 2(a). Relation between student criteria rated by Student S1

Criteria	SC1	SC2	SC3	SC4	SC5
SC1	0.5	0.65	0.6	0.4	0.65
SC2	0.35	0.5	0.65	0.9	0.65
SC3	0.4	0.35	0.5	0.15	0.15
SC4	0.6	0.1	0.85	0.5	0.65
SC5	0.35	0.65	0.85	0.35	0.5

Table: 3(b). Relation between student criteria rated by student S2

Criteria	SC1	SC2	SC3	SC4	SC5
SC1	0.5	0.55	0.3	0.4	0.65
SC2	0.45	0.5	0.3	0.9	0.35
SC3	0.7	0.7	0.5	0.15	0.65
SC4	0.6	0.1	0.85	0.5	0.65
SC5	0.35	0.65	0.35	0.35	0.5

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 4 (c). Relation between student criteria rated by faculty F1

Criteria	SC1	SC2	SC3	SC4	SC5
SC1	0.5	0.4	0.2	0.9	0.7
SC2	0.2	0.5	0.6	0.9	0.7
SC3	0.3	0.1	0.5	0.8	0.5
SC4	0.8	0.2	0.7	0.5	0.3
SC5	0.2	0.1	0.2	0.6	0.5

Table: 4 (d). Relation between student criteria rated by faculty F2

Criteria	SC1	SC2	SC3	SC4	SC5
SC1	0.5	0.8	0.4	0.1	0.7
SC2	0.2	0.5	0.6	0.9	0.7
SC3	0.7	0.8	0.5	0.9	0.5
SC4	0.8	0.2	0.7	0.5	0.6
SC5	0.2	0.2	0.2	0.6	0.5

Table: 4(a). Relationship between academic performance criteria rated by student S1

Criteria	AC1	AC2	AC3
AC1	0.5	0.3	0.6
AC2	0.7	0.5	0.35
AC3	0.4	0.65	0.5

Table: 5(b). Relationship between academic performance criteria rated by student S2

Criteria	AC1	AC2	AC3
AC1	0.5	0.3	0.6
AC2	0.7	0.5	0.35
AC3	0.4	0.65	0.5

Table: 5(c). Relationship between academic performance criteria rated by faculty F1

Criteria	AC1	AC2	AC3
AC1	0.5	0.5	0.3
AC2	0.3	0.5	0.4
AC3	0.4	0.5	0.5

Table: 5(d): Relationship between academic performance criteria rated by faculty F1

Criteria	AC1	AC2	AC3
AC1	0.5	0.5	0.3
AC2	0.3	0.5	0.4
AC3	0.4	0.5	0.5

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 6. Aggregated relationship between student criteria

Criteria	SC1	SC2	SC3	SC4	SC5
SC1	0.5	0.6	0.375	0.45	0.675
SC2	0.3125	0.4625	0.5375	0.9	0.6
SC3	0.525	0.5	0.5	0.5	0.45
SC4	0.7	0.65	0.775	0.5	0.55
SC5	0.275	0.4	0.4	0.475	0.5

Table: 7. Aggregated relationship between academic performance criteria

Criteria	AC1	AC2	AC3
AC1	0.5	0.4	0.45
AC2	0.5	0.5	0.375
AC3	0.4	0.575	0.5

Table: 8. Priority weights of student criteria

Criteria	Weight
SC1	0.3780
SC2	0.4219
SC3	0.3774
SC4	0.4451
SC5	0.3775

Table: 9. Priority weights of academic performance criteria

Criteria	Weight
AC1	0.667
AC2	0.661
AC3	0.673

Table: 10. Fuzzy rating of alternative with respect to criteria

	SC1	SC2	SC3	SC4	SC5	AC1	AC2	AC3
A1	(0.2,0.3,0.7,0.8)	(0.1,0.4,0.7,0.8)	(0.3,0.7,0.8,.9)	(0.1,0.3,0.7,0.5)	(0.5,0.6,0.8,1)	(0.4,0.5,0.8,1)	(0.3,0.7,0.8,1)	(0.2,0.5,0.8,1)
A2	(0.1,0.4,0.7,0.9)	(0.3,0.7,0.8,.9)	(0.2,0.5,0.8,1)	(0.1,0.5,0.8,0.9)	(0.2,0.5,0.8,1)	(0.2,0.3,0.7,0.8)	(0.5,0.6,0.8,1)	(0.2,0.3,0.7,0.9)
A3	(0.3,0.5,0.7,0.8)	(0.2,0.5,0.8,1)	(0.5,0.6,0.8,1)	(0.3,0.7,0.8,.9)	(0.2,0.3,0.7,0.8)	(0.1,0.5,0.8,0.9)	(0.2,0.3,0.5,0.9)	(0.5,0.6,0.8,1)
A4	(0.4,0.7,0.8,.9)	(0.5,0.6,0.8,1)	(0.2,0.6,0.7,0.8)	(0.2,0.6,0.7,0.8)	(0.5,0.6,0.8,1)	(0.3,0.6,0.8,1)	(0.2,0.5,0.8,1)	(0.5,0.6,0.8,1)
A5	(0.1,0.5,0.8,0.9)	(0.3,0.8,0.7,1)	(0.1,0.4,0.7,0.9)	(0.1,0.5,0.8,0.9)	(0.3,0.8,0.7,1)	(0.2,0.5,0.8,1)	(0.5,0.6,0.8,1)	(0.3,0.8,0.7,1)

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 11. Weighted normalized matrix

	Benefit criteria					Cost criteria		
	SC1	SC2	SC3	SC4	SC5	AC1	AC2	AC3
A1	0.1465	0.1455	0.1807	0.1402	0.1849	0.3286	0.3148	0.3070
A2	0.1538	0.1964	0.1673	0.2015	0.1594	0.2434	0.3260	0.2579
A3	0.1685	0.1819	0.1941	0.2366	0.1275	0.2799	0.2136	0.3561
A4	0.2051	0.2110	0.1539	0.2015	0.1849	0.3286	0.2810	0.3561
A5	0.1685	0.2037	0.1405	0.2015	0.1785	0.3043	0.3260	0.1965

Table: 12. Ranking of alternatives

Alternative	Relative closeness to ideal solution	Rank
A1	0.72	1
A2	0.48	4
A3	0.51	3
A4	0.61	2
A5	0.44	5

DISCUSSION

The integrated QFD-DEMATEL-TOPSIS methodology has been used for investigation of SNS usage on the academic performance of students from Asian countries and especially in VIT University, India. The data used in this study were collected from faculty and students of VIT University to explore the SNS usage on academic performance through survey questionnaires. The faculty and students have given their subjective judgments based on multiplicative preference/fuzzy preference relations as shown **Table– 3**. Totally 4 (5 alternatives) DMs' samples are represented in this study to explore the SNS effectiveness criteria using linguistic assessments on fuzzy preference/multiplicative preference relation. In addition, **Tables 6 and 7** depicts the aggregation of DMs ranking of each alternative with respect to criteria on Table-3 using fuzzy linguistic items as shown in **Table–2**. Subsequently, the relative importance of the criteria and output of QFD modeling has been represented in **Tables–8 and 9**. The weights of DEMATEL results address that Agreeableness (SC4) and Motives of SNS usage (AC3) are more significant that those other evaluation factors. Consequently, the QFD-DEMATEL modeling results have been applied in the TOPSIS method and their results are given in **Tables– 10-12**. From the above results, it can be concluded that alternative A1 is closest to the ideal solution while alternative A5 is farthest from it. Thereby, SNS usage on university students can facilitate the effectiveness of student learning performance in academic organizations.

CONCLUSION

The main objective of this paper is to provide an approach to evaluate SNS effectiveness on from the student academic performance perspective. In order to do that, a new MCDM approach combining QFD-DEMATEL-TOPSIS has been proposed in the fuzzy environment. Moreover, this proposed approach has been investigated among VIT University students to explore the influence of SNS effectiveness in academic performance. Furthermore, to demonstrate the applicability and creditability of QFD-DEMATEL-TOPSIS approach, the framework has been validated based on the data collected from the VIT university students. Consequently, this study has presents two valuable contributions: (i) a comprehensive overview of the factors influencing SNS usage on academic performance (ii) QFD-DEMATEL-TOPSIS approach to find the relative importance of the criteria and to rank the criteria. In this study, we have suggested a research framework based on QFD-DEMATEL-TOPSIS which can effectively validate and rate the SNS evaluation criteria in the context of students' academic performance. Subsequently, the case study results address that personality, technology, and cultural context factors have a significant impact on the evaluation of the SNS effectiveness of student academic performance in the academic environment. The prototype of proposed approach (QFD-DEMATEL-TOPSIS) can be developed in the future it can be enhanced into an efficient tool to handle MCDM in a real time settings.

THE IIOAB JOURNAL
ISSN: 0976-3104

Through MCDM approach we have revealed that SNS and its influential factors are the main contributors to enhance student learning performance in an academic setting.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Chamorro-Premuzic T, Furnham A. [2003] Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*(4): 319–338.

[2] Fan ZP, Hu GF, Xiao SH. [2004] A method for multiple attribute decision-making with the fuzzy preference relation on alternatives. *Computers & Industrial Engineering*, *46*(2): 321–327.

[3] O'Connor MC, Paunonen SV. [2007] Big Five personality predictors of post-secondary academic performance. *Personality and Individual differences*, *43*(5): 971–990.

[4] Lockyer L, Patterson J. [2008] Integrating social networking technologies in education: a case study of a formal learning environment. In *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on* (pp. 529-533). IEEE.

[5] HT Liu. [2009] "The extension of fuzzy QFD: from product planning to part deployment," *Expert Systems with Applications*, 36 (8): 11131–11144.

[6] Vasalou A, Joinson AN, Courvoisier D. [2010] Cultural differences, experience with social networks and the nature of "true commitment" in Facebook. *International journal of human-computer studies*, *68*(10): 719–728.

[7] Ryan T, Xenos S. [2011] Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, *27*(5): 1658–1664.

[8] Kim Y, Sohn D, Choi SM. [2011] Cultural difference in motivations for using social network sites: A comparative study of American and Korean college students. *Computers in Human Behavior* 27(1): 365–372.

[9] LZ Lin, LC Huang, HR Yeh. [2011] "Fuzzy group decision making for service innovations in quality function deployment," *Group Decision and Negotiatio*, 21(4):495–517.

[10] Thompson SH, Lougheed E. [2012] Frazzled by Facebook? an exploratory study of gender differences in social network communication among undergraduate men and women. *Coll Stud J* 46 (1): 88–99.

[11] Wang JL, Jackson LA, Zhang DJ, Su ZQ. [2012] The relationships among the Big Five Personality factors, self-esteem, narcissism, and sensation-seeking to Chinese University students' uses of social networking sites (SNSs). *Computers in Human Behavior*, *28*(6): 2313–2319.

[12] Chen B, Marcus J. [2012] Students' self-presentation on Facebook: An examination of personality and self-construal factors. *Computers in Human Behavior*, *28*(6): 2091–2099.

[13] Hughes DJ, Rowe M, Batey M, Lee A. [2012] A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2): 561–569.

[14] Lei CU, Krilavicius T, Zhang N, Wan K, Man KL. [2012] Using Web 2.0 tools to enhance learning in higher education: A case study in technological education. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2).

[15] Jackson LA, Wang JL. [2013] Cultural differences in social networking site use: A comparative study of China and the United States. *Computers in human behavior*, *29*(3): 910–921.

[16] Mitrofana N, Iona A. [2013] Predictors of Academic Performance. The Relation between the Big Five Factors and Academic Performance. *Procedia-Social and Behavioral Sciences*, *78:* 125–129.

[17] Karpinski AC, Kirschner PA, Ozer I, Mellott JA, Ochwo P. [2013] An exploration of social networking site use, multitasking, and academic performance among United States and European university students. *Computers in Human Behavior*, *29*(3): 1182–1192.

[18] Vedel A. [2014] The Big Five and tertiary academic performance: A systematic review and meta-analysis. *Personality and Individual Differences*, *71*:66–76.

[19] Statista – The statistics Portal [2014\ India: number of social network user 2012-2018. <http://www.statista.com/statistics/278407/number-of-social-network-users-in-india/>

[20] Ozer I, Karpinski AC, Kirschner PA. [2014] A Cross-cultural Qualitative Examination of Social-networking Sites and Academic Performance. *Procedia-Social and Behavioral Sciences*, *112*: 873–881.

[21] Michikyan M, Subrahmanyam K, Dennis J. [2014] Can you tell who I am? Neuroticism, extraversion, and online self-presentation among young adults. *Computers in Human Behavior*, *33*: 179–183.

[22] Li Y, Hu Y, Zhang X, Deng Y, Mahadevan S. [2014] An evidential DEMATEL method to identify critical success factors in emergency management. *Applied Soft Computing*, *22*:504–510.

[23] Shukla RK, Garg, D, Agarwal A. [2014] An integrated approach of Fuzzy AHP and Fuzzy TOPSIS in modeling supply chain coordination. *Production & Manufacturing Research*, *2*(1): 415–437.

[24] Sangaiah AK, Subramaniam PR, Zheng X. [2015] A combined fuzzy DEMATEL and fuzzy TOPSIS approach

for evaluating GSD project outcome factors. Neural Computing and Applications, *Springer Publishers*, Article in Press, DOI: 10.1007/s00521-014-1771-1.

[25] Altuntas S, Dereli T. [2015] A novel approach based on DEMATEL method and patent citation analysis for prioritizing a portfolio of investment projects. *Expert Systems with Applications*, *42*(3): 1003–1012.

[26] Sang X, Liu X, Qin J. [2015] An analytical solution to fuzzy TOPSIS and its application in personnel selection for knowledge-intensive enterprise. *Applied Soft Computing*, *30*: 190–204.

[27] Sangaiah, A., & Thangavelu, A. (2013). An exploration of FMCDM approach for evaluating the outcome/success of GSD projects. open engineering, 3(3):419–435.

ABOUT AUTHORS

Prof. Jesna James P *is pursuing 4th MS Software Engineering in VIT University .Her areas of interest include Software Engineering, Soft Computing and Cloud Computing. She is currently working under the guidance of Prof. Arun Kumar S in the field of multi-criteria decision making problems.*

Dr. C. Anusha *is pursuing 4th MS Software Engineering in VIT University. Her areas of interest include Software Engineering, Soft Computing, Software Inspection and Cloud Computing. She is currently working under the guidance of Prof. Arun Kumar S in the field of multi-criteria decision making problems.*

Dr. S. Arun Kumar *received Master of Engineering (M.E) Degree in Computer Science and Engineering from Government College of Engineering, Tirunelveli, Anna University Chennai. He had received Doctor of Philosophy (Ph.D.) Degree in Computer Science and Engineering from VIT University, Vellore, Tamil Nadu, India. He is presently working as Associate Professor in School of Computing Science and Engineering, VIT University, India. His areas of interest include Software Engineering, Soft Computing, Wireless Networks, Bio-Informatics, and Embedded Systems. He has author of more than 50 publications in different journals and conference of National and International repute. His teaching areas include Software Engineering, Wireless Networks, Data and Computer Communications, Soft Computing, Programming Languages, and etc. His current research work includes Global Software Development, Wireless Ad hoc and Sensor Networks, Machine learning, Cognitive Networks and Advances in Mobile Computing and Communications. He is active member for Compute Society of India. He has guided many research students and post-graduate students in the field of communication networks, ad hoc networks, database, and soft computing techniques.*

.

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Gurumurthy and Tripathy

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

PERFORAMNCE ENHANCEMENT FOR AUTOMATED ANALYSIS IN HUMAN BRAIN SIGNAL PROCESSING TO FINDING ALZHEIMER'S SYNDROME USING INTELLIGENT TECHNIQUES

Sasikumar Gurumurthy* **and B. K. Tripathy**

School of Computing Science and Engineering, VIT University, Vellore, INDIA

ABSTRACT

Today efficient brain signal recognition is limited to find different brain diseases; using hardware like Electroencephalogram (EEG), Magnetoencephalograms (MEG), and Functional MRI (fMRI). Alternatively, abnormal brain waves have shown to be associated with particular brain disorders (e.g., Alzheimer's disease and epilepsy). But the problem with this approach is that there are not many algorithms that could efficiently extract signals from a brain to find Alzheimer's disease. In this research paper, we suggest the design of new algorithm which could do this job of translating brain signals to digital text data for Alzheimer's disease. We propose a new approach using which the problem of recognizing brain signals for Alzheimer's disease can be solved. We also provide the implementation details of this software. For the implementation of our idea, we propose a new intelligent technique Architecture. The technique adopted by the EEG signals used in brain for Alzheimer's disease, guided and stimulated us to design this research work. Details regarding the automatic preprocessing and decoding interpretation that would be essential and the shortcomings are also included in the research work.

Received on: 10th-March-2015
Revised on: 18th-May-2015
Accepted on: 26th–May-2015
Published on: 8th –Aug-2015

KEY WORDS

Electroencephalograph (EEG); Alzheimer; Back Propagation Network (BPN); Independent Component analysis (ICA; Non-Linear Energy Operator (NLEO); Discrete Wavelet Transform (DWT).

***Corresponding author:** **Email:** sasichief@gmail.com; **Tel.:** +91-0416-2202750; **Fax:** +91-0416-2243092

INTRODUCTION

In this paper we propose a novel method for the classification of memory disorders and Alzheimer. Memory disorders now a days gain much more attention in the research area because it is the basic symptoms of some major diseases such as cancer and psychiatric problems [1]. After pain, memory disorder is the second most indicator of illness. Memory loss is the basic symptoms of some major diseases such as cancer and Psychiatry. So memory disorder should be detected earlier and thereby give better treatments for the patients. Alzheimer is the abnormal activity of the brain. It is noticed that one in hundreds of the populations are the victims of this diseases. Sometimes Convolutions and unconscious are the symptoms of Alzheimer. All the above two diseases are diagnosed by the encephalograph (EEG).

The electrical fauna of the human nervous classification has been predictable for more than a period. It remains well recognized that the difference of the external potential circulation on the scalp replicates functional happenings emerging from the essential brain. This external potential difference can be verified by sticking a group of electrodes toward the scalp, and computing the electrical energy among couple of these electrodes, which remain before clarified, amplified, and verified. The resultant information is titled the EEG [13]. The EEG needs the lesser signal generosity in the range of microvolts (μV). EEG frequency band are normally classified into five categories. The meaning of these different frequencies is not completely known. They are alpha, beta, theta, delta and gamma. Alpha waves are rhythmic waves occurring at a frequency between 8 and 13 hertz. These waves are found in the normal persons when they are awake in a quiet, resting state [1]. When the subject is in memory loss, the alpha waves disappear completely. Beta waves normally occur in the frequency range of 14 to 30 hertz. These waves can be divided into two type's beta 1 waves and beta 2 waves. Beta 1 is elicited by the mental activity and the other is inhibited by it. Theta waves have frequencies between 4 and 7 hertz. These waves occur during emotional stress, particularly during periods of disappointment and frustration. Delta waves occurs only once every 2 or 3 seconds.

THE IIOAB JOURNAL
ISSN: 0976-3104

They occur in deep Alzheimer, in infants and in serious organic brain diseases. Gamma waves consist of low-amplitude, high-frequency waves resulting from attention or sensory stimulation. Architecture for implicated processes is shown in **Figure-1**.

All these waves have their own particular shapes in the signals and if any kind of diseases occur then the normal wave's shapes have changed. Theta and delta waves are normally used to diagnosis the memory disorders. Sometimes the signal shape of Alzheimer and one kind of memory loss disorder say narcolepsy may overlap. Visual inspection of EEG signals is very time consuming and very laborious work. So the EEG signal parameters extracted and analyzed using computers, are highly useful in diagnostics [11]. The information can be as several for example 128 channels then usually 20 channels remain used and copies characteristically last used for 30 to 60 minutes of time domain data, through a bandwidth among 0.1 to 150Hz, which remains demonstrated digitally arranged on a computer screen [15].

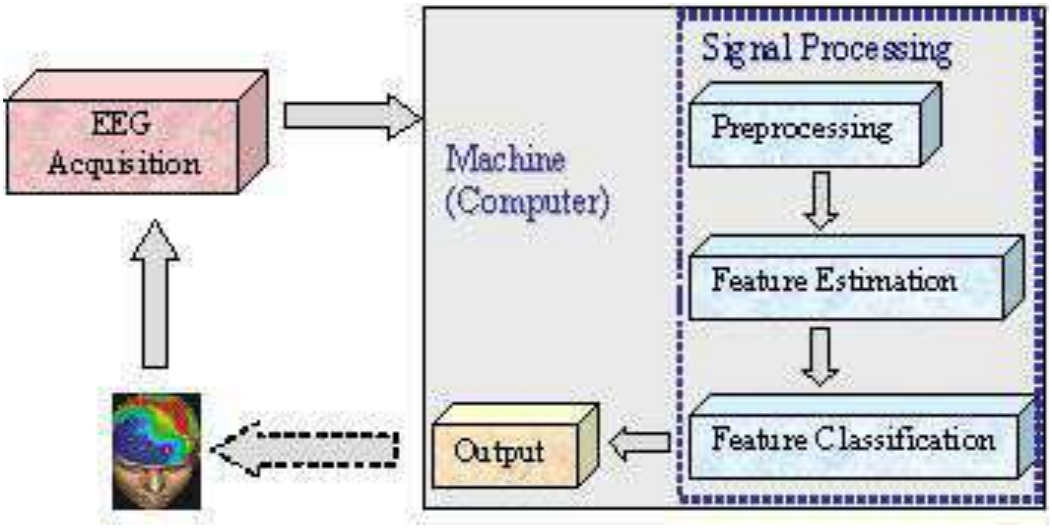


Fig: 1. Architecture for implicated processes

TECHNIQUES AND RESOURCES

The brain analysis and classification follows a specific method custom designed for the identification of brain diseases. **Figure –2** shows the block diagram of the proposed neural network [2] based automated brain signal analysis system.

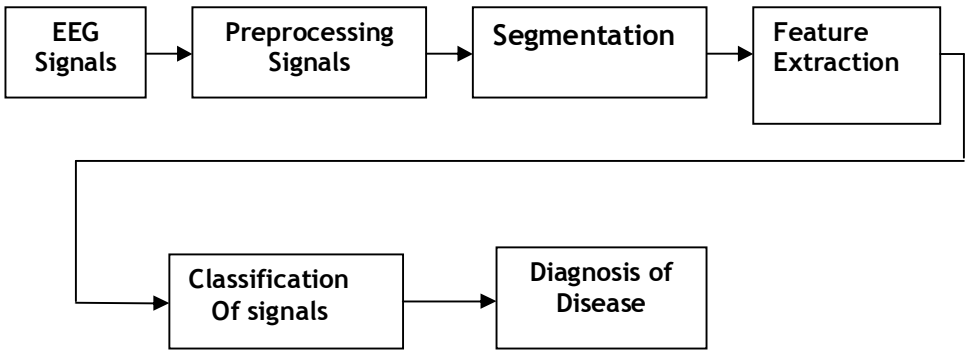


Fig: 2. Automated System design of the proposed system.

THE IIOAB JOURNAL
ISSN: 0976-3104

Signal attainment and EEG database

EEG acquisition collects these underlying electrical patterns from the scalp, and digitalizes them for computer storage. In our experiments, we have used the10-20 Classification of Electrode Location, which is constructed on the association among the position of an electrode besides the original part of cerebral cortex ["10" and "20" mention to the 10% or 20% interelectrode distance]. These signals include normal and abnormal signals of EEG. Divide them and name them as set A and set B. Set A contain normal signals. B contains abnormal signals such as memory disorders and Alzheimer signals. In the present study we classify them as two datasets.

Preprocessing

This Once the signals are acquired, it is necessary to clean them. Usually EEG signals are contaminated [5]. These contaminations may leads to misdiagnosis of diseases. So before analyzing the signals, it should be artifact free [6]. To remove the artifacts there are various methods. However, environments with large, low-frequency electromagnetic fields can cause a significant interference to EEG recordings. The second limitation is the activity of muscles in the head region (e.g. chewing, speaking, etc). The electrical activities of these muscles have, usually, much large amplitude then the EEG. With some effort in quality control and artifact processing the disadvantages can be compensated [14]. The best method for EEG signal preprocessing is Independent Component Analysis (ICA) [10]. ICA is a method for finding underlying factors or components from multivariate statistical data. **Figure-3** shows the ICA is a method for finding underlying factors or components from multivariate statistical data.

ICA outperforms the traditional principal component analysis (PCA) in numerous applications; in specific, it has been suitable in the place of removal of optical artifacts after the EEG, wherever principal PCA could not distinct eye artifacts from brain signals, specifically when they must comparable amplitudes [13]. Principal component analysis (PCA) is alternative technique that remains accessible, when associated with ICA for the selected application went short of effectiveness and efficiency [16].

Let s (t) is the signals generated. These signals are linearly combined through a memory less channel, mathematically described by the mixing matrix A. A has M rows, N columns, and xi (t) (I = 1 …M) are the signals observed at the channel output.

The mixing model is described by the following equation

$$X(t) = A.S(t) \quad \ldots\ldots (1)$$

Where s (t) and x (t) are the source and mixture signal vectors, respectively.

After mixing, the preprocessing of ICA involves the following steps. Centering, whitening and rotation. Centering make the signals centered in zero.
The main purpose of centering is to make the zero mean. Centering is achieved by simply subtracting the mean of signal from each reading of that signal.

$$X = x - E(x) \quad \ldots\ldots (2)$$

The next step is to whiten or sphere the data. This means that remove any correlations in the data [7]. Whitening is achieved by the eigen-value decomposition of the covariance matrix. Taking the covariance between every pair of signals can form a covariance matrix.

This matrix will be square and symmetric.

$$COV(X) = E(XXT)$$

Perform eigen-value decomposition on the covariance matrix and then transform the data so the covariance matrix of the transformed data is equal to the identity.

This procedure is called sphereing or whitening.

$$V = ED^{-1/2}E^{T} \quad \ldots\ldots (3)$$

(Eigen value decomposition of covariance matrix $E\{XX^{T}\} = EDE^{T}$ and D-1/2 = diag (d1-1/2, d2-1/ 2 … dn-1/2)]

Now rotation can be done by the inverse of whitening operation on the mixing matrix A.

$$S = \tilde{A}^{-1}Z \quad \ldots\ldots (4)$$

Where Z = VX.

The disadvantages of EEG are that the signal to noise ratio is poor; and it is necessary to deal with large subject-specific, inter- and intra-trial variability; hence, sophisticated data analysis has to be completed.

THE IIOAB JOURNAL
ISSN: 0976-3104

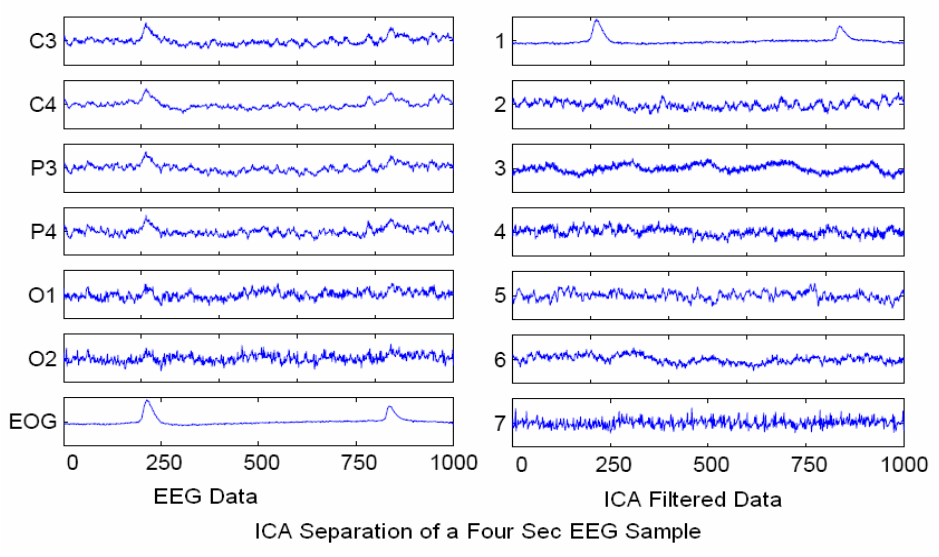


Fig: 3. ICA separation of a four sec EEG sample.

Segmentation

Novel way to examine non-stationary signals, it remains occasionally calmer to fragment signals hooked on pseudo-stationary sections. By means of adaptive segmentation algorithms [4], the non-stationary signals can be broken down into segments that are pseudo-stationary, and analyzed or processed separately. Signal segmentation is based on energy and frequency of signal and the signal has to be divided according to its characteristics. The main purpose segmentation is to find the edge and the energy of the signal. Among the various segmentation algorithm's NLEO is the best one.

A segmentation algorithm involving a non-linear energy operator (NLEO) is defined as follows.

$$E\{\ \Psi [\ x(n)\ ]\} = E [\ x(n-1)\ x(n-2) - x(n)\ \ x\ \ (n-3)\ ] \quad \ldots\ldots\ldots\ldots\ldots (5)$$

Where x (n) is the input EEG signal at current time n and $\Psi [\ (x (n)\ ] = x (n -1) x (n-2) -x (n) - x (n-3)$ is the nonlinear energy operator. Using NLEO, we can calculate the localized energy, and segment boundary. To detect a sudden change in the NLEO, a variable used for segmentation criteria is defined as follows.

$$Gnleo\ (n) = \sum_{m\ =\ n-N+1}^{n} \Psi\ (m) - \sum_{m\ =\ n+1}^{n+N} \Psi\ (m) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ (6)$$

Where 2N is the window size. Gnleo (n) reaches a peak when the Ψ (m) is discontinuous. The boundaries can be detected by the peaks of Gnleo (n). According to the windowing and thresholding equation the significant level can be obtained. The threshold equation can be defined as

$$T\ (n) = \max [\ Gnleo\ (n - L/2 : n+L/2\ )\ ],\ L \text{ window length})| \ldots\ldots\ldots (7)$$

Where Gnleo (n) is the energy change in the window obtained by using a moving window with length 2N at center n, and T (n) is the threshold value obtained by using another moving window with length L. G (n) represents the significant energy change after thresholding.

Feature extraction

After segmentation features should be extracted. Different features are suited for different diseases. The proposed system uses discrete wavelet transform (DWT) for memory disorders and spectral entropy for Alzheimer [8]. Spectral entropy quantifies the spectral complexity of the time series. Power spectral density is defined as

$$P\ (\omega) = I / N\ (|X\ (\omega)|) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where X (ω) represents fast Fourier transform of the signal. The normalization of equation (8) gives the spectral entropy and is defined as

$$H\ (\omega) = — \sum p\omega\ \ \log p\omega \quad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

Discrete wavelet transform (DWT) is suited to non-stationary signals and performs a multiresolution analysis of a signal. It builds on the concept of scales. The following features were found to be suitable for Alzheimer detection [3]. Zero crossing Extrema.

Neural network classifier

Artificial Neural Networks (ANN) is considered to be good classifier due to their inherent features such as adaptive learning, robustness, self-organization, and generalization capability [12]. ANN's are particularly useful for complex pattern recognition and classification tasks where enough data are available for training and where the simpler classification algorithms fail. In neural network [9], designing the architecture and network training are the main issues. If training is insufficient, then network will not learn properly. If excessive training of network is unable to generalize the training database. For each type of EEG signals, a corresponding output class is associated. In order to make neural network training [2] more efficient, the input features were normalized so that they fall in the range [0, 1]. Since the number of output class is 2, the ANN with one output is sufficient to produce a code for each class. **Figure-4** shows the co-efficient analysis.

The output are represented by

[0] = Normal memory states or normal states.
[1] = Abnormal that is memory disorders or Alzheimer seizure.

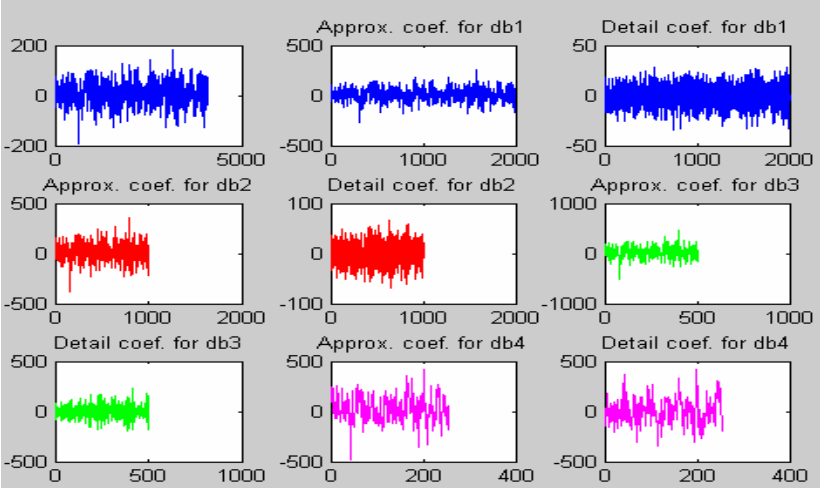


Fig: 4. Co-efficient analysis

The proposed system makes use of BPN. It is a multilayer, feed forward and supervised learning model. Back propagation remains the preeminent recognized training algorithm aimed at neural networks besides still one of the maximum suitable. It requires lesser memory necessities than utmost algorithms then frequently influences a suitable error level quite rapidly, even though it can be very slow to converge suitably on an error least possible. **Figure-5** shows the BP network.

.

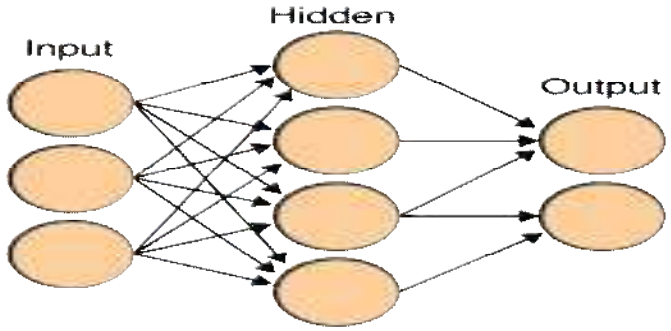


Fig: 5. Architecture of Back Propagation Network

THE IIOAB JOURNAL
ISSN: 0976-3104

Validation

The performance of BPN are evaluated by using the two parameters namely sensitivity (SE) and specificity (SP) which are defined as follows.

SE = TNcp/TNcp × 100… (10)

Where TNCP signifies the entire quantity of properly identified positive patterns and TNAP signifies the entire amount of actual positive patterns.

A positive pattern specifies a perceived seizure.

SP = TN_{CN} /TN_{CN} ×100……. (11)

Where TN_{CN} signifies the entire quantity of properly identified negative patterns and TN_{AN} signifies the entire quantity of real negative patterns.

A negative pattern specifies an identified non seizure. **Figure-6** shows the output screen shot of the final output.

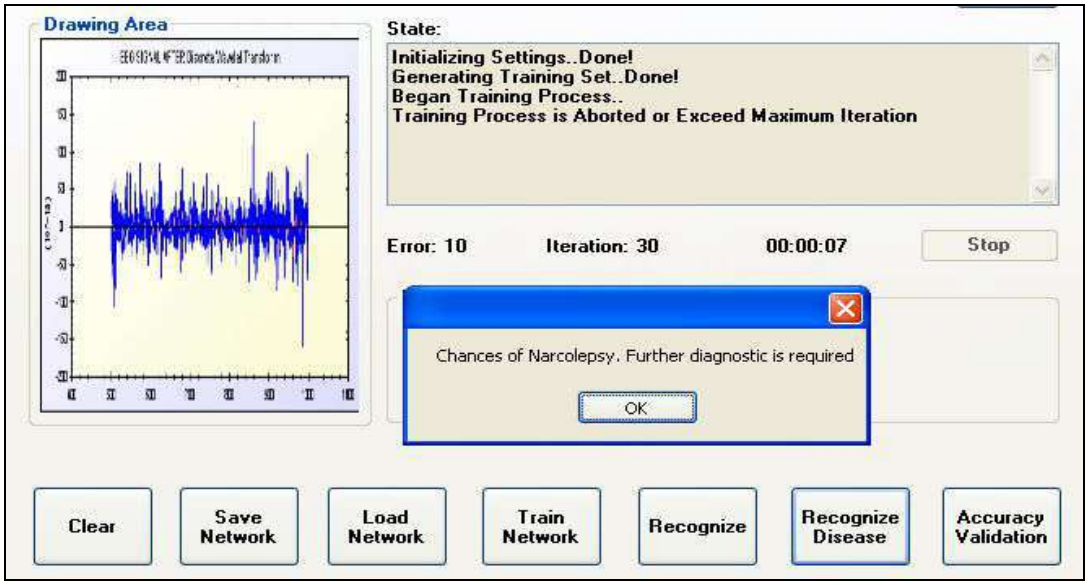


Fig: 6. Screen shot of the Final yield

CONCLUSION

The Work proposes an algorithm which will prove useful in biomedical signal processing where a specific underlying signal requires to be extracted from the possibly noisy multi-channel recordings. It is clear that the method is suitable for the extraction of independent components from the measured EEG. The algorithm worked efficiently in extracting memory spindle as well as Alzheimer seizures which were distributed throughout the measurement channels. We expect that this system will be valuable in the department of neurology and will help mankind. Current developments in computer hardware in addition to signal processing have made conceivable the consumption of EEG signals or "brain waves" for communication among humans and computers [13] and the same work can further be extended towards the domain also.

ACKNOWLEDGEMENT

The first author thanks to the School of Computing Science and Engineering Department of computer science, VIT University and Special thanks to Dean SCSE, for his kind guidance and support. This work has been (Partially) supported by the research program in SCSE, VIT University, India.

CONFLICT OF INTEREST

No conflict of interest

THE IIOAB JOURNAL
ISSN: 0976-3104

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Abdulhamit Subasi , M Kemal Kiymik, Ahmet Alkana, Etem Koklukaya. [2012] 'Neural Network Classification of EEG Signals by Using AR with MLE Preprocessing For Epileptic Seizure Detection' 57–70.

[2] Alois Schlög, Mel Slater, Gert Pfurtscheller. [2002] Presence research and EEG', Workshop on Presence, 2002 in collaboration with *The Institute of Biomedical Engineering, University of Technology Graz*. 1–7.

[3] CJ James, O Gibson. [2003] 'Electromagnetic Brain Signals analysis using Constrained ICA'. in the proceedings of Biomedical Engineering, *IEEE Transactions on* 50(9): 1108 – 1116

[4] DS Holder, RH Bayford, J Fritschy O Gilad, H Kaube, CD Binnie. [2003] " Development of generic software for analysis, archiving & Internet dissemination of brain and systems physiological data" , Fall 2003, James.N.Knight, "Signal fraction analysis and artifact removal in EEG". 1–13.

[5] EC Ifeachor, GT Henderson, C Goh, HSK Wimalaratna and N Hudson. [2005] 'Biopattern Analysis and Subject-Specific Diagnosis and Care of Dementia' Proceedings of the, "*IEEE Engineering in Medicine and Biology" 27th annual conference*. 2490–2493

[6] Hongzhi Qi, Baikun Wan, Li Zhao. [2004] 'Mutual Information entropy Research on Dementia EEG Signals', published on the Computer and Information Technology, 2004. CIT '04. The Fourth International Conference on IEEE transaction in biomedicine. 885 – 889

[7] Jakob stastny, Pavel Sovka, Andrej Stancak.[2003] 'EEG Signal Classification: Introduction to the Problem' Conference on Radio Engineering . 98–108.

[8] J Jamei, MB Shamsolahi. [2003]'Adaptive Zero-Tracking of EEG Signals for Detecting Epileptic Patients' in the proceedings of 2003 *Summer Bioengineering Conference*. 1255–1256.

[9] Mirghasemi H, Shamsollahi, MB Fazel-Rezai R. [2006] 'Assessment of Preprocessing of classifiers used in the P300 speller paradigm' International Conference on BCI Sep 2006. 1319 – 1322.

[10] M Ungureanu, C Bigan, R Strungaru, V Lazarescu. [2004] 'Independent Component Analysis Applied in Biomedical Signal Processing' *Measurement Science Review*, 4(2):1–8.

[11] MS Ghiyasvand, SK Guha, S Anand, KK Deepak. [1995] 'A new EEG Signal Processing Technique for Discrimination of Eyes close and Eyes Open' Proceedings RC IEEE-EMBS & 14th BMESI . 15–18.

[12] Parisa Shooshtari, Gelareh Mohamadi, Behnam Molaee Ardekani, Mohammad Bagher Shamsollah. [2006] ' Removing Ocular Artifacts from EEG signals using Adaptive Filtering and ARMAX Modeling' P*WASET* 11 : 277–280.

[13] Srinivasan V, Eswaran C, Sriraam N. (2007) 'Approximate Entropy-Based Epileptic EEG Detection Using Artificial Neural Networks', *IEEE transaction information technology in biomedicine,* 11(3): 288 – 295

[14] Touradj Ebrahimi. [2002] 'EEG Signal Classification for Brain Computer Interface Applications'. *Ecole Polytechnique Federale De Lausanne*. 45–55.

[15] Wong L, Abdulla W. [2006] 'Time-frequency evaluation of segmentation methods for neonatal EEG signals, '*IEEE EMBS Annual International Conference*'. 1303–1306

ABOUT AUTHORS

Prof. Sasikumar Gurumurthy *is an Assistant professor (Senior) in the school of computing sciences and engineering, VIT University, at Vellore, Tamil Nadu, India, has published more than 38 technical papers in international journals/ proceedings of international conferences. He is having more than 6 years of teaching experience. He is a member of international professional associations like CSI, IAENG and AIRCC and is a reviewer of AIRCC international journals. Also, he is in the editorial board of AIRCC. His current research directions include detecting technique and signal processing, intelligence computation and soft computing.mechatronic systems and robotics.*

Dr. BK Tripathy *is a senior professor in the school of computing sciences and engineering, VIT University, at Vellore, India, has published more than 155 technical papers in international journals/ proceedings of international conferences/ edited book chapters of reputed publications like Springer and guided 12 students for PhD. so far. He is having more than 30 years of teaching experience. He is a member of international professional associations like IEEE, ACM, IRSS, CSI, IMS, OITS, OMS, IACSIT, IST and is a reviewer of around 21 international journals which include IEEE, World Scientific, Springer and Science Direct publications. Also, he is in the editorial board of at least 11 international journals. His current research interest includes Fuzzy sets and systems, Rough sets and knowledge engineering, Granular computing, soft computing, Data clustering, Database anonymisation techniques, bag theory, list theory and social network analysis.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Thangakrishnan and Ramar

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A SURVEY ON FORENSIC SKETCH MATCHING

M. Suresh Thangakrishnan* and Kadarkaraiyandi Ramar

Einstein college of Engineering, Tirunelveli - 627012, Tamil Nadu, INDIA

ABSTRACT

Biometrics is one of the unique ways of identifying a person by the physiological features in the human body. Various biometric techniques includes features in the human body like the facial, iris, gestures, fingerprint, gene, key stroke biometrics, etc. In the facial recognition many algorithms are explored highly with various different orientations. The facial matching framework accepts the input as faces and the outputs the recognized faces from the image database. The objective of forensic sketch matching is the mapping between the image databases with the sketches. Since the information obtained from the victims is almost inadequate the mapping is very complicated. The selection of the features and modeling them for matching without any human intervention is still a challenging task. This paper gives a survey on the various sketch matching techniques that are used in the face matching and recognition. The complexity in forensic sketch matching is analyzed and a new model based on the neural network is proposed to automate the forensic sketch matching system.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 16th –Aug-2015

KEY WORDS

Biometrics; Forensic sketch; Sketch matching; Viewed sketches; Face recognition

***Corresponding author: Email:** suresh.nellai@gmail.com, **Tel.:** +91-9842521145; **Fax:** +91-0462-2487111

INTRODUCTION

Biometrics is a technique for identifying a person based on the physiological features of the human body. Some of the biometric techniques include fingerprint recognition, facial recognition, iris recognition, key stroke biometrics, signature, voice recognition, hand geometry, DNA. Numerous algorithms are available for face recognition and matching the faces. Various faces are identified automatically by the algorithms. The finger print is also an important feature to uniquely identify the persons. The advantage of finger print biometrics is in the case of twins the face resembles but the finger print of the twins cannot be same. In the forensics finger print plays a major role to identify the suspect easily. But in the forensic images it is impossible to depict the finger print by the artist from a victim. This also applies to the iris recognition where the iris pattern of the human is identified and recognized. In the case of the forensic matching the mapping is to be performed with the facial image and a sketch. So it leads to the development of various face matching algorithms with the forensic sketch and original image. Even the suspect can be identified by the voices, but sometimes it can be duplicated by the hackers. In the DNA a gene matching is done for identifying the suspect. In the key stroke biometrics the key press events of the users are identified and matched. It is useful in detecting uneven access by the fraudulent in the login pages. In forensics the Key stroke and gene plays a lesser role. But the finger print and facial features plays a vital role in identifying the suspect in a crime scene. The major motivation behind this paper is in all the existing models a manual inspection is needed to confirm the criminals. So it is very much necessary for an automated system without human intervention to identify and arrest the criminals before they commit the next assassination.

OUTLINE OF FACE MATCHING AND SKETCH MATCHING

Categories of face matching

Many face matching algorithms are proposed to uniquely identify and recognize the faces. The facial point's distances are collected from the face image and they are mapped with the input image for identifying the faces. Also researchers have explored various face detection algorithms with different categories. **Figure– 1** shows the various categories given below. It includes face recognition with multiple orientations like straight pose, side pose, with and without spectacles, under various face expressions (mouth open, mouth closed, mouth opened with teeth) and photos

THE IIOAB JOURNAL
ISSN: 0976-3104

under various illuminations. Also varieties of researches are going with the images under age variations which become a massive challenge.

Fig: 1. Various ways for face detection

Overview of forensics and sketches

Face matching is a difficult process as it is taken from various sources under various conditions. The sources of image are from surveillance cameras, social networks, and mobile cameras etc which are captured under different conditions. There must be a source to capture the images for the further process. But difficulties arise if there are no image capture sources. It becomes a big challenge for the police to find out the suspect without an image. So to identify the suspect the police identify the people in the crime environment and ask for the information about the person for the further investigation.

Forensic is a logical technique for collecting and analyzing the information about the crime in the earlier period. The aim is to detect the crime and identify the suspect based in the information collected. After collecting the information a sketch is drawn named forensic sketches. Sketches can be categorized into two types one is viewed sketches and another is forensic sketches. As for as viewed sketches in considered the sketch is drawn by viewing the photos of the particular person. So a qualified artist is needed to draw the face of the person. Also the task is not much complex as the photo of the person is available [13].

In the next case of the forensic sketches the information about the person is not available with the artist. All the information are to be collected by the people in the surrounding area where the incident has turn out. So the artist sits along with the people who were on the spot and analyses about the various facial features and draw the sketches. This seems to be complex problems where the image of suspect drawn will not be exactly matching. Many forensic artists are specialized in drawing the forensic sketches [3].

In the case of the forensic sketches certain methodology are to be followed strictly like more preprocessing techniques cannot be applied. The reason behind this is the people in the spot can remember more of the external features of the face than the internal features. External features include hair style, race, color, age factor, gender etc. Internal features are rarely identified. A high preprocessing may leave most of the external features. So a novel preprocessing method is always necessary during the process of forensic sketches. **Figure– 2** shows the steps in the forensic sketch matching for face matching [13].

OVERVIEW OF SKETCH MATCHING TECHNIQUES

There are various techniques available with face sketch recognition. The various images of a single person are combined into a single image and a model is created based on the combination to enhance the matching of the forensic sketches. Even though this method provides an acceptable result a human analysis is needed to pin point the final suspect [1]. The meaningful information from the facial attributes are obtained and cross modeled architecture was created, it works with caricature sketches but still it out performs with forensic sketches [2]. Availability of image data base to for the matching is very less so researchers started to create test set databases with sketches for the matching [4]. The face is converted into a sketch and the matching is performed over the image with the sketches. The comparison with the geometry and the Eigen face methods shows a better performance [5].

THE IIOAB JOURNAL
ISSN: 0976-3104

Also face photo match retrieval through the sketches are performed by an invariant descriptor called Gabor shape [6]. The features extracted from the viewed face sketches must be spotted in varying illumination, noise and scaling. So researchers has even developed detection using SIFT (Scale Invariant Feature Transform). As for the above said methods the complexity of the problem was not further explored with large set of images [7]. The methodology of converting image into a sketch and vice versa is also performed for the matching of the image. Multi scale markov random fields are used in synthesis and recognition. So a new combined sketch photo model is learned in this method [8].

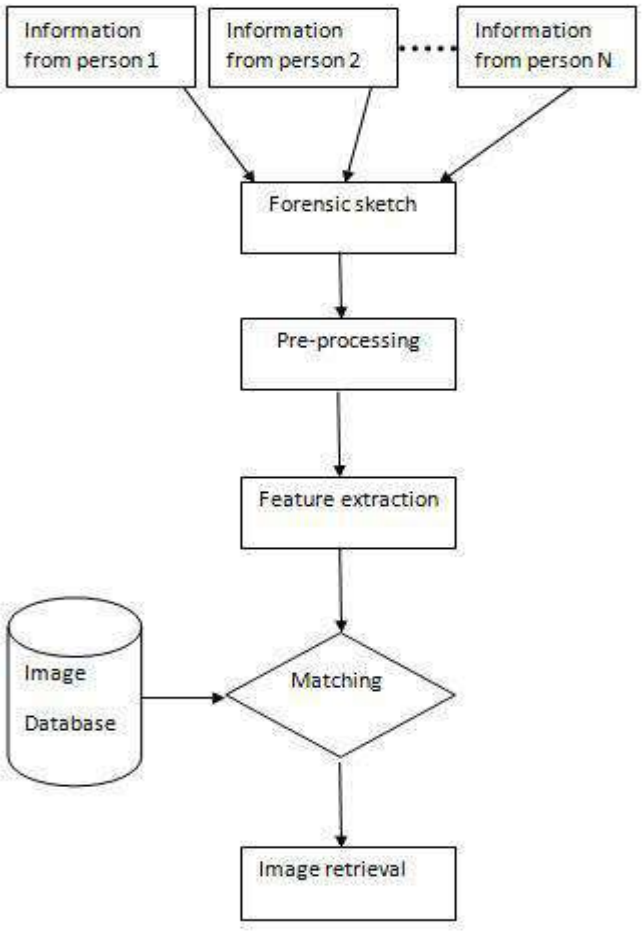


Fig: 2. Steps in forensic sketch matching

The external features of the suspect can easily identify and explained by the people in the spot. Researchers have developed methods to retrieve images using the tattoo. To eliminate the duplicates various metadata were used by the researchers. The system developed is based on unsupervised method. So further extensions can be done in the semi supervised methods and unsupervised methods [9].

In some methodologies a pseudo sketch based on the local linear safeguard of geometry between the photo and sketch images are generated and non discriminate analysis were used in the recognition of the probe sketch. The external feature hair is omitted in this methodology. Also the above said method will not work with different pose, illumination and photos with different conditions [10]. Certain features in the face can be remembered by the humans easily which includes marks in the faces. In order to identify the facial marks researchers have developed various methods using morphological operators. But still the accuracy can be improved with an automatic mark deduction system [11].

The above said methods were performed only for the viewed sketches. Only two large scale experiments were performed till date with the forensic sketches. The first large scale was done by Brendon klare et al for the

THE IIOAB JOURNAL
ISSN: 0976-3104

forensic sketches. The sketches and photos are distinguished by SIFT and Multi scale local binary patterns image descriptors. Filtering was done by race and gender also the results were compared which shows improvement in the accuracy [13]. The second large scale was performed by Kotha *et al* where the methodology based on Speeded Up Robust Features (SURF). Also an improved preprocessing method was proposed in order to preserve the external features. But the filters were not used for the improvement of the results [14]. At one point of time a human suggestion is needed to identify the images based on the comparison of top N rankings of the images. A fully automated forensic sketch matching is yet to be developed.

PROPOSED ARCHITECTURE

A proposed architecture that using fuzzy and neural network to reduce manual intervention is given in **Figure–3**

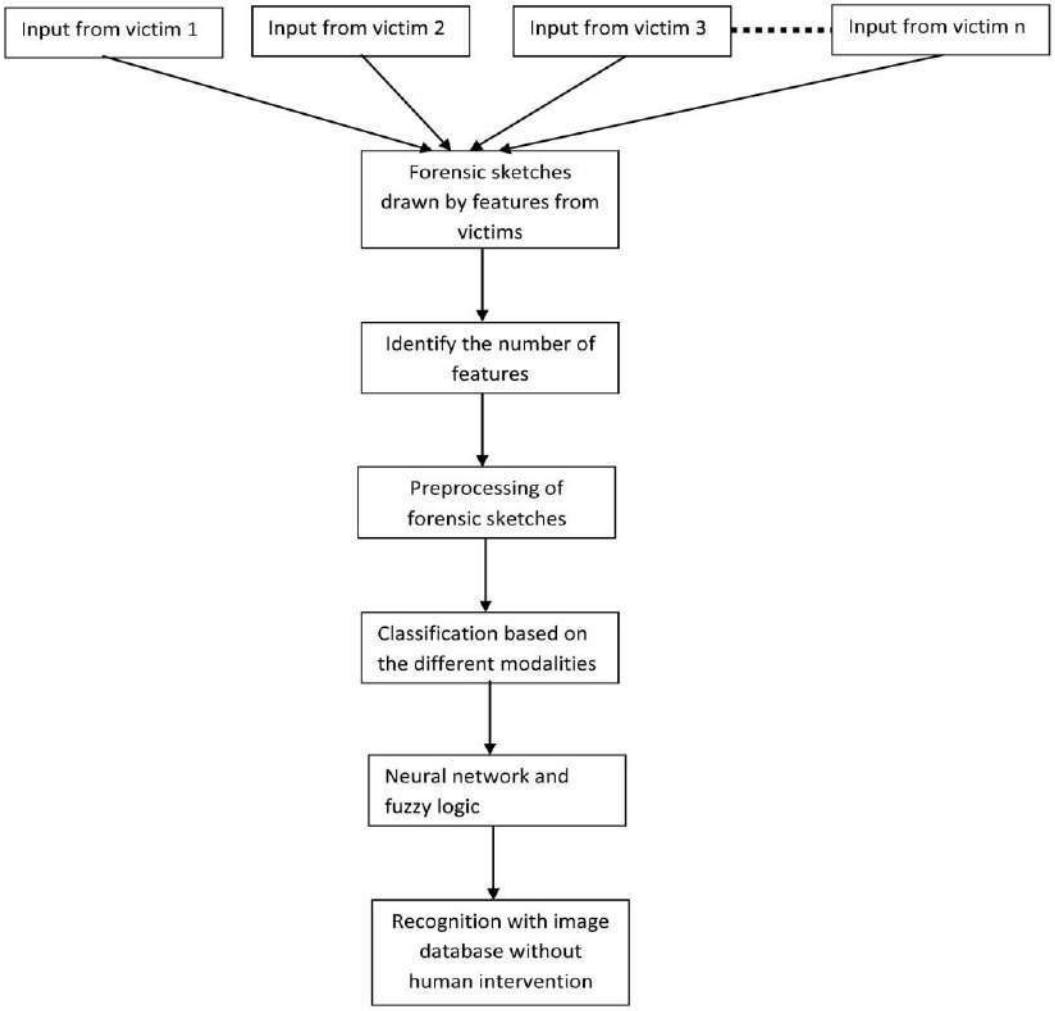


Fig: 3. Using fuzzy and neural network to reduce manual intervention

CONCLUSION

This paper gives a clear picture of all the existing methods and the research gaps in the existing methods. So the new framework will take in account of all the facial attributes with the balance of the neural network an automated forensic sketch facial recognition will give an accurate output. The systems complexities are to be analyzed. But due to the insufficient availability of the database the complexity analysis were not performed at all the cases. The

THE IIOAB JOURNAL ISSN: 0976-3104

forensic sketches were not performed over various poses with a higher efficiency. So in future an efficient multi pose forensic sketch recognition must be performed without any human intervention.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan Klare, Anil K Jain. [2014] Unconstrained Face Recognition: Identifying a Person of Interest From a Media Collection. *IEEE Transactions on Information Forensics and Security. 9(12).*

[2] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li. [2014] Cross-Modal Face Matching: Beyond Viewed Sketches. *Computer Vision -- ACCV 2014. 210– 225*

[3] Dipeeka S Mukane, SM Hundiwale, Pravin U.Dere. [2014] Emerging forensic face matching technology to apprehend criminals: A survey. *IJAET 7(1): 255-262.*

[4] Georgy Kukharev, Katarzyna Buda, Shchegoleva Nadegda. [2014] Sketch generation from photo to create test databases. *Przegląd Elektrotechniczny, 90 (2): 97-100.*

[5] Xiaoou Tang, Xiaogang Wang. [2004] Face Sketch Recognition. *IEEE Transactions on. Circuits and Systems for Video Technology*, 14(1): 50– 57.

[6] Hamed kiani galoogahi, Terence sim. [2012] Face photo Retrival by Sketch Example. *Proceedings of the 20th ACM international conference on Multimedia* Pages *949–952.*

[7] Mohd Ahmed, Farheen Bobere. [2012] Criminal Photograph Retrieval based on Forensic Face Sketch using Scale Invariant Feature Transform. *International Conference on Technology and Business Management 801 – 806.*

[8] Xiaogang Wang and Xiaoou Tang. [2009] Face Photo-Sketch Synthesis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(11): 1955-1967.

[9] Jung-Eun Lee, Rong Jin, and Anil K Jain. [2012] Michigan State University Wei Tong Carnegie Mellon University. Image Retrieval in Forensics:Tattoo Image Database Application. *IEEE Computer Society, 40– 49.*

[10] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, Songde Ma.[2005] A Nonlinear Approach for Face Sketch Synthesis and Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*

[11] Park AK Jain. [2010] Face Matching and Retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security 5(3).*

[12] Anil K Jain, Brendan Klare, Unsang Park Michigan State University. [2012] Face Matching and Retrieval in Forensics Applications. *IEEE Computer Society , 2– 10.*

[13] Brendan F Klare, Zhifeng Li, Anil K. Jain. [2011] Matching Forensic Sketches to Mug Shot Photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(3): 639–646.*

[14] Dileep Kumar Kotha, Santanu Rath. [2013] Forensic Sketch Matching Using SURF. *Proceedings of ICAdC, AISC 174:527–537.*

ABOUT AUTHORS

Prof. M Suresh Thangakrishnan *currently working as an assistant professor in the Computer Science and Engineering department at Einstein College of engineering. Also pursuing his research in Anna University Chennai. He has publications in international journal and national conferences. His area of interest includes image processing and computer networks. He is a life time member in computer society of India.*

Dr. K Ramar *currently working as an professor and principal at Einstein college of Engineering. He has twenty five years of research and teaching experience. He has received three research funded projects from government sector. Also he has received various seminar grants. He is an expert committee member in various journals and colleges. His research areas include image processing, computer networks and neural networks. He has a splendid international and national journal and conference publications. He is a member of various professional bodies.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Paneer and Tripathy.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

TUNED HYBRID SOFT CLUSTERING ALGORITHM FOR UNCERTAIN INFORMATION SYSTEM

Prabhavathy Paneer[1*] and Balakrishna Tripathy[2]

[1]*School of Information Technology and Engineering, VIT University, Vellore, INDIA*
[2]*School of Computing Science and Engineering, VIT University, Vellore, INDIA*

ABSTRACT

Clustering is an important mission in the field of machine learning, pattern recognition and web mining. Handling uncertain data in the information system is one of the key research topics in the vicinity of knowledge representation. Number of clustering algorithms are available [23][6][12]27]; but many of those algorithms are challenging when dealing with uncertain data. The aim of the paper is to tune two existing rough c-means and fuzzy c-means and integrate them into a tuned hybrid soft clustering algorithm termed as the tuned rough-fuzzy c-means algorithm. Rough c-means is extremely sensitive to the initial placement of the cluster centers. The proposed algorithm is enhanced by introducing dynamic centroid computation. The proposed algorithm performance is compared with the existing rough c means, fuzzy c-means, and rough fuzzy c –means approaches. The effectiveness of the algorithm is verified on real and synthetic datasets.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 16th –Aug-2015

KEY WORDS

Clustering; Uncertain; RoughSet; Fuzzy; C-means

***Corresponding author: Email:** pprabhavathy@vit.ac.in, **Tel.:** +91-9486236259; **Fax:** +91-416-2243092

INTRODUCTION

Cluster analysis [1] is a technique for finding natural groups present in the data. It divides a given data set into a set of clusters in such a way that two objects from the same cluster are as similar as possible and the objects from different clusters are as dissimilar as possible. Clustering techniques have been effectively applied to a wide range of engineering and scientific disciplines such as pattern recognition, machine learning, psychology, biology, medicine, computer vision, communications, and remote sensing. A number of clustering algorithms have been proposed to suit different requirements. Clustering is categorized as hard or soft in nature .Soft clusters may have fuzzy or rough boundaries. In hard clusters, the elements which are similar to each other are placed in the same cluster. The elements whose natures differ with each other drastically are placed in different clusters. Soft clustering [12] helps researchers to discover overlapping clusters in many applications such as web mining and text mining. Hence soft clusters may have two types of boundaries 1) Fuzzy boundary 2) Rough boundary. Fuzzy clusters need an association degree to distinguish each element present in the cluster. The elements in the rough clusters are distinguished with the help of boundary region. The relations between rough sets and fuzzy sets were compared [2, 4]. On the whole, both theories deal with the difficulty of information granulation: the theory of fuzzy sets is centered upon fuzzy information granulation, where as rough set theory is paying attention on crisp information granulation.

Data generation methods create uncertain, incompleteness, and granularity in information system which provides inaccurate result in data analysis. Rough set theory is a valuable tool for data mining. In the past few years the concept of basic rough sets has been extended in many different directions. The original rough set theory proposed by Pawlak [18-20] is based upon equivalence relations defined over a universe. It is the simplest formalization of indiscernibility. However, it cannot deal with a number of uncertain problems in real information systems. This has direct to numerous significant and motivating extensions of the original concept. Bezdek's fuzzy c-means [10, 11] is the another most popular soft clustering algorithm for many real life applications in a very diverse range of domains. K-means is one of the most extensively used partitioned based clustering algorithms and it is extremely sensitive to the initial placement of the cluster centers. Numerous initialization methods have been proposed [15] to deal with this problem. Efficient hybrid evolutionary data clustering algorithm K-MCI [9] has been presented to handle high dimensional data and large cluster. Fuzzy clustering is suitable to classify ordered sequences in human activity pattern analysis [22]. However, the majority of the present fuzzy clustering modules [3, 4, 16] packaged in both open source and commercial products

THE IIOAB JOURNAL ISSN: 0976-3104

have lack of enabling users to explore fuzzy clusters extremely and visually in terms of examination of different relations among clusters.

Attribute weighted fuzzy clustering has became a very active area of research and interval number has been introduced for attribute weighting in the weighted fuzzy c-means (WFCM) clustering approach [13, 25]. The existing fuzzy and rough clustering approaches have been refined based on the concept of shadowed sets. Shadowed clustering [26] has been presented which serves as a conceptual and algorithmic bridge between the FCM and RCM. Much work has been carried out using rough c-means, fuzzy c-means and rough fuzzy c-means in data clustering. The extensive survey of the significant extensions and derivatives of soft clustering approaches have been studied [5, 7]. In this paper, tuned rough fuzzy c-means clustering approach is proposed to resolve the uncertainty of information system.

This paper focuses on traditional rough fuzzy c-means and tuned rough fuzzy c-means approaches for handling uncertainty presents in the information system. The remainder of this paper is organized as follows. The introduction about the work is discussed in section 1. In Section 2, traditional soft clustering algorithms are discussed under materials and methods. Section 3 investigates experimental analysis of tuned soft clustering algorithm for uncertain data. Section 5 discusses performances of the proposed algorithm and this paper concludes in section 6.

MATERIALS AND METHODS

Traditional soft clustering

Fuzzy sets and roughsets [27-31] were incorporated in the c-means framework to develop the fuzzy c-means (FCM) [10], rough c-means [6, 8, 12, 23, 24] and rough–fuzzy c-means (RFCM) [21, 23] algorithms, respectively. While membership in FCM enables efficient handling of overlapping partitions, the roughest [17, 19] deal with uncertainty, vagueness and incompleteness of data in terms of upper and lower approximations.

Rough C-means

Rough c-means algorithm was introduced by Lingras, which describes a cluster by its centroid and its lower and upper approximations. In rough c- means, an object can belong completely in one cluster or can be in the uncertainty region or boundary of two clusters. The lower and upper approximations are weighted differently. In each iteration step of the algorithm, the distance of objects from the cluster centroids are computed and if the difference between the two lowest distances is less than a specified threshold value the element is placed in the boundary of the two clusters. Otherwise, the element is placed in the cluster for which the distance is the minimum.

Fuzzy C-means

Developed by Bezdek, the fuzzy C-means algorithm is a powerful method to classify fuzzy data by using the concept of objective function. This approach which minimizes the objective function is expressed in the form of an iterative algorithm makes it possible to reach at an optimal solution, where the solution space is of infinite cardinality. In fuzzy c-means data may belong to one or more than one clusters. It brings in the concept of having membership values. Each object will have a membership in every cluster; which represents the degree to which the element belongs to the cluster. So, here also the clusters are not disjoint. The multiple membership of data models uncertainty of elements belonging to clusters.

Rough-Fuzzy C-Means

It combines the concepts of rough set theory and fuzzy set theory. It has been established that the rough membership function is more general than the fuzzy membership function. However, this generalized membership function has some costs to pay as it does not provide a formula to find the membership values for union and intersection of rough sets. However, in fuzzy set theory we have definite formulae for the computation of the membership values. Thus the hybrid algorithms takes care of both the features by providing membership values to elements as well as modeling vagueness in data through the boundary concept. The concepts of lower and upper approximations in rough set deals with uncertainty and, vagueness whereas the concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters.

According to rough set theory if $x_j \in \underline{B}Ui$ then object x_j is contained completely in cluster U_i and if then object x_j belongs to cluster U_i and also belongs to another cluster. Hence according to fuzzy set theory the objects in boundary approximation should have different degree of membership on the clusters. So in RFCM the membership values of objects in lower approximation are μ_{ij} = 1 while for those in boundary region are determined by the membership values.

ISSN: 0976-3104

1. Assign initial means v_i, i=1, 2, 3,…., c. Choose values for fuzzifier m_1 and threshold ε and δ. Set iteration counter t=1.
2. Compute membership μ_{ij} by equation (1) for c clusters and n objects.
3. If μ_{ij} and μ_{ik} be the two highest membership value of x_j and $(\mu_{ij} - \mu_{kj}) \leq \delta$, then $x_j \in \overline{A}(\beta_i)$ and $x_j \in \overline{A}(\beta_k)$. Furthermore, x_j is not part of any lower bound.
4. Otherwise, $x_j \in \underline{A}(\beta_i)$. In addition, by properties of rough sets, $x_j \in \overline{A}(\beta_i)$.
5. Modify μ_{ij} considering lower and boundary regions for c clusters and n objects.
6. Compute new centroid as per equation (1).
7. Repeat steps 2 to 7, by incrementing t, until $|\mu_{ij}(t-1) - \mu_{ij}(t)| > \varepsilon$

$$v_i = \begin{cases} w \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$$

$$C_1 = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{1}{n_i} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j \quad where \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$$

$|\underline{A}(\beta_i)|$ represents the cardinality of $\underline{A}(\beta_i)$. 0 < w < w < 1

Tuned soft clustering

In particular the recent promising developments in the fusion of soft cluster algorithms show the need for approaches that holistically address uncertainty. Hence, soft clustering will continue in the attention of researchers and most likely attract yet more practitioners in the ground of data mining in support of their real life applications. The objective of this paper is to analyze Georg Peters [6] cluster algorithm rigorously and point out potential for further development. Based on the analyze we have presented a tuned rough fuzzy cluster algorithm and apply it to synthetic and real time market data.

Tuned rough C-means [6]

Lingras et.a[12] discussed rough clustering algorithm. Georg Peters evaluated Lingras et al. rough cluster algorithm and recommended some alternative solutions. This led to the new refined rough k-means algorithm.

Georg Peters cluster rough cluster algorithm goes as follows:

(a) Initialization: Randomly assign each data object to exactly one lower approximation. Hence, the data object will also belong to the upper approximation of the same cluster.

(b) Calculation of the new means. The means are calculated as follows:

$$\overline{m}_k = \omega_l \sum_{\underline{X}_n \in \underline{C}_k} \frac{\vec{X}_n}{|\underline{C}_k|} + \omega_u \sum_{\underline{X}_n \in \underline{C}_k} \frac{\vec{X}_n}{|\overline{C}_k|}$$

With $\omega_l + \omega_u = 1$

Now, the lower approximation of each cluster always has at least one member. Therefore $|\underline{C}_k| \neq \varphi, \forall k$ and by definition $|\overline{C}_k| \neq \varphi, \forall k$

(c) (i) Assign the data objects to the approximations. Assign the data object that represents a cluster to its lower and upper approximation.

1. Find the minimal distance between cluster k and all data objects n and assign data object l to lower and upper

THE IIOAB JOURNAL ISSN: 0976-3104

approximation of cluster h:

$$d(\vec{X_l},\vec{m_h}) = \min_{n,k} d(\vec{X_n},\vec{m_k}) \Rightarrow \vec{X_l} \in \underline{C_k} \wedge \vec{X_l} \in \overline{C_k}$$

2. Exclude $\vec{X_l}$ and $\vec{m_h}$. If clusters are left – so far, in the above step (a) no data object has been assigned to them – go back to Step (a). Otherwise continue with Step (ii).

(ii) For each remaining data point $\vec{X_m}'$ (m=1, 2, ….., M, with M=N-K) determine its closest mean $\vec{m_h}$:

$$d_{m,h}^{\min} = d(\vec{X_m}', \vec{m_h}) = \min_{k=1,\ldots,K} d(\vec{X_m}', \vec{m_k})$$

Assign $\vec{X_m}'$ to the upper approximation of cluster h.

(iii) Determine the mean $\vec{m_t}$ that are also close to $\vec{X_m}'$. Take the relative distance as defined above where ζ is a given relative threshold

$$T' = \left\{ t : \frac{d(\vec{X_m}', \vec{m_k})}{d(\vec{X_m}', \vec{m_h})} \leq \zeta \wedge h \neq k \right\}$$

If T'≠ ϕ ($\vec{x_m}'$ is also close to at least one other mean $\vec{m_t}$ besides $\vec{m_h}$).

Then $\vec{X_m}' \in \overline{C_t}, \forall t \in T'$.

Else $\vec{X_m}' \in \underline{C_h}$

(d) Check convergence for the algorithm. If the algorithm has not converged continue with step 2 else stop.

George peters refined rough c-means algorithm by replacing boundary into upper approximation in mean computation

Tuned rough fuzzy C-means

The algorithm as presented by Lingas et al. is numerical instable since there are data constellations where lower approximation is empty in some cases. The clusters will be weak if there is no representative the proposed algorithm ensures that each lower approximation has at least one member. It is implemented by assigning the data point that is closest to a mean to the lower approximation of the cluster. Otherwise the cluster seems to be weak since it has no sure representative. We have used relative distance represented by George peters instead of Lingras' et al. absolute distance measure to determine the set T. Rough C-means is one of the most extensively used partitioned based clustering algorithms and it is extremely sensitive to the initial placement of the cluster centers. Numerous initialization methods have been proposed [15] to deal with this problem. Here, we also addressed the solution for selection of cluster centers.

The tuned rough fuzzy c-means as follows:

Algorithm: Tuned Rough Fuzzy C-means

1 Assign initial means vi, i=1, 2, 3,…., c. Choose values for fuzzifier m_1 and threshold ε and δ. Set iteration counter t=1.

2 Compute membership μij by equation (2) for c clusters and n objects.

THE IIOAB JOURNAL ISSN: 0976-3104

3 If μij and μik be the two highest membership value of xj and (μij /μkj) ≤ δ, then xj $\in\overline{\underline{A}}(\beta_i)$ and xj $\in\overline{\underline{A}}(\beta_k)$. Furthermore, xj is not part of any lower bound.

4 Otherwise, xj $\in\underline{A}(\beta_i)$. In addition, by properties of rough sets, xj $\in\overline{\underline{A}}(\beta_i)$.

5 Modify μij considering lower and boundary regions for c clusters and n objects.

6 Compute new centroid as per equation (2).

7 Repeat steps 2 to 7, by incrementing t, until |μij(t-1)- μij(t)| > ε

$$v_i=\begin{cases} \tilde{w}\times C_1+w\times D_1 & if\ \underline{A}(\beta_I)\neq\varphi, B(\beta_i)\neq\varphi \\ C_1 & if\ \underline{A}(\beta_I)\neq\varphi, B(\beta_i)=\varphi \\ D_1 & if\ \underline{A}(\beta_I)=\varphi, B(\beta_i)\neq\varphi \end{cases}$$

$$C_1=\frac{1}{|\underline{A}(\beta_i)|}\sum_{x_j\in\underline{A}(\beta_i)} x_j \quad D_1=\frac{\sum_{x_j\in\underline{A}(\beta_i)} x_j+\sum_{x_j\in B(\beta_i)}(\mu_{ij})^{m_1}x_j}{n_i+\sum_{x_j\in B(\beta_i)}(\mu_{ij})^{m_1}} \in \quad where\ n_i=\sum_{x_j\in B(\beta_i)}(\mu_{ij})^{m_1}$$

$|\underline{A}(\beta_i)|$ represents the cardinality of $\underline{A}(\beta_i)$. $0 < \tilde{w} < w < 1$

Selection of initial centroid for rough fuzzy C-means algorithm

Step 1: From n objects calculate a point whose attribute values are average of n objects attribute values. Hence, first initial centroid is average on n- objects.
Step 2: Select next initial centroids from n-objects in such a way that the Euclidean distance of that object is maximum from other selected initial centroids.
Step 3: Repeat step 2 until we get k initial centroids.

From these steps the initial centroids are derived and tuned rough fuzzy c-means algorithm is tested for the dynamic centroids and random centroids.

RESULTS

Experimental analysis

The traditional soft clustering algorithms such as rough c-means(RCM), Fuzzy C-means (FCM), Rough-Fuzzy C-means(RFCM), Rough-Intuitionistic-fuzzy C-means (RIFCM) and proposed tuned rough fuzzy c-means (TRFCM) algorithms are implemented using Java.UCI Machine Learning Repository, Wholesale customers Data Set [29] is used to evaluate the performance of the above said algorithms. The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories. The centroid formulae for the algorithms are given in the Table 1. The traditional and tuned soft clustering algorithms are tested with random centroid selection and proposed centroid computation method and it's shown in Figure- 1.

Table: 1. Comparisons of Centroid formulae of the various soft clustering algorithms

Algorithm	Formula for Centroid calculation
RCM	$v_i=\begin{cases} \tilde{w}\times A+w\times B & if\ \underline{A}(\beta_I)\neq\varphi, B(\beta_i)\neq\varphi \\ A & if\ \underline{A}(\beta_I)\neq\varphi, B(\beta_i)=\varphi \\ B & if\ \underline{A}(\beta_I)=\varphi, B(\beta_i)\neq\varphi \end{cases}$ $A=\frac{1}{|\underline{A}(\beta_i)|}\sum_{x_j\in\underline{A}(\beta_i)} x_j \quad B=\frac{1}{|B(\beta_i)|}\sum_{x_j\in B(\beta_i)} x_j$,

THE IIOAB JOURNAL ISSN: 0976-3104

FCM	$v_i = \frac{1}{n_i} \sum_{j=1}^{n} (\mu_{ij})^{m_1} x_j ;\ where : n_i \cdot = \sum_{j=1}^{n} (\mu_{ij})^{m_1}$
RFCM	$v_i = \begin{cases} \tilde{w} \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$ $where \quad C_1 = \frac{1}{\lvert \underline{A}(\beta_i) \rvert} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{1}{n_i} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$,
RIFCM	$V_i = \begin{cases} w_{low} \frac{\sum_{x_k \in \underline{BU}_i} x_k}{\lvert \underline{BU}_i \rvert} + w_{up} \frac{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m x_k}{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m} & if \lvert \underline{BU}_i \rvert \neq \varphi \quad and \lvert BN(U_i) \rvert \neq \varphi \\ \frac{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m x_k}{\sum_{x_k \in BN(U_i)} \left(\mu'_{ik}\right)^m} & f \lvert \underline{BU}_i \rvert = \varphi \quad and \lvert BN(U_i) \rvert \neq \varphi \\ \frac{\sum_{x_k \in \underline{BU}_i} x_k}{\lvert \underline{BU}_i \rvert} & ELSE \end{cases}$
Tuned RCM	$\bar{m}_k = \omega_l \sum_{\underline{X}_n \in \underline{C}_k} \frac{\vec{X}_n}{\lvert \underline{C}_k \rvert} + \omega_u \sum_{\overline{X}_n \in \overline{C}_k} \frac{\vec{X}_n}{\lvert \overline{C}_k \rvert}, \omega_l + \omega_u = 1$
Tuned RFCM	$v_i = \begin{cases} \tilde{w} \times C_1 + w \times D_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) \neq \varphi \\ C_1 & if\ \underline{A}(\beta_I) \neq \varphi, B(\beta_i) = \varphi \\ D_1 & if\ \underline{A}(\beta_I) = \varphi, B(\beta_i) \neq \varphi \end{cases}$ $where\ C_1 = \frac{1}{\lvert \underline{A}(\beta_i) \rvert} \sum_{x_j \in \underline{A}(\beta_i)} x_j \quad D_1 = \frac{\sum_{x_j \in \underline{A}(\beta_i)} x_j + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1} x_j}{n_i + \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}} \quad n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{m_1}$,

THE IIOAB JOURNAL
ISSN: 0976-3104

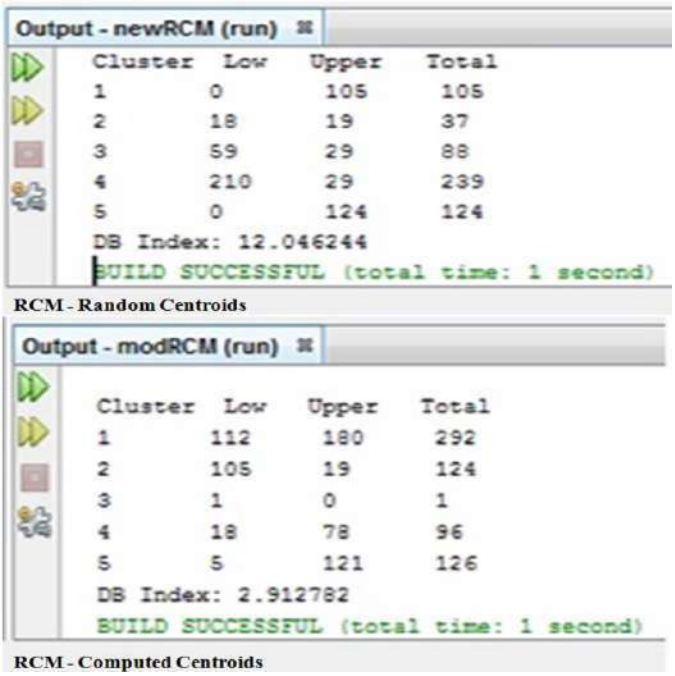


Fig: 1. a

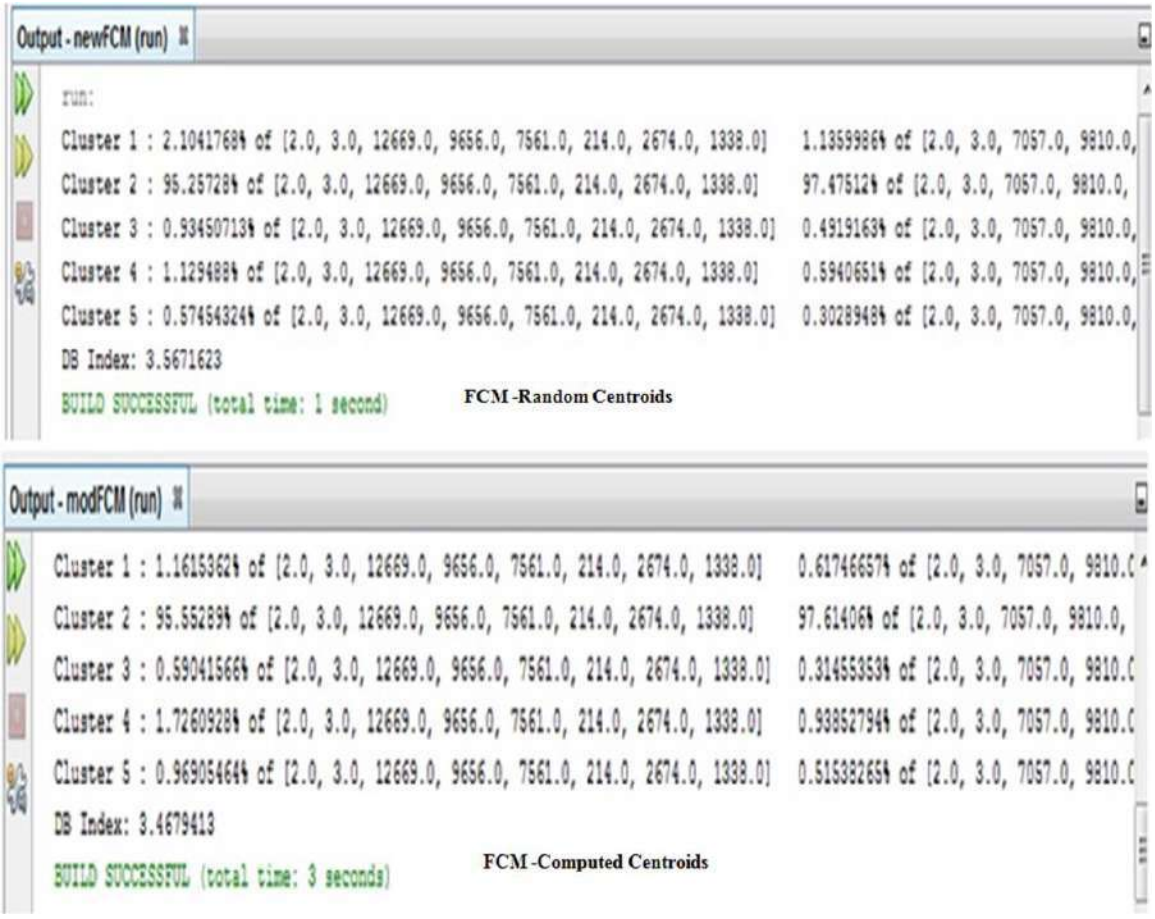
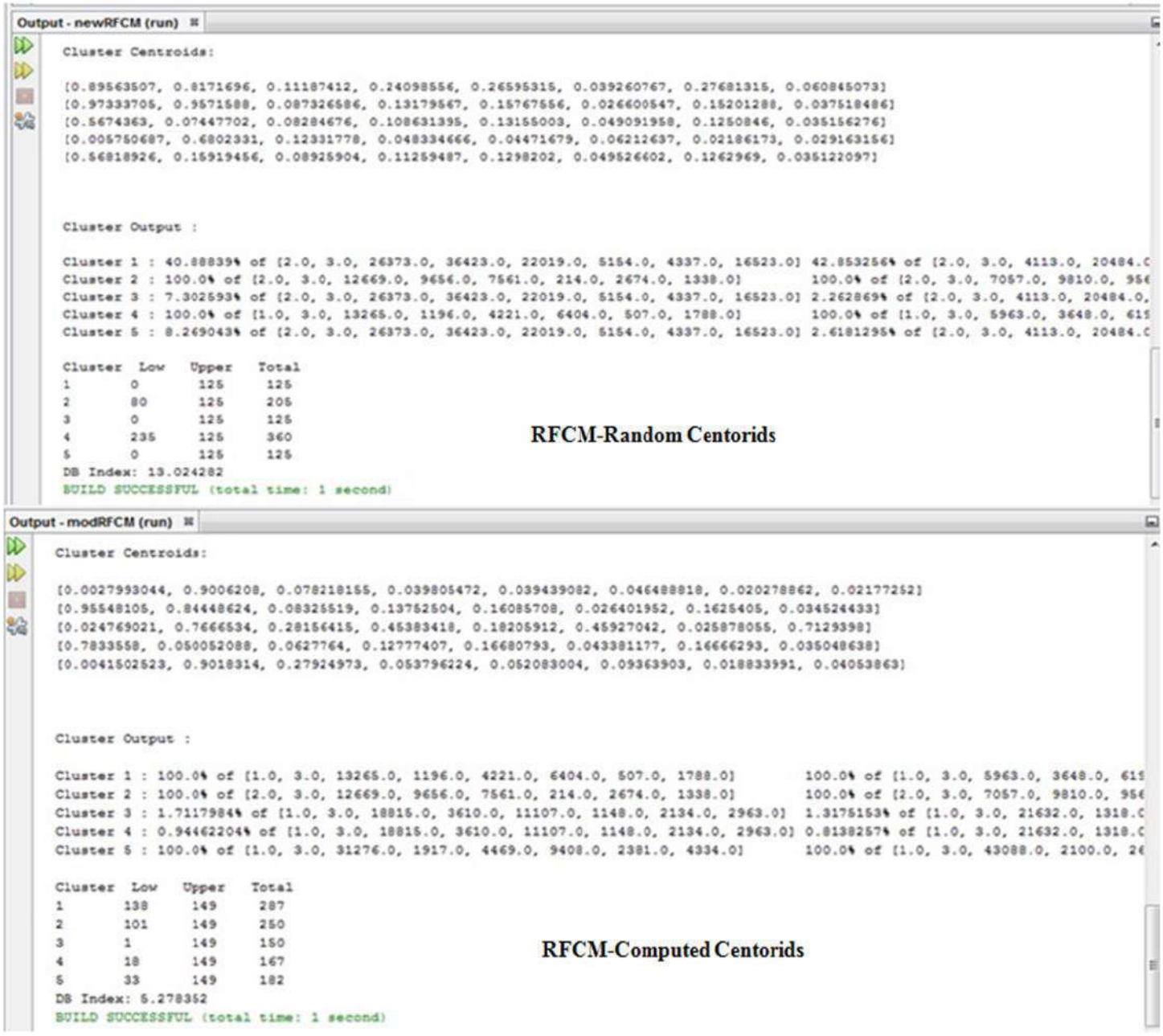


Fig: 1. b

THE IIOAB JOURNAL
ISSN: 0976-3104



```
Output - newRFCM (run)
Cluster Centroids:

[0.89563507, 0.8171696, 0.11187412, 0.24098556, 0.26595315, 0.039260767, 0.27681315, 0.060845073]
[0.97333705, 0.9571588, 0.087326586, 0.13179567, 0.15767556, 0.026600547, 0.15201288, 0.037518486]
[0.5674363, 0.07447702, 0.08284676, 0.108631395, 0.13155003, 0.049091958, 0.1250846, 0.035156276]
[0.005750687, 0.6802331, 0.12331778, 0.048334666, 0.04471679, 0.06212637, 0.02186173, 0.029163156]
[0.56818926, 0.15919456, 0.08925904, 0.11259487, 0.1298202, 0.049526602, 0.1262969, 0.035122097]

Cluster Output :

Cluster 1 : 40.88839% of [2.0, 3.0, 26373.0, 36423.0, 22019.0, 5154.0, 4337.0, 16523.0] 42.853256% of [2.0, 3.0, 4113.0, 20484.0
Cluster 2 : 100.0% of [2.0, 3.0, 12669.0, 9656.0, 7561.0, 214.0, 2674.0, 1338.0]        100.0% of [2.0, 3.0, 7057.0, 9810.0, 956
Cluster 3 : 7.302593% of [2.0, 3.0, 26373.0, 36423.0, 22019.0, 5154.0, 4337.0, 16523.0] 2.262869% of [2.0, 3.0, 4113.0, 20484.0,
Cluster 4 : 100.0% of [1.0, 3.0, 13265.0, 1196.0, 4221.0, 6404.0, 507.0, 1788.0]        100.0% of [1.0, 3.0, 5963.0, 3648.0, 615
Cluster 5 : 8.269043% of [2.0, 3.0, 26373.0, 36423.0, 22019.0, 5154.0, 4337.0, 16523.0] 2.61812958% of [2.0, 3.0, 4113.0, 20484.0

Cluster  Low   Upper   Total
1        0      125     125
2        80     125     205
3        0      125     125
4        235    125     360
5        0      125     125
DB Index: 13.024282
BUILD SUCCESSFUL (total time: 1 second)
```

RFCM-Random Centorids

```
Output - modRFCM (run)
Cluster Centroids:

[0.0027993044, 0.9006208, 0.078218155, 0.039805472, 0.039439082, 0.046488818, 0.020278862, 0.021772525]
[0.95548105, 0.84448624, 0.08325519, 0.13752504, 0.16085708, 0.026401952, 0.1625405, 0.034524433]
[0.024769021, 0.7666534, 0.28156415, 0.45383418, 0.18205912, 0.45927042, 0.025878055, 0.7129398]
[0.7833558, 0.050052088, 0.0627764, 0.12777407, 0.16680793, 0.043381177, 0.16666293, 0.035048638]
[0.0041502523, 0.9018314, 0.27924973, 0.053796224, 0.052083004, 0.09363903, 0.018833991, 0.04053863]

Cluster Output :

Cluster 1 : 100.0% of [1.0, 3.0, 13265.0, 1196.0, 4221.0, 6404.0, 507.0, 1788.0]        100.0% of [1.0, 3.0, 5963.0, 3648.0, 615
Cluster 2 : 100.0% of [2.0, 3.0, 12669.0, 9656.0, 7561.0, 214.0, 2674.0, 1338.0]        100.0% of [2.0, 3.0, 7057.0, 9810.0, 956
Cluster 3 : 1.7117984% of [1.0, 3.0, 18815.0, 3610.0, 11107.0, 1148.0, 2134.0, 2963.0] 1.3175153% of [1.0, 3.0, 21632.0, 1318.0
Cluster 4 : 0.94462204% of [1.0, 3.0, 18815.0, 3610.0, 11107.0, 1148.0, 2134.0, 2963.0] 0.8138257% of [1.0, 3.0, 21632.0, 1318.0
Cluster 5 : 100.0% of [1.0, 3.0, 31276.0, 1917.0, 4469.0, 9408.0, 2381.0, 4334.0]       100.0% of [1.0, 3.0, 43088.0, 2100.0, 26

Cluster  Low   Upper   Total
1        138    149     287
2        101    149     250
3        1      149     150
4        18     149     167
5        33     149     182
DB Index: 5.278352
BUILD SUCCESSFUL (total time: 1 second)
```

RFCM-Computed Centorids

Fig: 1. c

THE IIOAB JOURNAL
ISSN: 0976-3104

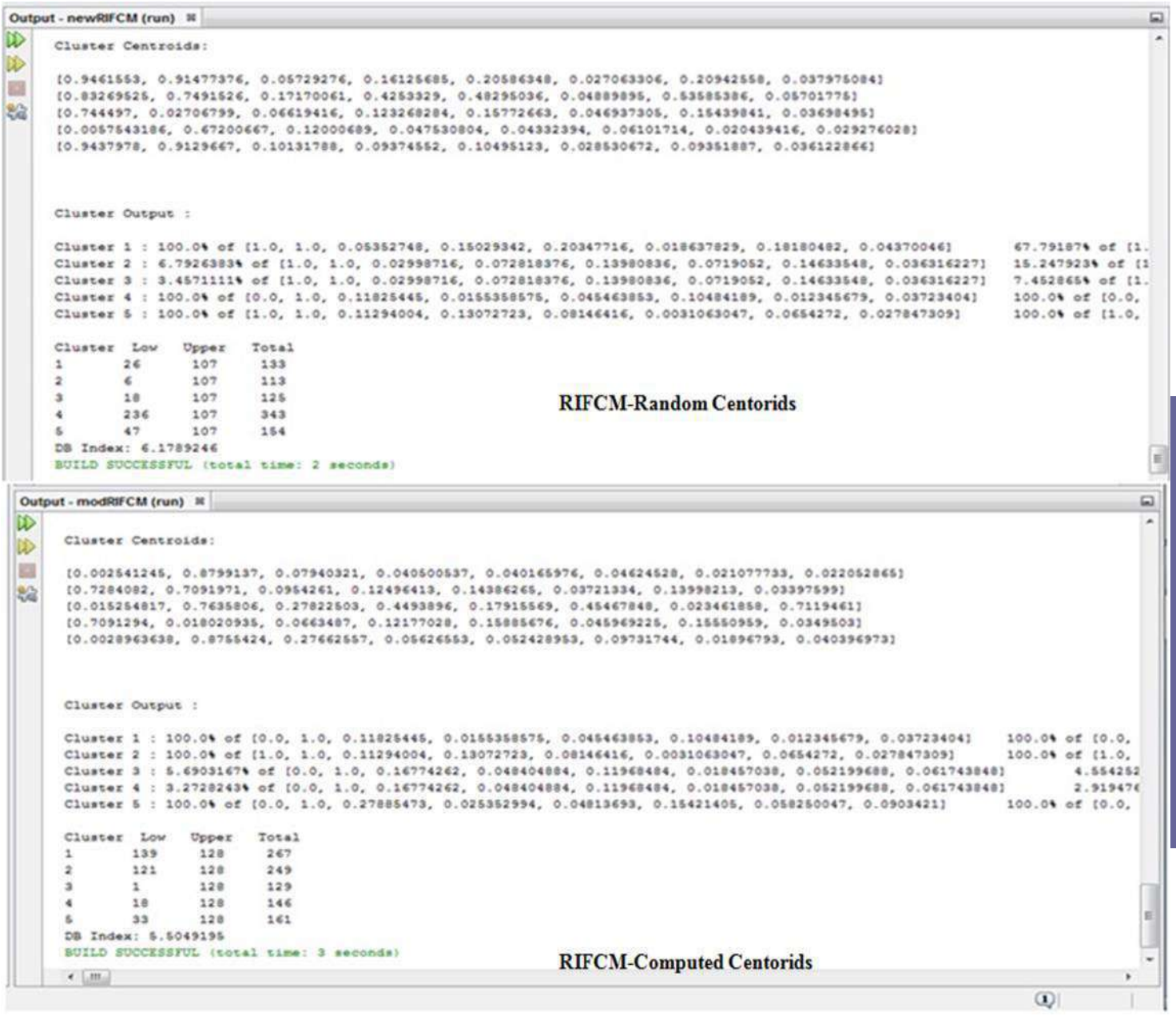
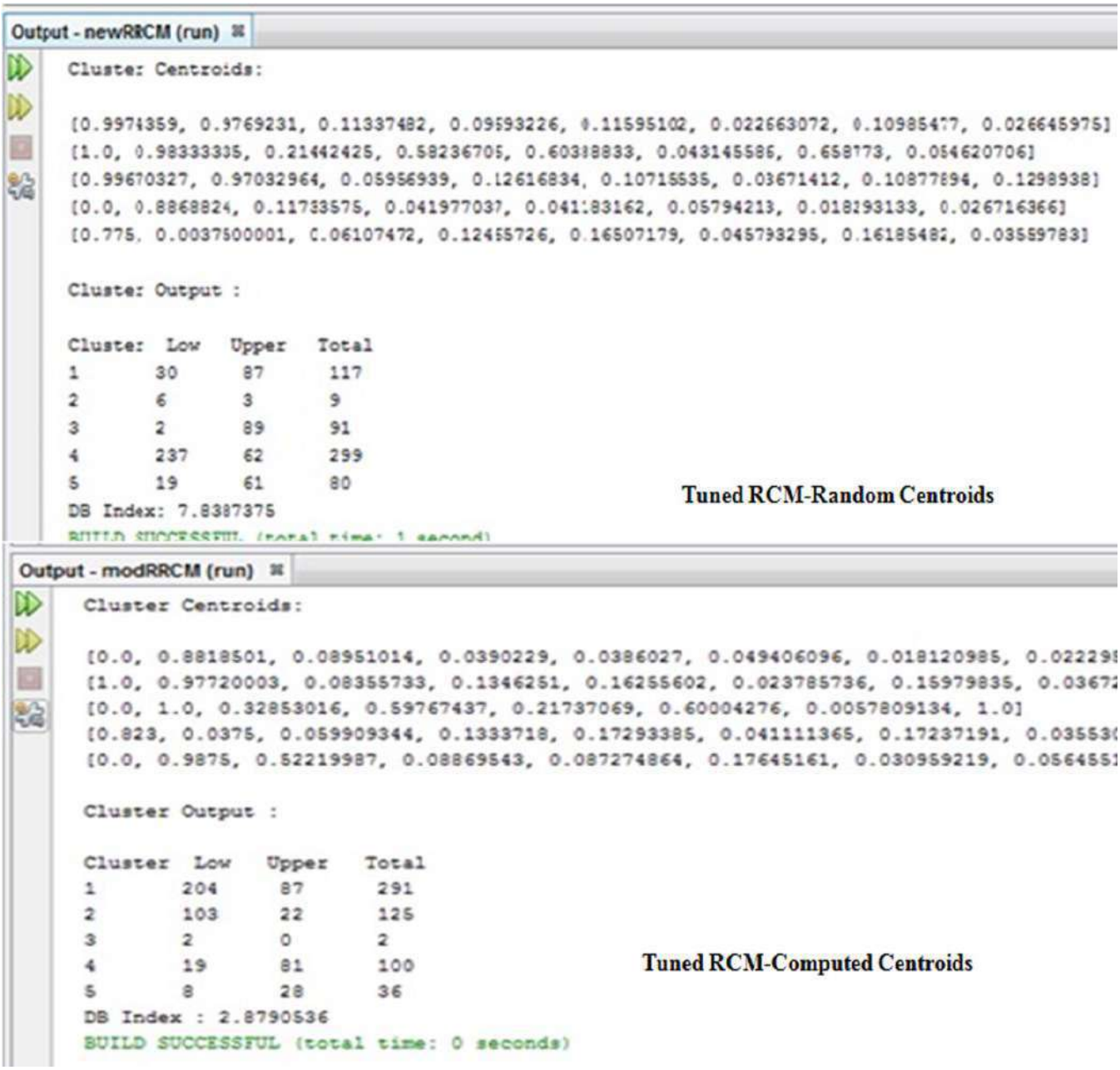


Fig: 1. d

THE IIOAB JOURNAL
ISSN: 0976-3104



Output - newRRCM (run)

```
Cluster Centroids:

[0.9974359, 0.9769231, 0.11337482, 0.09593226, 0.11595102, 0.022663072, 0.10985477, 0.026645975]
[1.0, 0.98333335, 0.21442425, 0.58236706, 0.60338833, 0.043145586, 0.658173, 0.054620706]
[0.99670327, 0.97032964, 0.05956939, 0.12616834, 0.10715535, 0.03671412, 0.10877894, 0.1298938]
[0.0, 0.8868824, 0.11733575, 0.041977037, 0.041183162, 0.05794213, 0.018193133, 0.026716366]
[0.775, 0.0037500001, 0.06107472, 0.12455726, 0.16507179, 0.045793295, 0.16185482, 0.035597833]

Cluster Output :

Cluster  Low   Upper   Total
1        30     87      117
2        6      3       9
3        2      89      91
4        237    62      299
5        19     61      80
DB Index: 7.8387375
BUILD SUCCESSFUL (total time: 1 second)
```

Tuned RCM-Random Centroids

Output - modRRCM (run)

```
Cluster Centroids:

[0.0, 0.8818501, 0.08951014, 0.0390229, 0.0386027, 0.049406096, 0.018120985, 0.022295
[1.0, 0.97720003, 0.08355733, 0.1346251, 0.16255602, 0.023785736, 0.15979835, 0.03672
[0.0, 1.0, 0.32853016, 0.59767437, 0.21737069, 0.60004276, 0.0057809134, 1.0]
[0.823, 0.0375, 0.059909344, 0.1333718, 0.17293385, 0.041111365, 0.17237191, 0.035530
[0.0, 0.9875, 0.52219987, 0.08869543, 0.087274864, 0.17645161, 0.030959219, 0.056455

Cluster Output :

Cluster  Low   Upper   Total
1        204    87      291
2        103    22      125
3        2      0       2
4        19     81      100
5        8      28      36
DB Index : 2.8790536
BUILD SUCCESSFUL (total time: 0 seconds)
```

Tuned RCM-Computed Centroids

Fig: 1. e

THE IIOAB JOURNAL ISSN: 0976-3104

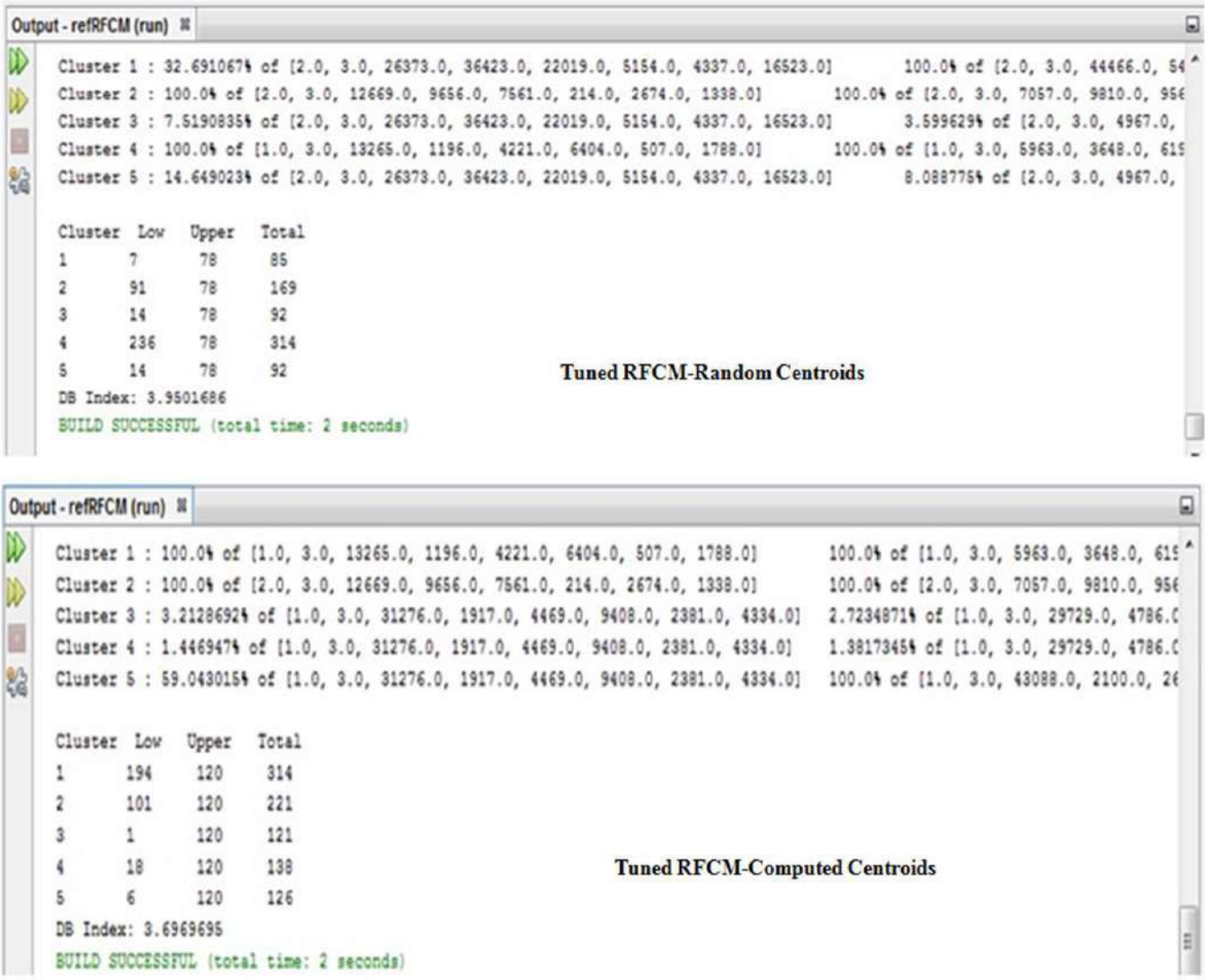


Fig: 1. f

Fig: 1. a-f. The Cluster formation comparisons using random centroid with proposed computation centroid for traditional soft clustering algorithms (RCM,FCM,RFCM,RIFCM) and Tuned hybrid soft clustering algorithms(TRFCM).

The clustering algorithms described are partitive, requiring pre-specification of the number of clusters. The results are dependent on the choice of c.

DISCUSSION

There exist validity indexes to evaluate the goodness of clustering, corresponding to a given value of c. In this paper, we compute the optimal number of clusters c0 in terms of the DB and D cluster validity indexes. The DB is a function of the ratio of the sum of within-cluster distance to between-cluster separation.

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{|ck|}\}$ be a set of patterns lying in a cluster U_k.. Then, the average distance between objects within the cluster U_k is expressed as:

$$S(U_k)=\frac{\Sigma_{i,i'}\|x_i - x_i'\|}{|c_k|(|c_k|-1)} \quad \text{where } \mathbf{x}_i, \mathbf{x}_{i'} \in U_k, \text{ and } i \neq i'$$

The between-cluster separation is defined as:

THE IIOAB JOURNAL ISSN: 0976-3104

$$d(U_k, U_l) = \frac{\sum_{i,j} \|x_i - x_j\|}{|c_k||c_l|}$$

Where $\mathbf{x}_i \in U_k$, $\mathbf{x}_j \in U_l$, such that $k = l$. The optimal clustering, for $c = c_0$, minimizes

$$DB = \frac{1}{c} \sum \max_{j \neq i} \left\{ \frac{S(U_i) + S(U_j)}{d(U_i, U_j)} \right\}$$

for $1 \leq i, j \leq c$. Thereby, the within-cluster distance $S(U_i)$ is minimized while the between-cluster separation $d(U_i, U_j)$ gets maximized. Like DB index, the D index is designed to identify sets of clusters that are compact and separated. Here, we maximize for $1 \leq i, j \leq c$. The inter-cluster separation is maximized, while minimizing intra-cluster distances. Note that the denominator of DB is analogous to the numerator of D.

$$D = \min_j \left\{ \min_{i \neq j} \left\{ \frac{d(U_i, U_j)}{\max_k S(U_k)} \right\} \right\}$$

The computation of the initial centroids of each cluster instead of random allocation generates a lower DB index resulting in clusters with greater accuracy. The traditional and tuned soft clustering algorithms are compared using DB index with default initial centroids and computed initial centroids. The results are shown in **Table– 2** for 5 cluster and **Table– 3** for **Table–4** clusters.

Table: 2. Comparisons of Clustering Algorithms using DB index with different centroids for 5 cluster

Algorithm	No. of Clusters	Del	Epsilon	DB index with default initial centroids	DB index with computed initial centroids
RCM	5	0.3	-	12.046244	2.912782
FCM	5	-	0.05	3.5671623	3.4679413
RFCM	5	0.2	0.05	6.4094977	5.577581
RIFCM	5	0.2	0.05	6.1789246	5.5049195
Tuned RCM	5	1.5	0.05	14.211797	2.0191371
Tuned RFCM	5	1.4	0.05	4.375386	2.335935

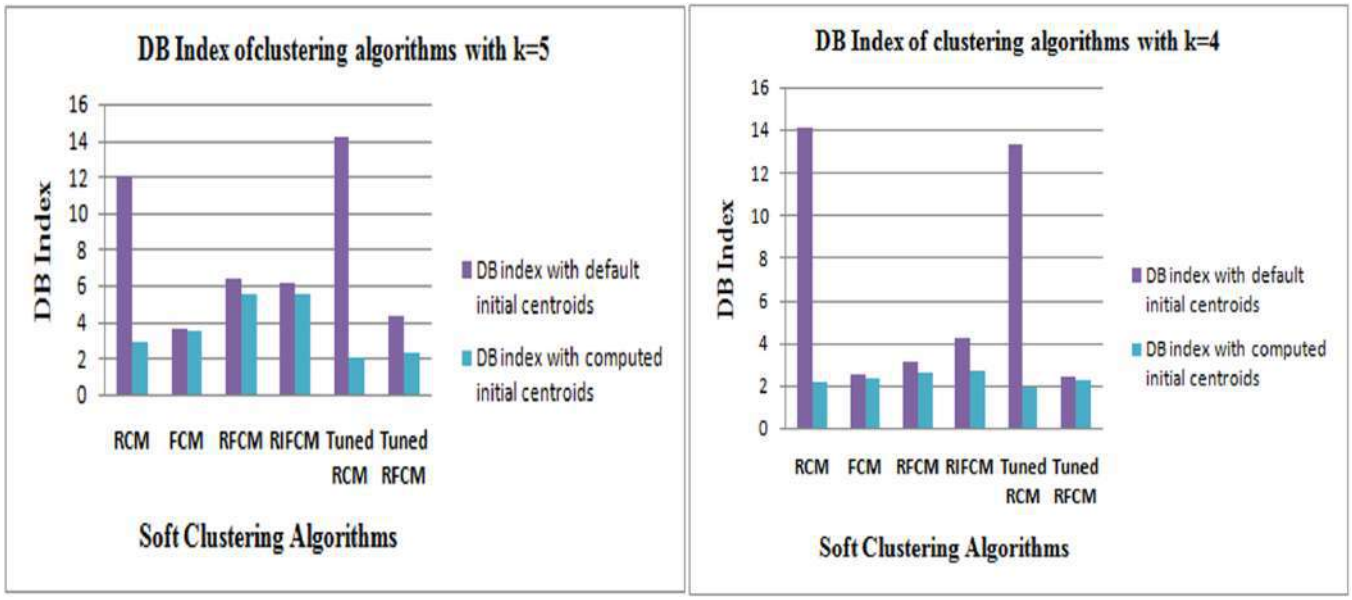
Table: 3. Comparisons of Clustering Algorithms using DB index with different centroids for 4 clusters

Algorithm	No. of Clusters	Del	Epsilon	DB index with default initial centroids	DB index with computed initial centroids
RCM	4	0.2	-	14.110096	2.1260233
FCM	4	-	0.05	2.4907343	2.3366437
RFCM	4	0.2	0.05	3.1227834	2.5758417

THE IIOAB JOURNAL ISSN: 0976-3104

RIFCM	4	0.2	0.05	4.22623	2.7078934
Tuned RCM	4	1.4	0.05	13.310548	1.8863555
Tuned RFCM	4	1.4	0.05	2.4247031	2.2060814

Upon analyzing the output produced by each algorithm in terms of DB index, it can be concluded that the efficiency of the algorithm is greatly affected by the parameters that are used for conclusion.



(a) **(b)**

Fig: 2. Performance of Clustering Algorithms using DB index with different centroids (a) for 5 cluster (b) for 4 Cluster

The performances of the various soft clustering algorithms are represented in **Figure– 2 (a) and (b)** respectively. The various soft clustering approaches are validated with number of cluster 4 and 5. All the approaches are tested for random centriod and dynamic centriod computation. The Result shows that, the proposed tuned RFCM algorithm performs very well than other soft clustering approaches with respect to number of cluster and dynamic centriod computation.

CONCLUSION

Data Clustering is one of the vital research domains with a number of issues. Much of the work done in hard clustering algorithms and a few work carried out in traditional soft clustering algorithms such as rough c-means and fuzzy c-means. In this paper, a tuned hybrid soft clustering algorithm termed as tuned rough fuzzy c-means algorithm is presented. The selection of initial centroid is one of the issues in c-means algorithm, which is resolved by dynamic computation in the proposed algorithm. UCI Machine Learning Repository, Wholesale customers Data Set has been used to compare and validate the performance of the proposed algorithm with traditional soft clustering approaches.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

This work is part of PhD Research work. It is not supported by any agency.

THE IIOAB JOURNAL ISSN: 0976-3104

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] AK Jain, MN Murty, PJ Flynn. [1999] Data clustering: a review, *ACM Computing Surveys* 31 (3) 264–323.

[2] Anna Maria Radzikowska, Etienne E. Kerre.[2002] A comparative study of fuzzy rough sets, *Fuzzy Sets and Systems*, 126(2): 137–155, ISSN 0165-0114.

[3] Selman Bozkir, Ebru Akcapinar Sezer.[2013] FUAT – A fuzzy clustering analysis tool, *Expert Systems with Applications*, 40(3): 15 842–849.

[4] D Dubois, H Prade.[1990] Rough fuzzy sets and fuzzy rough sets, *Internat J General Systems* 17 (2): 3191–209.

[5] Fan Li, Mao Ye, Xudong Chen. [2014] An extension to Rough c-means clustering based on decision-theoretic Rough Sets model, *International Journal of Approximate Reasoning*, 55(1): 116–129.

[6] G.Peters.[2006] Some refinements of rough k-means clustering, *Pattern Recognition* 39 :1481–1491.

[7] Georg Peters, Fernando Crespo, Pawan Lingras, Richard Weber.[2013] Soft clustering – Fuzzy and rough approaches and their extensions and derivatives, *International Journal of Approximate Reasoning*, 54(2):307–322.

[8] Georg Peters, Richard Weber, René Nowatzke,[2012] Dynamic rough clustering and its applications, Applied Soft Computing, 12 (10): 3193–3207.

[9] Ganesh Krishnasamy, Anand J Kulkarni, Raveendran Paramesran. [2014]A hybrid approach for data clustering based on modified cohort intelligence and K-means, *Expert Systems with Applications*, 41(13): 6009–6016.

[10] JC Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[11] Lin Zhu, Longbing Cao, Jie Yang, Jingsheng Lei.[2014] *Evolving soft subspace clustering, Applied Soft Computing*, 14(Part B) :210–228.

[12] Lingras C.[2004] West, Interval set clustering of web users with rough k-means, *Journal of Intelligent Information Systems* 23 (1): 5–16.

[13] Liyong Zhang ,Witold Pedrycz , Wei Lu , Xiaodong Liu , Li Zhang. [2014]" An interval weighed fuzzy c-means clustering by genetically guided alternating optimization", *Expert Systems with Applications* 41: 5960–5971.

[14] LA Zadeh.[(1994)] Fuzzy logic, neural networks, and soft computing, *Communica- tions of the ACM* 37:77–84.

[15] M Emre Celebi, Hassan A Kingravi, Patricio A Vela.[1994] A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications*, 40(1): 200–210.

[16] Matteo Brunelli, József Mezei.[2013] How different are ranking methods for fuzzy numbers? A numerical study, International *Journal of Approximate Reasoning*, 54(5): 627–639.

[17] BK Tripathy, GK Panda and A Mitra. "Covering Based Rough Equality of sets and Comparison of Knowledge", in: Proceedings of the Inter. *Conf. in Mathematics and Computer Science* (ICMCS 2009), 5-6 Jan. 09', Chennai, INDIA,2:438–443.

[18] Pawlak, Z. [1982]"Rough Sets", *Int Jour Inf Comp Sc,* 11:341–356.

[19] Pawlak Z. [1991] Rough Sets: Theoretical Aspects of Reasoning about Data, *Kluwer Academic Publishers.*

[20] Prakash Kumar, BK Tripathy.[2009] MMeR: an algorithm for clustering heterogeneous data using rough set theory,*International Journal of Rapid Manufacturing*, 1(2):189–207.

[21] P Maji, SK Pal. [2007] RFCM: a hybrid clustering algorithm using rough and fuzzy sets, *Fundamenta Informaticae* 79:1–22.

[22] Pierpaolo D'Urso, Riccardo Massari.[2013]Fuzzy clustering of human activity patterns, *Fuzzy Sets and Systems,* 215: 29–54.

[23] S Mitra, H. Banka, W Pedrycz. [2006] Rough–fuzzycollaborativeclustering,IEEE Transactionson Systems, Man, and Cybernetics—Part B36:795–805.

[24] S Mitra, T Acharya. [2003] Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley, New York

[25] Sotirios P Chatzis.[2011] A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional, Expert Systems with Applications, 38(7): 8684-8689.

[26] Tripathy BK, Tripathy HK. [2009]Covering Based Rough equivalence of Sets and Comparison of Knowledge, Proceedings of the IACSIT Spring Conference 2009, Singapore, 303–307.

[27] Tripathy BK, Mitra A and Ojha J. [2008] On Rough Equalities and Rough Equivalences of Sets, RSCTC 2008-Akron, U.S.A., Springer-Verlag Berlin Heidelberg LNAI 5306, 92–102.

[28] Tripathy BK, Tripathy HK. [2009]] Covering Based Rough Equivalence of Sets and Comparison of Knowledge. Computer Science and Information Technology - Spring Conference, 2009.IACSITSC '09. *International Association of* ,303–307, 17–20 April 2009

[29] https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

[30] BKTripathy and K.Govindarajulu. [2014] On Covering Based Pessimistic Multi Granular Rough Sets,2014 Sixth International Conference on Computational Intelligence and Communication networks,978-1-4799-6929-6 (14): 708–713.

[31] BK Tripathy, K. Govindarajulu. [2015] Some more properties of covering based multigranular rough sets, INDIA 2015,Kalayni University, J.K.Mandal et al(Eds),Information system design and applications,Advances in *Intelligent Systems and Computing*,339:555–564.

ABOUT AUTHORS

THE IIOAB JOURNAL
ISSN: 0976-3104

Prof. Prabhavathy Panneer *is working as Assistant Professor (Senior) in School of Information Technology and Engineering,VIT University,Vellore. Her research area includes computational intelligence, data mining, database. She has published 8 journal paper in her research filed. She is life member of CSI and IEEE. She is also part of various school activity committees. She has published number of papers in international conferences.*

Dr. B. K Tripathy *is a senior professor in the school of computing sciences and engineering, VIT University, at Vellore, India. He has been awarded with Gold Medals both at graduate and post graduate levels of Berhampur University, India. Also, he has been awarded with the best post graduate of the Berhampur University. He has received national scholarship, UGC fellowship, SERC visiting fellowship and DOE (Govt. of India) scholarship at various levels of his career. He has published more than 240 technical papers in various international journals, conferences, and Springer book chapters. He has produced 18 PhD's under his supervision. He is associated with many professional bodies like IEEE, ACM,IRSS, WSEAS, AISTC, ISTP, CSI, AMS, and IMS. His name also appeared in the editorial board of several international journals like CTA, ITTA, AMMS, IJCTE, AISS, AIT,and IJPS. His research interest includes fuzzy sets and systems, rough sets and knowledge engineering, data clustering, social network analysis, soft computing and granular computing.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Subhalakshmi et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

TRACKING OF CHNGES IN CANCER DATA USING SOFTCLUSTERING

Chatti Subbalakshmi[1*], G Rama Krishna[2], and S Krishna Mohan Rao[3]

[1]*Guru Nanak Institutions Technical Campus, Dept. of CSE, Hyderabad, Telangana, INDIA*
[2]*K L University, Dept. of CSE, Vijayawada, AP, INDIA*
[3]*Sidhartha Engineering College, Dept. of CSE, Hyderabad, Telangana, INDIA*

ABSTRACT

Data mining is process of extracting the knowledge from large amount of data and its methods are being used efficiently in biological and biomedical applications. Soft computing is for intelligent management systems and its components are fuzzy logic, evaluation computing and genetic algorithms. In recent years, combinations of data mining with soft computing approaches are more suitable for bio-informatics. Clustering is a process of unsupervised learning algorithm in data mining which can be implemented using soft computing approaches. Fuzzy Clustering or soft clustering is based on fuzzy set theory. It groups the data based on partial membership function and it assigns data point more than one cluster. It is used in many biomedical databases, like gene expression, protein sequences; image processing and image segmentation is main step for detection of cancer. In this paper, we proposed dynamic fuzzy clustering algorithm to identify changes in cluster structure. We applied on Wisconsin breast cancer data is collected periodically grouped into eight with class (benign / malignant). We executed fuzzy c-means clustering in individual groups and also executed dynamic fuzzy clustering by incrementally adding instances of groups. We presented our observation of results in both cases which can be support the analysis of cancer state between the instances of class.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 16th –Aug-2015

KEY WORDS

Data mining; soft computing; breast cancer data; dynamic fuzzy clustering;

***Corresponding author:** **Email:** subbalakshmichatti@gmail.com; **Tel.:** +91-40-9032312260

INTRODUCTION

In Data analysis, data mining plays major role in many application databases like business intelligence, science and engineering, bio-informatics, medical analytics. Data mining is part of "Knowledge Discovery in Databases" (KDD) [1] and it is computational process of finding patterns in large databases. It is set of methods with association of artificial intelligence, machine learning, statistics and database systems to handle different types of data. The main data mining tasks are supervised and unsupervised learning methods and the most frequently used unsupervised learning method is cluster analysis [2]; process of grouping most similar objects into clusters depend on similarity measures.

There are many clustering models are present in literature, connectivity models, centroid models, density models, graph-based models. All these algorithms are called hard clustering or exclusive clustering, as they assign data point to only one cluster as it uses conventional crisp set theory. The many complex applications in biology, medicine, the humanities, management sciences require mathematics and analytical methods with uncertain and unpredictable. But these algorithms do not handles application databases with uncertainty, imprecision, partial truth and approximation characteristics, hence soft computing approaches are introduced in cluster analysis to support these databases. The components of soft computing are Fuzzy logic, Neural networks [3], Evolutionary computation [4] and Support Vector Machines [5] and has been most frequently applied successfully in Bioinformatics and Biomedicine in recent years [6].

Fuzzy logic deals with approximate and value ranges in between 0 and 1and it was introduced with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh [7, 8]. It had been applied to several areas, from control areas to intelligence systems. The fuzzy clustering or soft clustering uses fuzzy set theory for grouping data objects and it assigns partial membership value for each data objects to each cluster. Fuzzy c-means is centroid based fuzzy clustering algorithm and it uses degree of membership value to cluster data point [9, 10].

IIOAB JOURNAL ISSN: 0976-3104

In this paper, we have used soft clustering method, i.e. Fuzzy c-means is to group the cancer data. Fuzz c-means takes number of clusters in prior to execution and it has to update as data change. The objective of this paper is, we consider the problem of clustering on cancer data set and identifying the changes in the data as data is added periodically. For that, we selected Wisconsin breast cancer database is collected from UCI repository which was obtained from University of Wisconsin hospitals, Madison from Dr. William H Walberg. They collected breast cancer instances continuous eight months and grouped into eight. Each instance defined by nine features and one class (benign / malignant). We implemented fuzzy c-means clustering on individual groups of instances and then applied dynamic fuzzy c-means algorithm by incrementally adding instances of groups in R data mining software.

The paper is organized as, related work is given in section 2, dynamic fuzzy clustering in section 3, results are given in section 3 and conclusion is mentioned in section 4.

RELATED WORK

Fuzzy clustering

The fuzzy clustering defined by the fuzzy set theory and it is a process of grouping the objects by allowing the concept of partial membership, in which each object can belong to multiple clusters. For all data object, it assigns the membership values between 0 to 1 represents fit in for each cluster and the sum of the membership values of each data objects to all clusters must be 1. The high membership value shows more likely that data object belongs to that cluster. The most widely used fuzzy clustering algorithm is Fuzzy C-Means (FCM) [11].

Given a set of n data objects, $p_k=p_1, p_2, p_3,, p_n$ the algorithm minimizes a weighted within group of sum of squared error an objective function shown in equation (1).

$$J = \sum_{i=1}^{n}\sum_{k=1}^{c} \mu_{ik}^{m} \|p_i - v_k\|^2 \qquad (1)$$

Where J is objective function, n is number of data objects, c is cluster number, μ_{ik} membership value, p_i is data object, v_k is center of cluster k, m is fuzziness factor value always greater than one. The center of k^{th} cluster can be calculated using equation (2) as,

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} p_i}{\sum_{i=1}^{n} \mu_{ik}^{m}} \qquad (2)$$

The fuzzy membership value can calculated using equation (3) as,

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (\|p_i - v_k\| \|p_i - v_j\|)^{\frac{2}{m-1}}} \qquad (3)$$

The fuzzy c-means is objective function-based clustering method, which takes cluster number determined before the execution of algorithm. The deficiency of algorithm is, it does not identify noise and outliers as it uses sum of square error objective function. The second deficiency FCM, defined by Krishnapuram and Keller is that due to the constraint on membership shows as degree of sharing, but not as degree of possibility of a point belonging to a class. Mainly it deals with similarity between perfectly described objects, i.e. all feature values are exactly known and it does not deal with the uncertainty included by missing or incorrect data.

Cluster validity

The internal cluster validity can be done by using Silhouette cluster validity index is defined as [12, 13]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (4)$$

Where *a (i)* is average dissimilarity between data point (*i*) *and* all other data within the similar cluster and *b(i)* is the minimum average dissimilarity of *i* to all other cluster. The positive value of s(i) indicates the correct clustering and negative value shows the incorrect clustering.

THE IIOAB JOURNAL ISSN: 0976-3104

DYNAMIC FUZZY CLUSTERING

The main objective of dynamic fuzzy clustering is to revise the initial parameters of the algorithm when new data is added. Fuzzy c-means clustering algorithm requires number of clusters as an input value and this value is defined by data size. When data size changes number of clusters might be changes. Therefore, for each new incoming data cycle of dynamic fuzzy clustering as illustrate in **Figure– 1**.

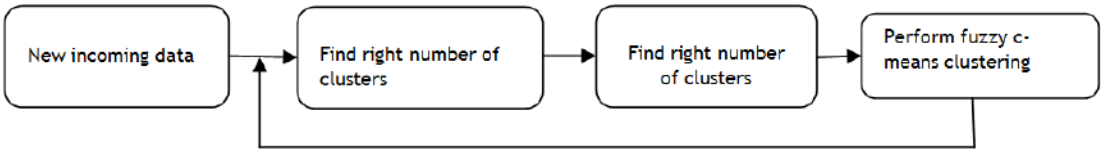


Fig: 1. Cycle of dynamic fuzzy clustering algorithm.

The general steps of algorithm are given as:

Initial clustering:

Step1: find the size of data set *n* for $D_{initial}$;
Step 2: for n= 1 to n-1/2
Calculate cluster average silhouette width S for each cluster;
Step 3: find c with maximum cluster average silhouette width;
Step 4: execute fuzzy c-means clustering algorithm;

For each new incoming data D_{new}

Step 5: add new data with old data as,
$D = D_{inital} + D_{new}$
Repeat step1 to step4
In this algorithm, we are adding the data incrementally and finding the cluster changes in data.

RESULTS

We have used Wisconsin breast cancer database is collected from UCI repository which was obtained from University of Wisconsin hospitals, Madison from Dr. William H Walberg. They collected samples periodically and made into eight groups as:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated database on 15 July 1992)
Each instance consists of 9 attributes and one class (benign/malignant) with range of values as:

#	Attribute	Domain
1.	Clump Thickness	1 - 10
2.	Uniformity of Cell Size	1 - 10
3.	Uniformity of Cell Shape	1 - 10
4.	Marginal Adhesion	1 - 10
5.	Single Epithelial Cell Size	1 - 10
6.	Bare Nuclei	1 - 10
7.	Bland Chromatin	1 - 10
8.	Normal Nucleoli	1 - 10
9.	Mitoses	1 – 10
10.	Class	2 for benign and 4 for malignant

THE IIOAB JOURNAL ISSN: 0976-3104

Results of fuzzy c-means on individual groups of cancer data

For each group of data, we executed fuzzy c-means clustering after finding right number of clusters using silhouette cluster validity index. Our observations are for all individual eight groups of data resulting only two right clusters which indicate only two class values (benign/malignant) exits. From the results of fuzzy c-means on group1 data, some of the instance's membership values partially belongs to benign and malignant. But in next seven groups, the Dunn's partition coefficient indicates, there is no overlapping instance which shows instance belongs to any one class. From the result of eight groups, we have identified that silhouette width gradually increasing, i.e. instances are very close to each other with same characteristic. The comparative results between each group are given in **Table–1**, the centers of eight group clusters are given **Table–2** and for each group fuzzy c-means cluster centers are shown in **Figure –2**.

Table: 1. Results of fuzzy c-means clustering algorithm for eight groups of cancer data

	Group 1 with 367 instances	Group 2 with 70 instances	Group 3 with 31 instances	Group 4 with 17 instances	Group 5 With 48 Instances	Group 6 With 49 instances	Group 7 With 31 Instances	Group 8 With 86 instances
Clus Avg Sil width	0.69340811	0.87437541	0.8765545	0.9372129	0.8434955	0.8871365	0.7294244	0.8878145
Right c value	2	2	2	2	2	2	2	2
Dunn_coeff	0.8041438	0.9035008	0.904723	0.9583406	0.897402	0.9212774	0.8312131	0.9178934
Belgian	207	14	8	3	11	09	19	74
Malignant	160	56	23	14	37	40	12	12

Table: 2. Fuzzy c-means cluster centers of eight groups of cancer instances

	Cluster centers	Clump thickness	Uniformity cell size	Uniformity cell shape	Marginal adhesion	Single Epithelial Cell Size	Bare nuclei	Bland chromatin	Normal nucleoli	Mitoses
Group 1	Cluster1	2.160705	2.838931	1.428396	1.54586	1.368898	2.23415	1.551031	2.635899	1.456337
	Cluster 2	3.898318	7.367233	6.570298	6.643348	5.579655	5.680445	8.065509	5.63092	6.320495
Group 2	Cluster1	2.015703	2.707998	1.578677	1.718243	1.390675	2.164551	1.185944	1.909195	1.315732
	Cluster 2	3.831131	8.082565	8.120871	7.860144	5.168349	5.772975	8.372444	6.79263	5.386354
Group 3	Cluster1	3.9925	7.772094	6.34029	6.367802	6.603301	5.100012	8.420575	8.448891	7.173032
	Cluster 2	2.061884	3.835446	1.178616	1.313346	1.882224	2.008565	1.580559	1.127177	1.069119
Group 4	Cluster1	3.997523	7.093996	8.31591	8.357888	8.232472	6.374376	8.301134	8.709004	8.441122
	Cluster 2	2.003191	4.294436	1.218299	1.290711	1.355207	1.791213	1.012295	1.150615	1.009591
Group 5	Cluster1	3.947392	6.252951	7.346187	7.155194	7.63559	4.293177	8.736356	6.53448	4.95016
	Cluster 2	2.063439	3.20521	1.211239	1.190368	1.313407	1.837813	1.287205	1.704877	1.044458
Group 6	Cluster1	2.017287	3.464124	1.237472	1.303173	1.165077	1.956136	1.217999	2.229199	1.198536
	Cluster 2	3.977982	7.076954	7.993649	7.721948	6.686502	4.84257	8.063018	7.784607	6.933361
Group 7	Cluster1	2.088273	3.903984	1.265267	1.599191	1.55003	2.040429	1.259187	1.774887	1.156555
	Cluster 2	3.99606	6.438842	7.514016	7.442704	7.437029	5.106654	7.721833	7.746245	5.172701
Group 8	Cluster1	2.013597	2.863381	1.216343	1.392772	1.305293	2.085176	1.162901	1.657168	1.114918

THE IIOAB JOURNAL
ISSN: 0976-3104

	Cluster 2	3.93992	5.889319	9.011659	8.241414	6.810015	5.551078	5.622764	7.155445	3.05486

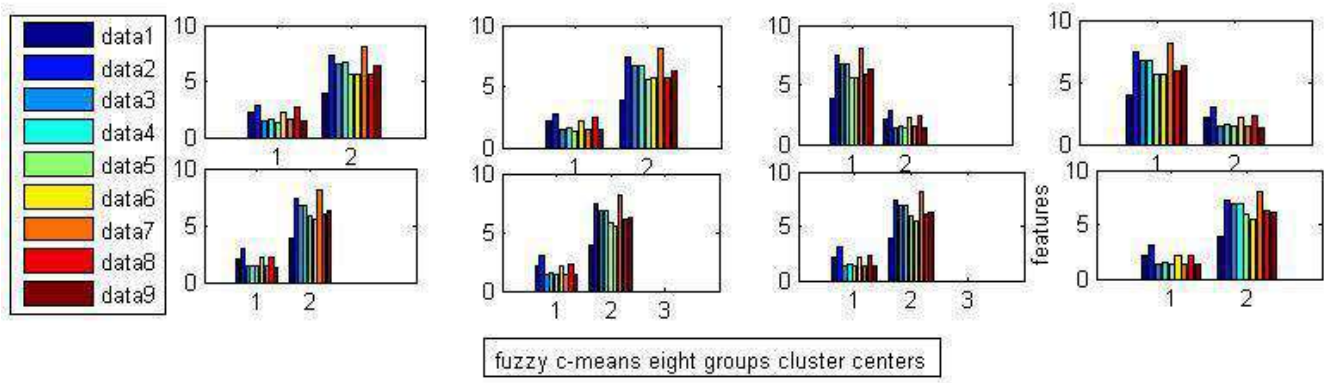


Fig: 2. Fuzzy c-means cluster centers.

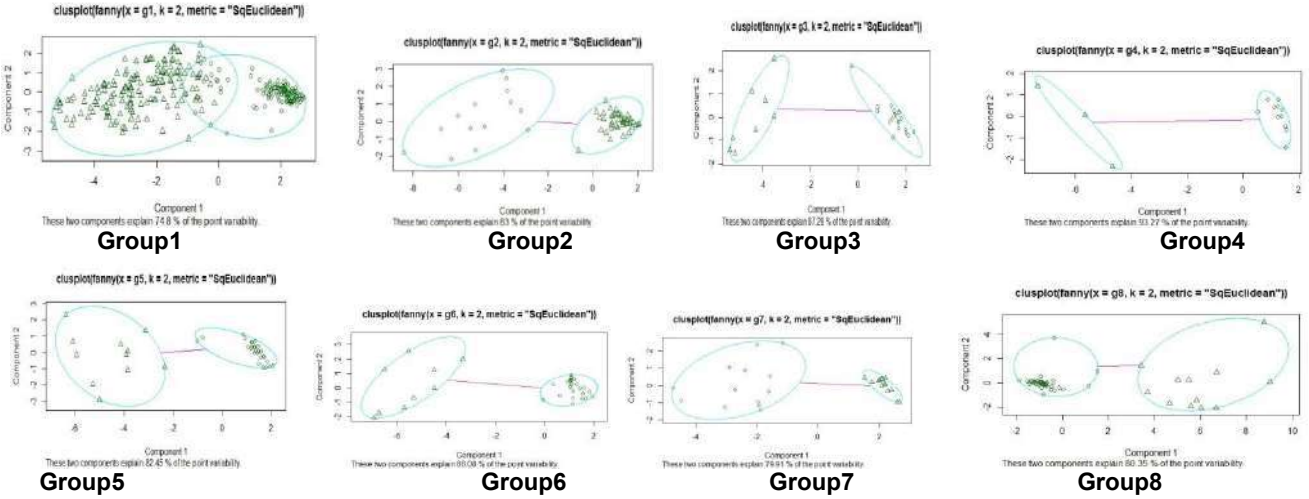


Fig: 3. The outputs of fuzzy c-means cluster points.

Results of dynamic fuzzy clustering algorithm

We have executed initial clustering on group1 data with 316 instances and incrementally we added remaining group of data in the next cycles. The results are given **Table– 3** and clustering points

Table: 3. Results of dynamic fuzzy c-means clustering algorithm

	Cycle-1	Cycle-2	Cycle-3	Cycle-4	Cycle-5	Cycle-6	Cycle-7	Cycle-8
Data size	367	437	468	485	533	582	613	697
Clus Avg Sil width	0.69340811	0.7204372	0.72818879	0.73489439	0.7443762	0.7564407	0.7547007	0.7687378
Right c value	2	2	2	2	2	2	2	2
Dunn_coeff	0.8041438	0.8187395	0.8222193	0.825793	0.8315249	0.838496	0.8375065	0.8455845
Belgian	207	263	286	301	338	378	396	469
Malignant	160	174	182	184	195	204	217	228

THE IIOAB JOURNAL
ISSN: 0976-3104

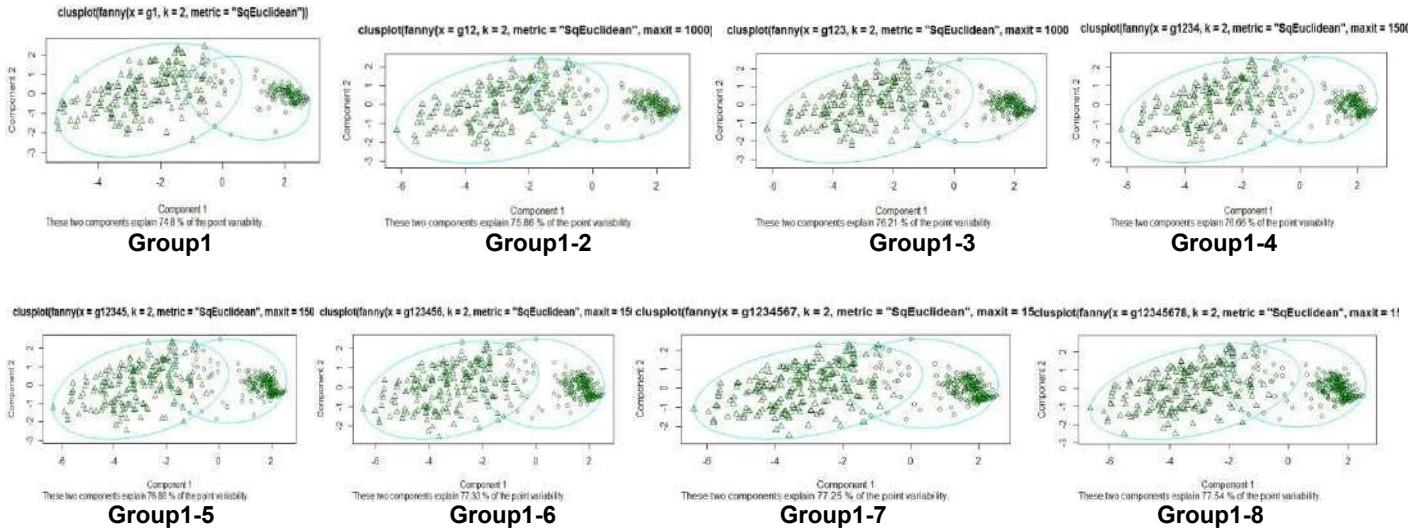


Fig: 4. For each cycle of dynamic fuzzy c-means cluster points.

Table: 4. Dynamic fuzzy c-means cluster centers of eight groups of cancer instances

	Cluster centers	Clump thickness	Uniformity cell size	Uniformity cell shape	Marginal adhesion	Single Epithelial Cell Size	Bare nuclei	Bland chromatin	Normal nucleoli	Mitoses
Group1	Cluster1	2.160703	2.838925	1.428393	1.545856	1.368896	2.234148	1.551025	2.635896	1.456334
	Cluster 2	3.898317	7.367231	6.570284	6.643336	5.579643	5.680435	8.065507	5.630912	6.320481
Group 2	Cluster1	2.129815	2.810183	1.459899	1.582505	1.373803	2.219787	1.473415	2.4792	1.425762
	Cluster 2	3.892513	7.433734	6.707197	6.747639	5.537787	5.68313	8.089441	5.734023	6.247491
Group 3	Cluster1	3.897594	7.450822	6.688592	6.727644	5.597415	5.651275	8.105787	5.878557	6.303923
	Cluster 2	2.12504	2.890006	1.44051	1.563888	1.413542	2.205181	1.483058	2.374672	1.399523
Group 4	Cluster1	3.899821	7.438435	6.727208	6.764603	5.658186	5.668592	8.114747	5.939162	6.355033
	Cluster 2	2.120565	2.96046	1.431706	1.553691	1.411136	2.187144	1.464081	2.318261	1.382631
Group 5	Cluster1	2.114216	2.987723	1.408163	1.513945	1.40153	2.148456	1.443685	2.249238	1.344811
	Cluster 2	3.902375	7.367682	6.754794	6.778681	5.773704	5.585881	8.15409	5.975062	6.275716
Group 6	Cluster1	2.104058	3.040041	1.3894	1.492056	1.376849	2.127336	1.41911	2.247866	1.3284
	Cluster 2	3.906056	7.353648	6.818226	6.82056	5.813291	5.557155	8.14977	6.058938	6.309007
Group 7	Cluster1	2.104699	3.08218	1.387464	1.500128	1.386206	2.125459	1.414027	2.229006	1.322869
	Cluster 2	3.911752	7.302867	6.859551	6.859199	5.917126	5.529439	8.137364	6.16135	6.248692
Group 8	Cluster1	2.091352	3.049704	1.365231	1.486661	1.37489	2.12169	1.375652	2.142978	1.290061
	Cluster 2	3.912348	7.235365	6.968129	6.930414	5.972788	5.531019	8.01619	6.215421	6.08946

DISCUSSIONS

In The Wisconsin breast cancer database consists of eight groups of cancer instances with one class. These instances are already classified into benign or malignant. In our experiment, initially we have done fuzzy c-means on individual groups. For that, we removed the class filed and we applied fuzzy c-means clustering on nine features. From the results it shows that all groups of instance are clustered into only two groups and it was similar to classification according database. We identified that in group1 instances having overlapping, i.e. some instances partially belongs both groups. But in next groups does not have overlapping instances and they almost belong to one group. The cancer instances in overlapping portion between classes, it is difficult to find state of

THE IIOAB JOURNAL ISSN: 0976-3104

patient. In next step, we executed dynamic fuzzy c-means where we added gradually groups of instances and every time we checked the number of clusters. In all cycles, instances are grouped into two and overlapping between them. We identified the changes in class centers and silhouette widths. With these results, one can identify the similarity between the breast cancer patients within the same group and other group patients.

CONCLUSION

We consider the problem of tracking of changes in cancer data using soft clustering algorithm. The clustering can be performed using hard computing or soft computing approaches. Hard clustering is not efficient to handle impression, uncertainty, partial truth and approximation data set. Soft computing approaches successful in handling this type of data and it supports many complex applications. We apply the fuzzy c-means clustering or soft clustering on Wisconsin breast cancer data on individual groups, after that for each cycle incrementally added group cases and apply dynamic fuzzy c-means clustering algorithm. We projected results in each case which can be useful for finding the comparison between group instances. This method can be implemented using any other soft clustering methods to improve the results.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

REFERENCES

[1] Fayyad Usama, Piatetsky-Shapiro, Gregory Smyth, Padhraic. [1996] From Data Mining to Knowledge Discovery in Databases.

[2] Data Mining Curriculum ACM SIGKDD. 2006-04–30.

[3] Ferreira C. [2006] Designing Neural Networks Using Gene Expression Programming, Applied Soft Computing Technologies: *The Challenger of Complexity*, 517–536, *Springer-Verlag*.

[4] AE Eiben and M Schoenauer. [2002*] Evolutionary computing, Information Processing Letters*, 82(1):1–6,

[5] Zadeh Lotfi A.[1994] Fuzzy Logic, Neural Network, and Soft Computing, Communication of the ACM, March 1994, 37 (3): 77–84.

[6] Yudong Zhang, Saeed Balochian, Vishal Bhatnagar. [2014] "Emerging Trends in Soft Computing Models in Bioinformatics and Biomedicine". *The Scientific World Journal* 3.doi:10.1155/2014/683029.

[7] Fuzzy Logic. *Stanford Encyclopedia of Philosophy*. 2006-07–23.

[8] Zedeh LA. [1965] Fuzzy Set . *Information and Control* 8 (3): 338–353.

[9] J Bezdek, S Pal. [1992] Fuzzy models for pattern recognition, IEEE press, New York ,

[10] H- J Zimmermam, Fuzzy set theory and it applications, [1991].

[11] Nock R, Nielsen F. [2006] On Weighting Clustering, *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 28 (8): 1–13.

[12] Peter J Rousseeuw. [1987] "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*20: 53–65.

[13] Chatti Subbalakshmi. [2015] A method to find optimal number of clusters based on fuzzy silhouette on dynamic dataset. *Elsevier Science Direct*, 2015,doi: 10.1016 /j.procs.2015.02.030.

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Nagaraju and Tripathy.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

APPROXIMATE EQUALITIES FOR COVERING BASED OPTIMISTIC MULTI GRANULAR ROUGH SETS AND THEIR PROPERTIES

M. Nagaraju* and B. K. Tripathy

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, INDIA

ABSTRACT

Imprecision in modern day data has become a common feature and in order to efficiently handle them many uncertainty based models have been put forth in the literature. Rough set model introduced by Pawlak has established itself to be an efficient model in many real life situations. But the basic rough set model of Pawlak has limited applications because of the constraint of it being dependent on equivalence relations. The equivalent mathematical concept to equivalence relation is that of a partition. A cover is a generalization of the notion of partition and this led to the development of covering based rough sets, which has better modeling power than basic rough sets. Following the concepts of granular computing rough set introduced by Pawlak is single granulation. So, in order to handle multi-granularity, two types of multi-granularities called optimistic multi-granulation and pessimistic multi-granulation were introduced in 2006 and 2010 respectively. Recently these two concepts of multi-granulation and covering based rough sets were combined to define covering based multi-granular rough sets. The equality of sets in mathematics is too redundant to have any fruitful real life application as it does not include user knowledge into it, which is normally done in practice. In order to handle this rough equalities were defined by Pawlak et al, which was extended by Tripathy in2008 to define rough equivalence. In this paper we introduce and study covering based optimistic multi-granular approximate equalities and study their properties. We study two types of properties called general properties and replacement properties. A real life example is used for illustration of the concepts and also to aid in the construction of counter examples in the proofs of the properties.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 20th –Aug-2015

KEY WORDS

Rough Sets, Granulation, Multigranulation, Cover, Covering Based Optimistic Multi-Granular Rough Sets (CBOMGRS).

***Corresponding author: Email:** mnagaraju@vit.ac.in; **Tel.:** +91-9894803487

INTRODUCTION

Data in real life are mostly imprecise in nature and so the conventional tools for formal modeling, reasoning and computing, which are crisp, deterministic and precise in characteristics, are inadequate to handle them. This gives rise to the development of several imprecise models, of which rough sets introduced by Pawlak [5, 6] is one of the most efficient one. It has been proved to be very efficient to capture impreciseness in data. According to Pawlak, the knowledge of human beings depends upon their capability to classify objects of universes. Equivalence relations on any universe induce classifications through the equivalence classes associated with them. So, Pawlak had taken equivalence relations to define rough sets and related notions.
A pair of crisp sets called the lower and upper approximations are associated with every rough set. Lower approximation comprising of elements certainly belong to it and upper approximation comprising of elements certainly or possibly belong to it, with respect to the available information.

This basic rough set has been extended further in many directions. These extensions are actually either based on tolerance relations or any such relations that do not require the stringent restrictions of an equivalence relation.

From the point of view of granular computing, basic rough set theory deals with a single granulation [17]. However, in some application areas we need to handle more than one granulation at a time and this necessitated the development of multi-granular rough sets (MGRS)[7], where at least two equivalence relations are taken for granulation of a universe. This concept is further extended by considering covers and this lead to the development of covering based multi granular rough sets(CBMGRS). They are of two types namely, optimistic and pessimistic. In this paper optimistic one is considered. Four types of CBOMGRS are defined and their properties, general and replacement, are established.

THE IIOAB JOURNAL ISSN: 0976-3104

Equality of sets in mathematics is a very stringent notion and its application is also limited. In real life situations we use our own knowledge about the universe of discourse to determine the equality of sets. However, such interception of user in deciding the equality of sets is very much expected. In order to make a place for user knowledge in deciding the equality of sets fully or partially, three kinds of rough equalities were introduced by Novotny and Pawlak [5,6]. In fact three notions were introduced, called the top rough equality, the bottom rough equality and the rough equality. Now the sets can be equal or not from the user point of view. They had established several properties of these notions. They tried to interchange the concepts of top rough equality and bottom rough equality in the properties to find their validity and commented that these properties do not hold true under such circumstances. .

This paper is organized into four sections. First section provides the over view and related literatures. Section two presents various definitions and notions required. Section three introduces multi granular rough equalities and its general and replacement properties. In this section real life examples are considered to prove few properties as sample. In section four conclusion to the work are presented.

DEFINITIONS AND NOTATIONS

Rough set

The notion of rough set was introduced by Z.Pawlak in the year 1982 ([5]). We extract the definition and present below.

Let *U* be a universe of discourse and *R* be an equivalence relation over *U*. By *U/R* we denote the family of all equivalence classes of *R*, referred to as categories or concepts of *R* and the equivalence class of an element $x \in U$ is denoted by $[x]_R$. By a knowledge base, we understand a relational system $K = (U, P)$, where *U* is as above and P is a family of equivalence relations over *U*. For any subset Q ($\neq \phi$) $\subseteq$ P, the intersection of all equivalence relations in Q is denoted by IND(Q) and is called the indiscernibility relation over Q. Given any $A \subseteq U$ and R $\in$ IND (K), we associate two subsets, $\underline{R}A = \bigcup\{B \in U / R : B \subseteq A\}$ and $\overline{R}A = \bigcup\{B \in U / R : B \bigcap A \neq \phi\}$, called the R-lower and R-upper approximations of 'A' respectively. The R-boundary of 'A' is denoted by $BN_R(A)$ and is given by $BN_R(A) = \overline{R}A - \underline{R}A$. The elements of $\underline{R}$ A are those elements of U, which can certainly be classified as elements of A, and the elements of $\overline{R}$ A are those elements of U, which can possibly be classified as elements of 'A', employing knowledge of *R*. We say that A is rough with respect to R if and only if $\underline{R}A \neq \overline{R}A$, equivalently $BN_R(A) \neq \phi$. 'A' is said to be R-definable if and only if $\underline{R}A = \overline{R}A$, or $BN_R(A) = \phi$.

Covering based rough sets

Basic rough sets introduced by Pawlak have been extended in many ways. One such extension is the notion of covering based rough sets, where the notion of partitions is replaced by the general notion of covers [16]. In this section we introduce the basics of these sets.

Definition 2.2.1: ([23, 25]) Let U be a universe and C = {C_1,C_2,….., C_n} be a family of non-empty subsets of U that may be overlapping in nature. If $\bigcup C$ = U, then C is called a covering of U. The pair (U, C) is called covering approximation space. For any A $\subseteq$ U, the covering lower and upper approximations of 'A' with respect to C can be defined as follows

(2.2.1) $\underline{C}(A) = \bigcup\{C_i \subset A, i \in \{1, 2, ..., n\}\}$

(2.2.2) $\overline{C}(A) = \bigcup\{C_i \bigcap A \neq \phi, i \in \{1, 2, ..., n\}\}$

The pair $(\underline{C}(A), \overline{C}(A))$ is called covering based rough set associated with X with respect to cover C if $\underline{C}(A) \neq \overline{C}(A)$, i.e., A is said to be roughly definable with respect to C. Otherwise A is said to be C-definable.

Definition 2.2.2: ([23,25]) Given a covering approximation space (U, C) for any x $\in$ U, sets $md_c(x)$ and $MD_c(x)$ are respectively called minimal and maximal descriptors of x with respect to C,

(2.2.3) $md_c(x) = \{S \in C / x \in S \text{ and } (\forall T \in C \text{ if } (x \in T \text{ and } T \subseteq S) \text{ then } S = T)\}$

It is a set of all minimal covers containing x where a minimal cover containing x be one for which no proper sub cover containing x exists.

(2.2.4) $MD_c(x) = \{S \in C / x \in S \text{ and } (\forall T \in C \text{ if } (x \in T \text{ and } T \supseteq S) \text{ then } S = T)\}$

THE IIOAB JOURNAL ISSN: 0976-3104

It is a set of all maximal covers containing x where a maximal cover containing x be one for which no proper super cover containing x exists.

Multi granular rough sets

In the view of granular computing (proposed by L. A. Zadeh), an equivalence relation on the universe can be regarded as a granulation, and a partition on the universe can be regarded as a granulation space [5, 6]. For an incomplete information system, similarly, a tolerance relation on the universe can be one regard as a granulation, and a cover induced by the relation can be regarded as a granulation space. Several measures in knowledge base closely associated with granular computing, such as knowledge granulation, granulation measure, information entropy and rough entropy. On research of rough set method based on multi-granulations, Y. H. Qian and J. Y. Liang introduced a rough set model based on multi-granulations [7], which is established by using multi equivalence relations.

Definition 2.3.1 ([23, 25]) Let K= (U, **R**) be a knowledge base, **R** be a family of equivalence relations, T, S $\in$ **R.** We define the optimistic multi-granular lower approximation and upper approximation of X in U as

$$(2.3.1)\quad \underline{T+S}(A) = \bigcup\{x / [x]_T \subseteq A \text{ or } [x]_S \subseteq A\} \quad \text{and}$$

$$(2.3.2)\quad \overline{T+S}(A) = \underline{(T+S}(A^C))^C$$

Covering based optimistic multi granular rough sets

The notion of Multi-granular rough sets have been extended to covering approximation spaces. They can be of two types; namely, optimistic and pessimistic. By employing minimal and maximal descriptors four types of CBOMGRS are possible. The definitions of four types of CBOMGRS are as follows [4].

Let (U, C) be a covering approximation space, C_1 and C_2 be two covers in C and A be any subset of U, The four types of optimistic covering based optimistic multi granular rough sets, are defined as follows.

Definition 2.4.1([16]): The first type CBOMGRS lower and upper approximations with respect to C_1 and C_2 are defined as

$$(2.4.1)\quad \underline{O_{C_1+C_2}}(A) = \{x \in U / \cap md_{C_1}(x) \subseteq A \text{ or } \cap md_{C_2}(x) \subseteq A\} \text{ and}$$

$$(2.4.2)\quad \overline{O_{C_1+C_2}}(A) = \{x \in U / (\cap md_{C_1}(x)) \cap A \neq \phi \text{ and } (\cap md_{C_2}(x)) \cap A \neq \phi\}$$

Definition 2.4.2([16]): The second type CBOMGRS lower and upper approximations with respect to C_1 and C_2 are defined as

$$(2.4.3)\quad \underline{S_{C_1+C_2}}(A) = \{x \in U / \cup md_{C_1}(x) \subseteq A \text{ or } \cup md_{C_2}(x) \subseteq A\} \text{ and}$$

$$(2.4.4)\quad \overline{S_{C_1+C_2}}(A) = \{x \in U / (\cup md_{C_1}(x)) \cap A \neq \phi \text{ and } (\cup md_{C_2}(x)) \cap A \neq \phi\}$$

Definition 2.4.3([16]): The third type CBOMGRS lower and upper approximations with respect to C_1 and C_2 are defined as

$$(2.4.5)\quad \underline{T_{C_1+C_2}}(A) = \{x \in U / \cap MD_{C_1}(x) \subseteq A \text{ or } \cap MD_{C_2}(x) \subseteq A\} \text{ and}$$

$$(2.4.6)\quad \overline{T_{C_1+C_2}}(A) = \{x \in U / (\cap MD_{C_1}(x)) \cap A \neq \phi \text{ and } (\cap MD_{C_2}(x)) \cap A \neq \phi\}$$

Definition 2.4.4([16]): The fourth type CBOMGRS lower and upper approximations with respect to C_1 and C_2 are defined as

$$(2.4.7)\quad \underline{L_{C_1+C_2}}(A) = \{x \in U / \cup MD_{C_1}(x) \subseteq A \text{ or } \cup MD_{C_2}(x) \subseteq A\} \text{ and}$$

$$(2.4.8)\quad \overline{L_{C_1+C_2}}(A) = \{x \in U / (\cup MD_{C_1}(x)) \cap A \neq \phi \text{ and } (\cup MD_{C_2}(x)) \cap A \neq \phi\}$$

Properties of covering based optimistic multi granular rough sets

The following are the properties of covering based optimistic multi granular rough sets. Here 'w' denotes any of the four types first, second, third or fourth of optimistic multigranulation. Let A and B be any two subsets of U. We omit the proofs of these properties as these are more or less trivial. The proofs can also be found in [15,16].

$$(2.5.1)\quad A \subseteq B \Rightarrow \underline{W_{C_1+C_2}}(A) \subseteq \underline{W_{C_1+C_2}}(B)$$

$$(2.5.2)\quad A \subseteq B \Rightarrow \overline{W_{C_1+C_2}}(A) \subseteq \overline{W_{C_1+C_2}}(B)$$

$$(2.5.3)\quad \underline{W_{C_1+C_2}}(\sim A) = \sim \overline{W_{C_1+C_2}}(A)$$

$$(2.5.4)\quad \overline{W_{C_1+C_2}}(\sim A) = \sim \underline{W_{C_1+C_2}}(A)$$

THE IIOAB JOURNAL ISSN: 0976-3104

(2.5.5) $\underline{W_{C_1+C_2}}(A \cup B) \supseteq \underline{W_{C_1+C_2}}(A) \cup \underline{W_{C_1+C_2}}(B)$

(2.5.6) $\overline{W_{C_1+C_2}}(A \cup B) \supseteq \overline{W_{C_1+C_2}}(A) \cup \overline{W_{C_1+C_2}}(B)$

(2.5.7) $\underline{W_{C_1+C_2}}(A \cap B) \subseteq \underline{W_{C_1+C_2}}(A) \cap \underline{W_{C_1+C_2}}(B)$

(2.5.8) $\overline{W_{C_1+C_2}}(A \cap B) \subseteq \overline{W_{C_1+C_2}}(A) \cap \overline{W_{C_1+C_2}}(B)$

RESULTS

Approximate equalities

The equality of sets or domains used in mathematics is too stringent. In most of the real life situations we often consider equality of sets or domains, as approximately equal under the existing circumstances. These existing circumstances serve as user knowledge about the set or domain. So, they play a significant role in approximate reasoning. Also, one can state that it mostly depends upon the knowledge the assessors have about the set of domains under consideration as a whole but not on the knowledge about individuals of the sets or domains.

As a step to include user knowledge in considering likely equality of sets, Novotny and Pawlak [5,6] introduced the following rough equalities of two sets A and B which are subsets of U.

Let K= (U, R) be a knowledge base, $A, B \subseteq U$ *and* $R \in IND(K)$.

Definition 3.1: We say that,

(3.1.1) A and B are bottom rough equal (A $\underline{\underline{R}}$ B) if and if only $\underline{R}A = \underline{R}B$.

(3.1.2)A and B are top rough equal (A $\overline{\overline{R}}$ B) if and if only $\overline{R}A = \overline{R}B$.

(3.1.3)A and B are rough equal (A R_eq B) if and if only $\underline{R}A = \underline{R}B$ and $\overline{R}A = \overline{R}B$ i.e., (A $\underline{\underline{R}}$ B) and (A $\overline{\overline{R}}$ B).

There are several properties of these approximate equalities established by Novotny and Pawlak in the form of general and replacement properties. The replacement properties are those properties obtained from the general properties by interchanging the top and bottom equalities. As noted by them, all these approximate equalities of sets are relative in character; that is, sets are equal or not equal from our point of view depending on what we know about them. So, in a sense the definition of rough equality incorporates user knowledge about the universe in arriving at equality of sets or domains. However, these notions of approximate equalities of sets boil down to equality of sets again. So, in order to make the equalities more general, a notion called rough equivalences was introduced by Tripathy in 2008 [16]. These notions are more general and more applicable in real life situations. An example of cattle in a society is taken by him to explain the drawbacks in the earlier notion and also to establish the superiority of the new notions in the real life scenario. These two different forms of approximate equalities have been generalized to the context of multi-granulation by Tripathy along with coauthors in a series of papers [14-17, 24-26].

In this paper we shall introduce the concepts of approximate equalities and rough equivalence to the context of covering based optimistic multi granulations and prove their properties (both general and replacement). We establish both the direct as well as the replacement properties for both these notions. In fact two types of covering based multi granular rough equalities and equivalences are possible, namely, optimistic and pessimistic. In this paper optimistic ones are considered. First type covering based optimistic multi granular rough set is considered and its rough equalities and equivalences are studied. The direct properties of such sets are stated and proved first. Later its replacement properties are also studied and proved. To substantiate better understanding of these concepts few of these properties are studied and interpreted in terms of one real life example.

Covering based optimistic multi granular approximate equalities

We introduce in the following the different covering based optimistic multi granular rough equalities for first type, CBOMGRS, and study their properties. For the other types of multi granulations similar definitions hold good.

THE IIOAB JOURNAL ISSN: 0976-3104

Let C_1 and C_2 be two covers on U and let O denotes first type of CBOMGRS.

Definition 3.2: We say that,

(3.2.1) A and B are bottom C_1+C_2 rough equal to each other (A $\overset{=}{}_{C_1+C_2}$ B) iff $\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B)$.

(3.2.2) A and B are top C_1+C_2 rough equal to each other (A $\overset{C_1+C_2}{=}$ B) iff $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B)$.

(3.2.3) A and B are optimistic total rough equal to each other with respect to C_1 and C_2 (A $r\_C_1+C_2\_eq$ B) iff $\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B)$ and $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B)$.

Properties for first type of covering based optimistic multi granular approximate equalities

The general properties of first type of covering based rough equalities are stated, proved and substantiated with few proofs and examples wherever necessary.

Let C_1 and C_2 be two covers on U and $C_1, C_2 \in C$ *and* $A, B \subseteq U$. Let F denotes first type CBOMGRS. Then

(3.3.1) A $\overset{=}{}_{C_1+C_2}$ B if $A \cap B$ $\overset{=}{}_{C_1+C_2}$ A and B both. But the converse may not be true in general.

Proof: $Given \quad A \cap B \quad \overset{=}{}_{C_1+C_2} \quad A \Rightarrow \underline{O_{C_1+C_2}}(A \cap B) = \underline{O_{C_1+C_2}}(A)$ and

$Given \quad A \cap B \quad \overset{=}{}_{C_1+C_2} \quad B \Rightarrow \underline{O_{C_1+C_2}}(A \cap B) = \underline{O_{C_1+C_2}}(B)$

From the above two expressions we have

$\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B) \Rightarrow$ A $\overset{=}{}_{C_1+C_2}$ B.

For the converse part logical equivalence of the statements $(a \wedge b) \vee (c \wedge d)$ and $(a \vee c) \wedge (b \vee d)$, where a, b, c and d are any four logical statements. However, from their truth values we find that these two statements are not equivalent to each other in the following case.

a	B	c	d
True	False	False	True
False	True	True	False

So, examples can be provided which satisfy any of the above cases to show that the converse is not true.

(3.3.2) A $\overset{C_1+C_2}{=}$ B if $A \cup B$ $\overset{C_1+C_2}{=}$ A and B both. The converse may not be true in general.

Proof: $\text{Given} \quad A \text{UB} \quad \overset{C_1+C_2}{=} \quad A \Rightarrow \overline{O_{C_1+C_2}}(\text{AUB}) = \overline{O_{C_1+C_2}}(\text{A})$ and

$\text{Given} \quad A \text{UB} \quad \overset{C_1+C_2}{=} \quad B \Rightarrow \overline{O_{C_1+C_2}}(\text{AUB}) = \overline{O_{C_1+C_2}}(\text{B})$

From the above two expressions we have

$\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B) \Rightarrow$ A $\overset{C_1+C_2}{=}$ B.

The converse part is not true as in property 1. We note that the truth of the converse depends upon the logical equivalence of the two statements, $(a \vee b) \wedge (c \vee d)$ and $(a \wedge c) \vee (b \wedge d)$. However, we find the statements quoted are not true in the following cases.

THE IIOAB JOURNAL ISSN: 0976-3104

a	b	c	d
True	False	False	True
False	True	True	False

So, examples can be constructed such that the above two cases occur to show that the converse part does not hold.

(3.3.3) A $\overset{C_1+C_2}{=} A'$ and B $\overset{C_1+C_2}{=} B'$ may not imply that $A \cup B \overset{C_1+C_2}{=} A' \cup B'$

Proof: The converse part is not true as in property 1. We note that the truth of the converse depends upon the logical equivalence of the two statements, $(a \vee b) \wedge (c \vee d)$ and $(a \wedge c) \vee (b \wedge d)$. However, we find the statements quoted are not true in the following cases.

a	B	c	d
True	False	False	True
False	True	True	False

So, examples can be constructed such that the above two cases occur to show that the converse part does not hold.

(3.3.4) A $\overset{=}{C_1+C_2} A'$ and B $\overset{=}{C_1+C_2} B'$ may not imply that $A \cap B \overset{=}{C_1+C_2} A' \cap B'$

Proof: Let us consider the following real life example to prove the above property.

Example 1: Consider the following data table. Let us consider 3 columns of it, such as, Faculty name, Roles and Project Numbers. Roles column specifies different roles each faculty play in the school, such as, Program chair-1, Division chair-2, and Year co-ordinator-3. Project Numbers column specifies number of the project on which faculty works on.

Table:1. Faculty Information

S.No.	Faculty Name	Division	Collection. Experience (yrs)	Distribution Experience (yrs)	Sex	Roles	Project Numbers
1	Alia –x_1	1	2	0	Female	1	1
2	Brinda-x_2	2	1	0	Female	3	2
3	Cris-x_3	1	3	3	Male	2	2
4	Danya-x_4	2	1	1	Male	2	3
5	Esha-x_5	1	3	3	Female	1, 2	2
6	Feroz-x_6	2	3	0	Male	2	1, 3
7	Gokul-x_7	1	1	4	Male	3	4
8	Harsha-x_8	2	2	4	Male	3	4

THE IIOAB JOURNAL ISSN: 0976-3104

Based on roles and project number columns two sets of covers are obtained as given below.

Let $U = \text{Set of faculties} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ and the following two covers $C_1 and\ C_2$, are generated as given below.

U / C_1 = Covers obtained based on roles of faculties $= \{\{x_1, x_5\}, \{x_3, x_4, x_5, x_6\}, \{x_2, x_7, x_8\}\}$

U / C_2 = Covers obtained based on project numbers they work on $= \{\{x_1, x_6\}, \{x_2, x_3, x_5\}, \{x_4, x_6\}, \{x_7, x_8\}\}$,

Interpretation of approximate equalities

Consider subsets $A, B \subseteq U$. Then the lower approximation of any set can be interpreted as a group of people who are certainly part of the committee and the upper approximation of any set can be interpreted as a group of people who are either certainly or possibly be part of the committee.

Two sets A and B are said to be optimistic bottom equivalent to each other with respect to C_1 and C_2 if their lower approximations with respect to C_1+C_2 are the same. That is the set of faculties who are certainly in A with respect to C_1 or with respect to C_2 is same as the set of faculties who are certainly in B with respect to C_1 or with respect to C_2.

Two sets A and B are said to be optimistic top equivalent to each other with respect to C_1 and C_2 if their upper approximations with respect to C_1+C_2 are the same. That is the set of faculties who are certainly or possibly be in A with respect to C_1 and with respect to C_2 is same as the set of faculties who are certainly or possibly be in B with respect to C_1 and with respect to C_2.

Table: 2. Table of minimal descriptors generated for C_1 and C_2

Elements Minimal Descriptors	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$md_{C_1}(x)$	{x1, x5}	{x2,x7, x8}	{x3,x4,x5,x6}	{x3,x4,x5,x6, }	{x5}	{x3,x4,x5,x 6,}	{x2,x7,x8}	{x2, x7,x8}
$md_{C_2}(x)$	{x1, x6}	{x2,x3, x5}	{x2,x3,x5}	{x4,x6}	{x2,x3, x5}	{x6}	{x7, x8}	{x7, x8}

Let $A = \{x_3, x_4, x_5, x_6\}$, $A' = \{x_1, x_3, x_4, x_5, x_6\}$, $B = \{x_5, x_6, x_7, x_8\}$, and $B' = \{x_1, x_5, x_6, x_7, x_8\}$

$\underline{O_{C_1+C_2}}(A) = \{x_3, x_4, x_5, x_6\}$ *and* $\underline{O_{C_1+C_2}}(A') = \{x_3, x_4, x_5, x_6\}$

$\underline{O_{C_1+C_2}}(B) = \{x_5, x_6, x_7, x_8\}$ *and* $\underline{O_{C_1+C_2}}(B') = \{x_5, x_6, x_7, x_8\}$

$A \cap B = \{x_5, x_6\}$ *and* $A' \cap B' = \{x_1, x_5, x_6\}$

$\underline{O_{C_1+C_2}}(A \cap B) = \{x_5, x_6\}$ *and* $\underline{O_{C_1+C_2}}(A' \cap B') = \{x_1, x_5, x_6\}$

$\underline{O_{C_1+C_2}}(A \cap B) \neq \underline{O_{C_1+C_2}}(A' \cap B')$. *Thus* $A \cap B \overset{\neq}{}_{C_1+C_2} A' \cap B'$

(3.3.5) $A \overset{C_1+C_2}{=} B \Rightarrow A \cup B^c \overset{C_1+C_2}{=} U$

Proof: Given $A \overset{C_1+C_2}{=} B \Rightarrow \overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B)$. But by (2.5.6)

$\overline{O_{C_1+C_2}}(A \cup B) \supseteq \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(B)$. Thus we have

$\overline{O_{C_1+C_2}}(A \cup B^c) \supseteq \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(B^c)$

THE IIOAB JOURNAL ISSN: 0976-3104

$= \overline{O_{C_1+C_2}}(B) \cup ((\underline{O_{C_1+C_2}}(B))^C) = \overline{O_{C_1+C_2}}(B) \cup ((\overline{O_{C_1+C_2}}B \setminus BN_{C_1+C_2}B)^C)$

$= \overline{O_{C_1+C_2}}(B) \cup \left(\overline{O_{C_1+C_2}}(B)\right)^C = U \Rightarrow A \cup B^c \overset{C_1+C_2}{=} U$

This completes the proof.

(3.3.6) A $\overset{=}{C_1+C_2}$ B => $A \cap B^c \overset{C_1+C_2}{=} \phi$

Proof: Given A $\overset{=}{C_1+C_2}$ B => $\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B)$. But by (2.5.7)

$\underline{O_{C_1+C_2}}(A \cap B) \subseteq \underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(B)$. Thus we have

$\underline{O_{C_1+C_2}}(A \cap B^c) \subseteq \underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(B^c) = \underline{O_{C_1+C_2}}(A) \cap \left(\overline{O_{C_1+C_2}}(B)\right)^C$

$= \underline{O_{C_1+C_2}}(A) \cap (U - \underline{O_{C_1+C_2}}(B) - BN_{C_1+C_2}(B)) \subseteq \underline{O_{C_1+C_2}}(A) \cap (U - \underline{O_{C_1+C_2}}(B)) = \phi$.

=> $A \cap B^c \overset{C_1+C_2}{=} \phi$

(3.3.7)If $A \subseteq B$ and B $\overset{C_1+C_2}{=} \phi$ then A $\overset{C_1+C_2}{=} \phi$

Proof: Given $A \subseteq B$ and B $\overset{C_1+C_2}{=} \phi$. So we have $\overline{O_{C_1+C_2}}(B) = \phi$. As $A \subseteq B$ =>

=> $\overline{O_{C_1+C_2}}(A) \subseteq \overline{O_{C_1+C_2}}(B) = \phi$ => A $\overset{C_1+C_2}{=} \phi$.

(3.3.8) If $A \subseteq B$ and A $\overset{C_1+C_2}{=}$ U then B $\overset{C_1+C_2}{=}$ U

Proof: Given $A \subseteq B$ and A $\overset{C_1+C_2}{=}$ U. So we have $\overline{O_{C_1+C_2}}(A) = U$. As $B \supseteq A$ =>

=> $\overline{O_{C_1+C_2}}(B) \supseteq \overline{O_{C_1+C_2}}(A) = U$ => B $\overset{C_1+C_2}{=}$ U.

(3.3.9) A $\overset{C_1+C_2}{=}$ B iff $A^C \overset{=}{C_1+C_2} B^C$

Proof: Given A $\overset{C_1+C_2}{=}$ B $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B)$

But we know that

$\overline{O_{C_1+C_2}}(A) = (\underline{O_{C_1+C_2}}(A^c))^c \Leftrightarrow (\underline{O_{C_1+C_2}}(A^c))^c = (\underline{O_{C_1+C_2}}(B^c))^c \Leftrightarrow \underline{O_{C_1+C_2}}(A^c) = \underline{O_{C_1+C_2}}(B^c)$

<=> $A^C \overset{=}{C_1+C_2} B^C$

(3.3.10) If A $\overset{=}{C_1+C_2}$ ϕ or B $\overset{=}{C_1+C_2}$ ϕ then $A \cap B \overset{=}{C_1+C_2} \phi$

Proof: Given A $\overset{=}{C_1+C_2}$ ϕ or B $\overset{=}{C_1+C_2}$ ϕ => $\underline{O_{C_1+C_2}}(A) = \phi$ *or* $\underline{O_{C_1+C_2}}(B) = \phi$

=> $\underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(B) = \phi$. But by (2.5.7)

$\underline{O_{C_1+C_2}}(A \cap B) \subseteq \underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(B)$ => $\underline{O_{C_1+C_2}}(A \cap B) = \phi$ => $A \cap B \overset{=}{C_1+C_2} \phi$.

(3.3.11) If A $\overset{C_1+C_2}{=}$ U or B $\overset{C_1+C_2}{=}$ U then $A \cup B \overset{C_1+C_2}{=}$ U

Proof: Given A $\overset{C_1+C_2}{=}$ U or B $\overset{C_1+C_2}{=}$ U => $\overline{O_{C_1+C_2}}(A) = U$ *or* $\overline{O_{C_1+C_2}}(B) = U$

THE IIOAB JOURNAL ISSN: 0976-3104

$\Rightarrow \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(B) = U$. But by (2.5.6)

$\overline{O_{C_1+C_2}}(A \cup B) \supseteq \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(B)$ $\overline{O_{C_1+C_2}}(A \cup B) = U$

$\Rightarrow A \cup B \overset{C_1+C_2}{=} U.$

Replacement properties for first type of covering based optimistic multi granular approximate equalities

These properties are also called as interchange properties. We have stated above the observation of Novotny and Pawlak in connection with holding of the properties for rough equalities when the bottom and top equalities are interchanged. They categorically told that the properties do not hold under this change. However, it is shown by Tripathy et al [16] that some of these properties hold under the interchange where as some other hold with some additional conditions which are sufficient but not necessary. They are stated as below along with their proofs. We use a real life example as detailed below, which shall be used to illustrate the properties as well as provide counter examples whenever necessary.

Example 2: Let us consider **Table–1**. Assume that an exam committee for the school of computing science and engineering (SCSE) is to be constituted to carry out activities such as collecting the question paper bundles and distributing the answer sheet bundles. Assume that there are 8 faculties available for the purpose. Their collection and distribution experiences in years along with their sex and division they belong to are considered for forming two sets of covers as given below.

C_1=Cover obtained by defining similarity relation between two faculties such that they are related to each other iff they belong to different divisions with their average collection experience as exactly 2 years and at most one them be a female faculty

$U/C_1 = \{\{x_1, x_8\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_4, x_5\}, \{x_6, x_7\}\}$

C_2= Cover obtained by defining similarity relation between two faculties such that they are related to each other iff they belong to different divisions with their average distribution experience as exactly 2 years and at most one of them be a female faculty

$U/C_2 = \{\{x_1, x_8\}, \{x_2, x_7\}, \{x_3, x_4\}, \{x_4, x_5\}, \{x_6, x_7\}\}$

Interpretation of approximate equalities

Consider subsets $A, B \subseteq U$. Then the lower approximation of any set can be interpreted as a group of people who are certainly part of the committee and the upper approximation of any set can be interpreted as a group of people who are either certainly or possibly be part of the committee.

Two sets A and B are said to be optimistic bottom equivalent to each other with respect to C_1 and C_2 if their lower approximations with respect to C_1+C_2 are the same. That is the set of faculties who are certainly in A with respect to C_1 or with respect to C_2 is same as the set of faculties who are certainly in B with respect to C_1 or with respect to C_2.

Two sets A and B are said to be optimistic top equivalent to each other with respect to C_1 and C_2 if their upper approximations with respect to C_1+C_2 are the same. That is the set of faculties who are certainly or possibly be in A with respect to C_1 and with respect to C_2 is same as the set of faculties who are certainly or possibly be in B with respect to C_1 and with respect to C_2.

Let us consider first type of CBOMGRS. Its lower and upper approximations are determined based on minimal descriptors.

THE IIOAB JOURNAL ISSN: 0976-3104

The minimal descriptor table for the two covers for the above example is as shown below.

Table: 3. Table of minimal descriptors for C_1 and C_2

Elements Minimum Descriptors	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$md_{C_1}(x)$	{x_1, x_8}	{x_2,x_3}	{x_3}	{x_4}	{x_4,x_5}	{x_6,x_7}	{x_6,x_7}	{x_1,x_8}
$md_{C_2}(x)$	{x_1, x_8}	{x_2,x_7}	{x_3,x_4}	{x_4}	{x_4,x_5}	{x_6,x_7}	{x_6,x_7}	{x_1,x_8}

(3.4.1) A $\overset{C_1+C_2}{=}$ B if $A \cap B \overset{C_1+C_2}{=}$ A and B both. The converse need not be true.

Proof: *Given* $A \cap B \overset{C_1+C_2}{=} A \Rightarrow \overline{O_{C_1+C_2}}(A \cap B) = \overline{O_{C_1+C_2}}(A)$

Given $A \cap B \overset{C_1+C_2}{=} B \Rightarrow \overline{O_{C_1+C_2}}(A \cap B) = \overline{O_{C_1+C_2}}(B)$

From the above two expressions we have, $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B) \Rightarrow$ A $\overset{C_1+C_2}{=}$ B.

This part can be interpreted as, if the set of faculty certainly or possibly be in committee for $A \cap B$ with respect to C_1+C_2 is the same as that of A and B, then the set of faculty certainly or possibly be in committee for A with respect to C_1+C_2 will be same as that of B. It means that a committee obtained through common people from sets A and B having same group of people who are either certainly or possibly be in the committee is the same as the committees obtained from A and B having same group of people who are either certainly or possibly be in the committee, then those committees obtained from A and B that way will be equal.

The following example shows that converse need not be true.

Let $A = \{x_3, x_6\}, B = \{x_3, x_7\}$ *and*

$\overline{O_{C_1+C_2}}(A) = \{x_3, x_6, x_7\}$, $\overline{O_{C_1+C_2}}(B) = \{x_3, x_6, x_7\}$ *and* $\overline{O_{C_1+C_2}}(A \cap B) = \{x_3\}$

Thus $\overline{O_{C_1+C_2}}(A \cap B) \neq \overline{O_{C_1+C_2}}(A)$ *and* $\overline{O_{C_1+C_2}}(A \cap B) \neq \overline{O_{C_1+C_2}}(B)$

though $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B)$

The converse part can be interpreted as, though the sets of faculty certainly or possibly be in committee with respect to C_1+C_2 are the same for A and B but the set of faculty certainly or possibly be in committee for $A \cap B$ with respect to C_1+C_2 is not same as that of A and B. It means that a committee obtained through common people from sets A and B having same group of people who are either certainly or possibly be in the committee , may not be the same as the committee obtained from A and B having same group of people who are either certainly or possibly be in the committee. Then the committees obtained from A and B that way need not be equal.

(3.4.2)A $\overset{=}{C_1+C_2}$ B if $A \cup B \overset{=}{C_1+C_2}$ A and B both. The converse need not be true

Proof: *Given* $A \cup B \overset{=}{C_1+C_2} A \Rightarrow \underline{O_{C_1+C_2}}(A \cup B) = \underline{O_{C_1+C_2}}(A)$ and

Given $A \cup B \overset{=}{C_1+C_2} B \Rightarrow \underline{O_{C_1+C_2}}(A \cup B) = \underline{O_{C_1+C_2}}(B)$

From the above two expressions we have

$\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B)$ *and so* $A \overset{=}{C_1+C_2} B$.

This part can be interpreted as, if the set of faculty certainly be in committee for $A \cup B$ with respect to C_1+C_2 is the same as that of A and B, then the set of faculty certainly be in committee for A with respect to C_1+C_2 will be same as that of B. It means that a committee obtained through the people from sets A and B having same group of people who are certainly be in the committee is the same as the committees obtained from A and B having same

THE IIOAB JOURNAL ISSN: 0976-3104

group of people who are certainly be in the committee, then the committees obtained from A and B that way will be equal.

The following example shows that the converse need not be true.

Let $A = \{x_4, x_6\}, B = \{x_4, x_7\}$ *and* $A \cup B = \{x_4, x_6, x_7\}$

$\underline{O_{C_1+C_2}}(A) = \{x_4\}$ *and* $\underline{O_{C_1+C_2}}(B) = \{x_4\}$ $\underline{O_{C_1+C_2}}(A \cup B) = \{x_4, x_6, x_7\}$

Thus $\underline{O_{C_1+C_2}}(A \cup B) \neq \underline{O_{C_1+C_2}}(A)$ *and* $\underline{O_{C_1+C_2}}(A \cup B) \neq \underline{O_{C_1+C_2}}(B)$

though $\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B)$

The converse part can be interpreted as, though the sets of faculty certainly in committee with respect to C_1+C_2 are the same for A and B, but the set of faculty certainly in committee for $A \cup B$ with respect to C_1+C_2 is not the same as that of A and B It means that a committee obtained through common people from sets A and B having same group of people who are certainly in the committee , may not be the same as the committee obtained from A and B having same group of people who are certainly in the committee. Then the committees obtained from A and B that way need not be equal.

(3.4.3)A $\overset{=}{}_{C_1+C_2}$ A' and B $\overset{=}{}_{C_1+C_2}$ B' may not imply that $A \cup B$ $\overset{=}{}_{C_1+C_2}$ $A' \cup B'$

Proof: The following example establishes the above proof.

Let A = $\{x_3, x_6\}$, A' = $\{x_1, x_3\}$, B = $\{x_6, x_8\}$ and B' = $\{x_7, x_8\}$. Then

$\underline{O_{C_1+C_2}}(A) = \{x_3\}$ *and* $\underline{O_{C_1+C_2}}(A') = \{x_3\}$

$\underline{O_{C_1+C_2}}(B) = \phi$ *and* $\underline{O_{C_1+C_2}}(B') = \phi$

$\Rightarrow \underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(A')$ *and* $\underline{O_{C_1+C_2}}(B) = \underline{O_{C_1+C_2}}(B')$

Thus A $\overset{=}{}_{C_1+C_2}$ A' and B $\overset{=}{}_{C_1+C_2}$ B'.

Now, $A \cup B = \{x_3, x_6, x_8\}$, so $\underline{O_{C_1+C_2}}(A \cup B) = \{x_3\}$

Also, $A' \cup B' = \{x_1, x_3, x_7, x_8\}$, so $\underline{O_{C_1+C_2}}(A' \cup B') = \{x_1, x_8\}$

Thus $\underline{O_{C_1+C_2}}(A \cup B) \neq \underline{O_{C_1+C_2}}(A' \cup B')$. $\Rightarrow A \cup B$ $\overset{\neq}{}_{C_1+C_2}$ $A' \cup B'$.

(3.4.4)A $\overset{C_1+C_2}{=}$ A' and B $\overset{C_1+C_2}{=}$ B' may not imply that $A \cap B$ $\overset{C_1+C_2}{=}$ $A' \cap B'$

Proof: The following example establishes the proof.

Let A = $\{x_3, x_4\}$, A' = $\{x_3, x_5\}$, B = $\{x_4, x_6\}$ and B' = $\{x_4, x_7\}$. Then

$\overline{O_{C_1+C_2}}(A) = \{x_3, x_4, x_5\}$ *and* $\overline{O_{C_1+C_2}}(A') = \{x_3, x_4, x_5\}$;

$\overline{O_{C_1+C_2}}(B) = \{x_4, x_6, x_7\}$ *and* $\overline{O_{C_1+C_2}}(B') = \{x_4, x_6, x_7\}$.

So, $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(A')$ *and* $\overline{O_{C_1+C_2}}(B) = \overline{O_{C_1+C_2}}(B')$

But $A \cap B = \{x_4\} \Rightarrow \overline{O_{C_1+C_2}}(A \cap B) = \{x_4\}$

and $A' \cap B' = \phi \Rightarrow \overline{O_{C_1+C_2}}(A' \cap B') = \phi$.

Thus, $\overline{O_{C_1+C_2}}(A \cap B) \neq \overline{O_{C_1+C_2}}(A' \cap B') \Rightarrow A \cap B$ $\overset{C_1+C_2}{\neq}$ $A' \cap B'$.

(3.4.5) A $\overset{=}{}_{C_1+C_2}$ B => $A \cup B^c$ $\overset{=}{}_{C_1+C_2}$ U

THE IIOAB JOURNAL ISSN: 0976-3104

Proof: Given A $\overset{=}{C_1+C_2}$ B => $\underline{O_{C_1+C_2}}(A)=\underline{O_{C_1+C_2}}(B)$. But from (2.5.5) $\underline{O_{C_1+C_2}}(A\cup B)\supseteq\underline{O_{C_1+C_2}}(A)\cup\underline{O_{C_1+C_2}}(B)$. Thus, we have

$\underline{O_{C_1+C_2}}(A\cup B)\supseteq\underline{O_{C_1+C_2}}(A)\cup\underline{O_{C_1+C_2}}(B) = \underline{O_{C_1+C_2}}(B)\cup(\overline{O_{C_1+C_2}}(B))^c$

$= \underline{O_{C_1+C_2}}(B)\cup(U-(\underline{O_{C_1+C_2}}(B)-BN_{C_1+C_2}B))$

$\subseteq \underline{O_{C_1+C_2}}(B)\cup(U-\underline{O_{C_1+C_2}}(B))=U \Rightarrow A\cup B' \overset{=}{C_1+C_2}$ U

(3.4.6) A $\overset{C_1+C_2}{=}$ B => $A\cap B^c \overset{=}{C_1+C_2}\ \phi$

Proof: Given A $\overset{C_1+C_2}{=}$ B => $\overline{O_{C_1+C_2}}(A)=\overline{O_{C_1+C_2}}(B)$. But from (2.5.8)

$\overline{O_{C_1+C_2}}(A\cap B)\subseteq\overline{O_{C_1+C_2}}(A)\cap\overline{O_{C_1+C_2}}(B)$. Thus, we have

$\overline{O_{C_1+C_2}}(A\cap B^c)\subseteq\overline{O_{C_1+C_2}}(A)\cap\overline{O_{C_1+C_2}}(B^c)=\overline{O_{C_1+C_2}}(A)\cap(\underline{O_{C_1+C_2}}(B))^c=\overline{O_{C_1+C_2}}(A)\cap(U-\underline{O_{C_1+C_2}}(B))$

$\subseteq BN_{C_1+C_2}(B)$

$\Rightarrow A\cap B' \overset{C_1+C_2}{\neq}\ \phi$.

(3.4.7) If $A\subseteq B$ and B $\overset{=}{C_1+C_2}\ \phi$ then A $\overset{=}{C_1+C_2}\ \phi$

Proof: Given $A\subseteq B$ and B $\overset{=}{C_1+C_2}\ \phi$. So, we have $\underline{O_{C_1+C_2}}(B)=\phi$. As $A\subseteq B \Rightarrow A=\phi \Rightarrow \underline{O_{C_1+C_2}}(A)=\phi$

=> A $\overset{=}{C_1+C_2}\ \phi$

(3.4.8) If $A\subseteq B$ and A $\overset{=}{C_1+C_2}$ U then B $\overset{=}{C_1+C_2}$ U

Proof: Given $A\subseteq B$ and B $\overset{=}{C_1+C_2}$ U. So $\underline{O_{C_1+C_2}}(A)=U$. As $A\subseteq B \Rightarrow \underline{O_{C_1+C_2}}(A)\subseteq\underline{O_{C_1+C_2}}(B)$, we have $\underline{O_{C_1+C_2}}(B)=U$

=> B $\overset{=}{C_1+C_2}$ U.

(3.4.9) A $\overset{=}{C_1+C_2}$ B iff $A^c \overset{C_1+C_2}{=} B^c$

Proof: Given A $\overset{=}{C_1+C_2}$ B

But we know that $\underline{O_{C_1+C_2}}(A)=(\overline{O_{C_1+C_2}}(A^c))^c$

$\Rightarrow(\overline{O_{C_1+C_2}}(A^c))^c=(\overline{O_{C_1+C_2}}(B^c))^c \Rightarrow \overline{O_{C_1+C_2}}(A^c)=\overline{O_{C_1+C_2}}(A^c)$

$\underline{O_{C_1+C_2}}(A)=\underline{O_{C_1+C_2}}(B) \Rightarrow A^c \overset{C_1+C_2}{=} B^c$

In a similar way converse will also be proved.

(3.4.10) If A $\overset{C_1+C_2}{=}\ \phi$ or B $\overset{C_1+C_2}{=}\ \phi$ then $A\cap B \overset{C_1+C_2}{=}\ \phi$

Proof: Given A $\overset{C_1+C_2}{=}\ \phi$ or B $\overset{C_1+C_2}{=}\ \phi$. So, $\overline{O_{C_1+C_2}}(A)=\phi$ *or* $\overline{O_{C_1+C_2}}(B)=\phi \Rightarrow \overline{O_{C_1+C_2}}(A)\cap\overline{O_{C_1+C_2}}(B)=\phi$.

But from (2.5.8) $\overline{O_{C_1+C_2}}(A\cap B)\subseteq\overline{O_{C_1+C_2}}(A)\cap\overline{O_{C_1+C_2}}(B) \Rightarrow \overline{O_{C_1+C_2}}(A\cap B)=\phi \Rightarrow A\cap B \overset{C_1+C_2}{=}\ \phi$

(3.4.11) If A $\overset{=}{C_1+C_2}$ U or B $\overset{=}{C_1+C_2}$ U then $A\cup B \overset{=}{C_1+C_2}$ U

THE IIOAB JOURNAL ISSN: 0976-3104

Proof: Given A $\underset{=}{}^{C_1+C_2}$ U or B $\underset{=}{}^{C_1+C_2}$ U. So, $\underline{O_{C_1+C_2}}(A) = U \text{ or } \underline{O_{C_1+C_2}}(B) = U \Rightarrow \underline{O_{C_1+C_2}}(A) \cup \underline{O_{C_1+C_2}}(B) = U$.

But from (2.5.7) $\underline{O_{C_1+C_2}}(A \cup B) \supseteq \underline{O_{C_1+C_2}}(A) \cup \underline{O_{C_1+C_2}}(B) \Rightarrow \underline{O_{C_1+C_2}}(A \cup B) = U \Rightarrow A \cup B \underset{C_1+C_2}{=} U$.

In a similar manner we can prove the necessity condition.

Approximate rough equivalence for covering based optimistic multi granulation

We now introduce the different rough equivalences for the first type of covering based optimistic multi granular rough set (CBOMGRS). These definitions will be the same for other types. Then we study, prove and provide counter examples for its direct and replacement properties as per requirement.

Let C_1 and C_2 be two covers on U and $C_1, C_2 \in C \text{ and } A, B \subseteq U$. Let F denotes first type CBOMGRS.

Definition 3.5: We say that,

(3.5.1) A and B are bottom rough $C_1 + C_2$ equivalent to each other (A $\overset{\approx}{}_{C_1+C_2}$ B) iff

$\underline{O_{C_1+C_2}}(A) \text{ and } \underline{O_{C_1+C_2}}(B) \text{ are } \phi \text{ or not } \phi \text{ together}.$

(3.5.2) A and B are top rough $C_1 + C_2$ equivalent to each other (A $\underset{\approx}{}^{C_1+C_2}$ B) iff

$\overline{O_{C_1+C_2}}(A) \text{ and } \overline{O_{C_1+C_2}}(B) \text{ are } U \text{ or not } U \text{ together}.$

(3.5.3) A and B are total rough $C_1 + C_2$ equivalent to each other (A $r\_C_1+C_2\_eqv$ B) iff

$\underline{O_{C_1+C_2}}(A) \text{ and } \underline{O_{C_1+C_2}}(B) \text{ are } \phi \text{ or not } \phi \text{ together}$ and $\overline{O_{C_1+C_2}}(A) \text{ and } \overline{O_{C_1+C_2}}(B)$ are U or not U together.

Following are the generalization of the approximate rough inclusions introduced by Pawlak [5,6] and approximate rough comparisons introduced by Tripathy et al[16]. We define these concepts in the context of first type of covering based
optimistic multi granulation as below.

Definition 3.6:

Let K=(U,R) be a knowledge base and $A, B \subseteq U \text{ and } C_1, C_2 \in C$. Then

(i) We say A is bottom $C_1 + C_2$ rough included in B ($A \overset{\subseteq}{}_{C_1+C_2} B$) *iff* $\underline{O_{C_1+C_2}}(A) \subseteq \underline{O_{C_1+C_2}}(B)$.

(ii) We say A is bottom $C_1 + C_2$ rough included in B ($A \overset{C_1+C_2}{\subseteq} B$) *iff* $\overline{O_{C_1+C_2}}(A) \subseteq \overline{O_{C_1+C_2}}(B)$.

(iii) We say A is rough $C_1 + C_2$ included in B *iff* $\underline{O_{C_1+C_2}}(A) \subseteq \underline{O_{C_1+C_2}}(B)$ *and* $\overline{O_{C_1+C_2}}(B) \subseteq \overline{O_{C_1+C_2}}(A)$.

Definition3.7:

Let K=(U,R) be a knowledge base and $A, B \subseteq U \text{ and } C_1, C_2 \in C$. Then

(i) We say that $A, B \subseteq U$ are bottom $C_1 + C_2$ comparable iff $A \overset{\subseteq}{}_{C_1+C_2} B$ *or* $B \overset{\subseteq}{}_{C_1+C_2} A$.

(ii) We say that $A, B \subseteq U$ are top $C_1 + C_2$ comparable iff $A \overset{C_1+C_2}{\subseteq} B$ *or* $B \overset{C_1+C_2}{\subseteq} A$.

(iii) We say that $A, B \subseteq U$ are $C_1 + C_2$ comparable iff A and B are both bottom $C_1 + C_2$ and top $C_1 + C_2$ comparable.

THE IIOAB JOURNAL ISSN: 0976-3104

Properties for covering based optimistic multi granular approximate equivalence

(3.6.1)(i) If $A \cap B \overset{\approx}{}_{C_1+C_2}$ A and $A \cap B \overset{\approx}{}_{C_1+C_2}$ B then A $\overset{\approx}{}_{C_1+C_2} B$.

(ii) The converse of (i) is not necessarily true.

(iii) The converse is true in (iii) if A and B is bottom $C_1 + C_2$ comparable.

(iv) the condition in (iii) is not necessary.

Proof:

$(i)\ \underline{O}_{C_1+C_2}(A \cap B)$ *and* $\underline{O}_{C_1+C_2}(A)$ *are either* ϕ *or not* ϕ *together* (*given*).

$\underline{O}_{C_1+C_2}(A \cap B)$ *and* $\underline{O}_{C_1+C_2}(B)$ *are either* ϕ *or not* ϕ *together* (*given*).

Then $\underline{O}_{C_1+C_2}(A)$ *and* $\underline{O}_{C_1+C_2}(B)$ *are either* ϕ *or not* ϕ *together* (*derived*).

Thus $A \overset{\approx}{}_{C_1+C_2} B$.

(ii) Continuing with example2 by taking A={x_3} and B={x_6, x_7}, we have

$\underline{O}_{c_1+c_2}(A) = \{x_3\} \neq \phi$ *and* $\underline{O}_{c_1+c_2}(B) = \{x_6, x_7\} \neq \phi \Rightarrow A \overset{\approx}{}_{C_1+C_2} B$. But $A \cap B = \phi$. Then

$\underline{O}_{C_1+C_2}(A \cap B) = \phi \Rightarrow A \cap B$ *not* $\overset{\approx}{}_{C_1+C_2} A$ *and* B *both*.

(iii) Even if A and B is bottom $C_1 + C_2$ comparable, we have $\underline{O}_{C_1+C_2}(A \cap B) \subseteq \underline{O}_{C_1+C_2}(A)$ *or* $\underline{O}_{C_1+C_2}(B)$ as the case

may be. So, if both $\underline{O}_{C_1+C_2}(A)$ and $\underline{O}_{C_1+C_2}(B)$ are ϕ, we have $\underline{O}_{C_1+C_2}(A \cap B) = \phi$. But when both are not ϕ, we cannot say the same for

$\underline{O}_{C_1+C_2}(A \cap B)$.

(iv) Continuing with example2 by taking A={x_4, x_5, x_6, x_7} and B={x_1, x_4, x_8}, we have

$\underline{O}_{c_1+c_2}(A) = \{x_3, x_5, x_6, x_7\} \neq \phi, \underline{O}_{c_1+c_2}(B) = \{x_1, x_4, x_8\} \neq \phi \Rightarrow A \overset{\approx}{}_{C_1+C_2} B$.

Also A *and* B *are not bottom* $C_1 + C_2$ *comparable*.

But $A \cap B = \{x_4\}$. *Then* $\underline{O}_{C_1+C_2}(A \cap B) \neq \phi \Rightarrow A \cap B \overset{\approx}{}_{C_1+C_2} A$ *and* $A \cap B \overset{\approx}{}_{C_1+C_2} B$

though A *and* B *are not bottom* $C_1 + C_2$ *comparable*.

(3.6.2)(i) If $A \cup B \overset{C_1+C_2}{\approx}$ A and $A \cup B \overset{C_1+C_2}{\approx}$ B then A $\overset{C_1+C_2}{\approx}$ B.

(ii)The converse of (i) is not necessarily true.

(iii) The converse cannot be true even if A and B are top $C_1 + C_2$ comparable.

(iv) The conditions in (iii) is not necessary.

Proof:

$(i)\ \overline{O}_{C_1+C_2}(A \cup B)$ *and* $\underline{O}_{C_1+C_2}(A)$ *are either* U *or not* U *together* (*given*).

$\overline{O}_{C_1+C_2}(A \cup B)$ *and* $\underline{O}_{C_1+C_2}(B)$ *are either* U *or not* U *together* (*given*).

Then $\underline{O}_{C_1+C_2}(A)$ *and* $\underline{O}_{C_1+C_2}(B)$ *are either* U *or not* U *together* (*derived*).

Thus $A \overset{C_1+C_2}{\approx} B$.

(ii) Continuing with example 2 by taking A={x_1, x_2, x_3, x_4} and B={x_5, x_6, x_7}, we have

$\overline{O}_{C_1+C_2}(A) = \{x_1, x_2, x_3, x_4, x_8\} \neq U$ *and* $\overline{O}_{C_1+C_2}(B) = \{x_5, x_6, x_7\} \neq U \Rightarrow A \overset{C_1+C_2}{\approx} B$. But

$A \cup B = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$. *Then* $\overline{O}_{C_1+C_2}(A \cup B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U$

$\Rightarrow A \cup B$ *not* $\overset{C_1+C_2}{\approx}$ A *and* B *both*.

(iii) Even if A and B is bottom $C_1 + C_2$ comparable, we have $\underline{O}_{C_1+C_2}(A \cup B) \subseteq \underline{O}_{C_1+C_2}(A)$ *or* $\underline{O}_{C_1+C_2}(B)$ as the

THE IIOAB JOURNAL ISSN: 0976-3104

case may be. So, if both $\underline{O_{C_1+C_2}}(A)$ and $\underline{O_{C_1+C_2}}(B)$ are U, we have $\underline{O_{C_1+C_2}}(A \cup B) = U$. But when both are not U,
we cannot say the same for $\underline{O_{C_1+C_2}}(A \cup B)$.

(iv) Continuing with example 2 by taking A={x_1, x_2, x_3, x_4} and B={x_5, x_6}we have

$\overline{O_{C_1+C_2}}(A) = \{x_1, x_2, x_3, x_4, x_8\} \neq U$ *and* $\overline{O_{C_1+C_2}}(B) = \{x_5, x_6\} \neq U$

$\Rightarrow \overline{O_{C_1+C_2}}(A) \not\subset \overline{O_{C_1+C_2}}(B)$ *or* $\overline{O_{C_1+C_2}}(B) \not\subset \overline{O_{C_1+C_2}}(A)$

$\Rightarrow A$ *and* B *are not top* $C_1 + C_2$ *comparable.*

And $A \cup B = \{x_1, x_2, x_3, x_4\}$. *Then* $\overline{O_{C_1+C_2}}(A \cup B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_8\} \neq U$

$\Rightarrow A \cup B \overset{C_1+C_2}{\approx}$ A *and* B *both though* A *and* B *are not top* $C_1 + C_2$ *comparable.*

(3.6.3)(i) If A $\overset{C_1+C_2}{\approx}$ A' and B $\overset{C_1+C_2}{\approx}$ B' then it may or may not be true that $A \cup B \overset{C_1+C_2}{\approx} A' \cup B'$.

(ii) A sufficient condition for the result in (i) to be true is that A and B are top $C_1 + C_2$ comparable and A' and B' are top $C_1 + C_2$ comparable.

(iii) The conditions in (ii) are not necessary for result in (i) to be true.

Proof: (i) The result fails to be true when all $\overline{O_{C_1+C_2}}(A), \overline{O_{C_1+C_2}}(A'), \overline{O_{C_1+C_2}}(B)$, *and* $\overline{O_{C_1+C_2}}(B')$ are not U and exactly one of $A \cup B$ *and* $A' \cup B'$ *is* U, then result will fail. The following example shows that.
Continuing with example 2 by taking A={x_1, x_2, x_3, x_4}, A'= {x_5, x_6, x_7}, B={x_1, x_5, x_6, x_7}, and B'={x_3, x_4, x_5}, we have

$\overline{O_{C_1+C_2}}(A) = \{x_1, x_2, x_3, x_4, x_8\} \neq U$ *and* $\overline{O_{C_1+C_2}}(A') = \{x_5, x_6, x_7\} \neq U$. This implies that A $\overset{C_1+C_2}{\approx}$ A'.

$\overline{O_{C_1+C_2}}(B) = \{x_1, x_5, x_6, x_7\} \neq U$ *and* $\overline{O_{C_1+C_2}}(B') = \{x_3, x_4, x_5\} \neq U$. This implies that B $\overset{C_1+C_2}{\approx}$ B'.

But $A \cup B = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ *and* $A' \cup B' = \{x_1, x_3, x_4, x_5, x_6, x_7\}$

$\overline{O_{C_1+C_2}}(A \cup B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U$ *and* $\overline{O_{C_1+C_2}}(A' \cup B') = \{x_1, x_3, x_4, x_5, x_6, x_7, x_8\} \neq U$.

$\Rightarrow A \cup B$ *not* $\overset{C_1+C_2}{\approx}$ $A' \cup B'$.

(ii) We have $\overline{O_{C_1+C_2}}(A) \neq U, \overline{O_{C_1+C_2}}(A') \neq U, \overline{O_{C_1+C_2}}(B) \neq U$, *and* $\overline{O_{C_1+C_2}}(B') \neq U$. So, under the hypothesis,

$\overline{O_{C_1+C_2}}(A \cup B) \supseteq \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(B) = \overline{O_{C_1+C_2}}(A)$ *or* $\overline{O_{C_1+C_2}}(B)$, which is not equal to U. Similarly, $\overline{O_{C_1+C_2}}(A' \cup B') \neq U$. Hence $A \cup B \overset{C_1+C_2}{\approx} A' \cup B'$.

(iii) Continuing with example 2 by taking A={x_1, x_2, x_3, x_4}, A'= {x_5, x_6, x_7},B={x_1, x_5, x_6, x_7}and B'={x_2, x_3, x_4, x_5}, we have

$\overline{O_{C_1+C_2}}(A) = \{x_1, x_2, x_3, x_4, x_8\} \neq U$ *and* $\overline{O_{C_1+C_2}}(A') = \{x_5, x_6, x_7\} \neq U$. This implies that A and A' are not top rough comparable.

$\overline{O_{C_1+C_2}}(B) = \{x_1, x_5, x_6, x_7\} \neq U$ *and* $\overline{O_{C_1+C_2}}(B') = \{x_3, x_4, x_5\} \neq U$. This implies that B and B' are not top rough comparable.

But $A \cup B = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ *and* $A' \cup B' = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

$\overline{O_{C_1+C_2}}(A \cup B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U$ *and* $\overline{O_{C_1+C_2}}(A' \cup B') = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U$

$\Rightarrow A \cup B \overset{C_1+C_2}{\approx} A' \cup B'$.

(3.6.4)(i) If A $\overset{\approx}{C_1+C_2}$ A' and B $\overset{\approx}{C_1+C_2}$ B' then it may or may not be true that $A \cap B \overset{\approx}{C_1+C_2} A' \cap B'$.

THE IIOAB JOURNAL ISSN: 0976-3104

(ii) A sufficient condition for the result in (i) to be true is that A and Bare bottom $C_1 + C_2$ comparable and A' and B'
are bottom $C_1 + C_2$ comparable.
(iii) The conditions in (ii) are not necessary for result in (i) to be true.

Proof: (i) The result fails to be true when all $\underline{O_{C_1+C_2}}(A), \underline{O_{C_1+C_2}}(A'), \underline{O_{C_1+C_2}}(B), \text{ and } \underline{O_{C_1+C_2}}(B')$ are not ϕ and exactly
one of $A \cap B \text{ and } A' \cap B' \text{ is } \phi$, then result will fail. The following example shows that.
Continuing with example 2 by taking A={x_3}, A'= {x_6, x_7}, B={x_4}, and B'={x_4, x_6, x_7}, we have
$\underline{O_{C_1+C_2}}(A) = \{x_3\} \neq \phi \text{ and } \underline{O_{C_1+C_2}}(A') = \{x_6, x_7\} \neq \phi$. This implies that A $\overset{\approx}{C_1+C_2}$ A'.
$\underline{O_{C_1+C_2}}(B) = \{x_4\} \neq \phi \text{ and } \underline{O_{C_1+C_2}}(B') = \{x_6, x_7\} \neq \phi$. This implies that B $\overset{\approx}{C_1+C_2}$ B'.
$But\ A \cap B = \phi \text{ and } A' \cap B' = \{x_6, x_7\}. \underline{O_{C_1+C_2}}(A \cap B) = \phi \text{ and } \underline{O_{C_1+C_2}}(A' \cap B') = \{x_6, x_7\} \neq \phi.$
$\Rightarrow A \cap B \text{ not } \overset{\approx}{C_1+C_2} A' \cap B'.$
(ii) We have $\underline{O_{C_1+C_2}}(A) \neq \phi, \underline{O_{C_1+C_2}}(A') \neq \phi, \underline{O_{C_1+C_2}}(B) \neq \phi, \text{ and } \underline{O_{C_1+C_2}}(B') \neq \phi$. So, under the hypothesis,
$\underline{O_{C_1+C_2}}(A \cap B) \subseteq \underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(B) = \underline{O_{C_1+C_2}}(A) \text{ or } \underline{O_{C_1+C_2}}(B) \neq \phi$. Similarly,
$\underline{O_{C_1+C_2}}(A' \cap B') \neq \phi$.
Hence, $A \cap B \overset{\approx}{C_1+C_2} A' \cap B'$.

(iii) Continuing with example 2 by taking A={x_6, x_7}, A'= {x_1, x_8}, B={x_2, x_6, x_7}, and B'={x_1, x_2, x_8}, we have
$\underline{O_{C_1+C_2}}(A) = \{x_6, x_7\} \neq \phi \text{ and } \underline{O_{C_1+C_2}}(A') = \{x_1, x_8\} \neq \phi$. This implies that A and A' are not bottom $C_1 + C_2$
comparable.
$\underline{O_{C_1+C_2}}(B) = \{x_2, x_6, x_7\} \neq \phi \text{ and } \underline{O_{C_1+C_2}}(B') = \{x_1, x_6\} \neq \phi$. This implies that B and B' are not bottom $C_1 + C_2$
comparable.
$But\ A \cap B = \{x_6, x_7\} \text{ and } A' \cap B' = \{x_1, x_8\}\ \underline{O_{C_1+C_2}}(A \cap B) = \{x_6, x_7\} \neq \phi \text{ and } \underline{O_{C_1+C_2}}(A' \cap B') = \{x_1, x_8\} \neq \phi.$
$\Rightarrow A \cap B \overset{\approx}{C_1+C_2} A' \cap B'.$

(3.6.5) (i)A $\overset{\approx}{C_1+C_2}$ B may or may not imply that $(A \cup \sim B) \overset{\approx}{C_1+C_2}$ U.
(ii) A sufficient condition for the result in (i) to hold is that A and B are bottom $C_1 + C_2$ equal.
(iii) The conditions in (ii) are not necessary for the result in (i) to hold.

Proof: (i) The result fails to hold true when $\overline{O_{C_1+C_2}}(A) \neq U, \overline{O_{C_1+C_2}}(B) \neq U$ and still $\overline{O_{C_1+C_2}}(A \cup \sim B) = U$.
(ii) The condition in (ii) is not sufficient as we have
$\overline{O_{C_1+C_2}}(A \cup \sim B) \supseteq \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(\sim B) = \overline{O_{C_1+C_2}}(A) \cup \overline{O_{C_1+C_2}}(\sim A) \supseteq \overline{O_{C_1+C_2}}(A \cup \sim A) = U$
(iii) Continuing with example 2 by taking A={x_1, x_3, x_4, x_7} and B={x_1, x_8}, we have
$\underline{O_{C_1+C_2}}(A) = \{x_2, x_3, x_4\} \text{ and } \underline{O_{C_1+C_2}}(B) = \{x_1, x_8\} \Rightarrow$ A and B is not bottom rough equal.
~B={x_2, x_3, x_4, x_5, x_6, x_7} and $A \cup \sim B = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$
$\overline{O_{C_1+C_2}}(A \cup \sim B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U \Rightarrow A \cup \sim B \overset{\approx}{C_1+C_2} U$.

THE IIOAB JOURNAL ISSN: 0976-3104

(3.6.6) (i) A $\overset{\approx}{}_{C_1+C_2}$ B may or may not imply that $(A \cup \sim B) \overset{\approx}{}_{C_1+C_2}$ U.

(ii) A sufficient condition for the result in (i) to hold is that A and B are top $C_1 + C_2$ equal.

(iii) The conditions in (ii) are not necessary for the result in (i) to hold.

Proof: (i) The result fails to hold true when $\underline{O_{C_1+C_1}}(A) \neq \phi, \underline{O_{C_1+C_1}}(B) \neq \phi$ and still $\underline{O_{C_1+C_1}}(A \cap B) = \phi$.

(ii) The condition in (ii) is not sufficient as we have

$\underline{O_{C_1+C_2}}(A \cap \sim B) \subseteq \underline{O_{C_1+C_2}}(A) \cap \underline{O_{C_1+C_2}}(\sim B) = \underline{O_{C_1+C_2}}(A) \cup \underline{O_{C_1+C_2}}(\sim A) \subseteq \underline{O_{C_1+C_2}}(A \cap \sim A) = \phi$

(iii) Continuing with example 2 by taking A={x_1, x_2, x_5} and B={x_2, x_3, x_4}, we have

$\overline{O_{C_1+C_2}}(A) = \{x_1, x_2, x_5\}$ *and* $\overline{O_{C_1+C_2}}(B) = \{x_2, x_3, x_5\} \Rightarrow$ A and B is not top rough equal.

~B={x_1, x_5, x_6, x_7, x_8} and $A \cap \sim B = \{x_1, x_5\}$

$\underline{O_{C_1+C_2}}(A \cap \sim b) = \phi \Rightarrow A \cap \sim B \overset{\approx}{}_{C_1+C_2} \phi$.

(3.6.7) If $A \subseteq B$ and B $\overset{\approx}{}_{C_1+C_2}$ ϕ then A $\overset{\approx}{}_{C_1+C_2}$ ϕ.

Proof: As B $\overset{\approx}{}_{C_1+C_2}$ ϕ, we have $\underline{O_{C_1+C_2}}(B) = \phi$. So, if $A \subseteq B$, $\underline{O_{C_1+C_2}}(A) \subseteq \underline{O_{C_1+C_2}}(B) = \phi$. Thus A $\overset{\approx}{}_{C_1+C_2}$ ϕ.

(3.6.8) If $A \subseteq B$ and A $\overset{C_1+C_2}{\approx}$ U then B $\overset{C_1+C_2}{\approx}$ U.

Proof: As A $\overset{C_1+C_2}{\approx}$ U, we have $\overline{O_{C_1+C_2}}(A) = U$. So, if $A \subseteq B$, $\overline{O_{C_1+C_2}}(B) \supseteq \overline{O_{C_1+C_2}}(A) = U$. Thus B $\overset{C_1+C_2}{\approx}$ U.

(3.6.9) A $\overset{C_1+C_2}{\approx}$ B iff ~A $\overset{\approx}{}_{C_1+C_2}$ ~B.

Proof: The proof follows from the property, $\underline{O_{C_1+C_2}}(\sim A) = \sim \overline{O_{C_1+C_2}}(A)$.

(3.6.10) A $\overset{\approx}{}_{C_1+C_2}$ ϕ, B $\overset{\approx}{}_{C_1+C_2}$ ϕ implies that $A \cap B \overset{\approx}{}_{C_1+C_2}$ ϕ.

Proof: The proof follows directly from the fact that under the hypothesis the only possibility is $\underline{O_{C_1+C_2}}(A) = \underline{O_{C_1+C_2}}(B) = \phi$.

(3.6.11) A $\overset{C_1+C_2}{\approx}$ U, B $\overset{C_1+C_2}{\approx}$ U implies that $A \cup B \overset{C_1+C_2}{\approx}$ U.

Proof: The proof follows directly from the fact that under the hypothesis the only possibility is $\overline{O_{C_1+C_2}}(A) = \overline{O_{C_1+C_2}}(B) = U$.

Replacement properties for covering based optimistic multi granular approximate equivalence

(3.7.1) (i) if $A \cap B \overset{C_1+C_2}{\approx}$ A and $A \cap B \overset{C_1+C_2}{\approx}$ B then A $\overset{C_1+C_2}{\approx}$ B.

(ii) The converse of (i) is not necessarily true.

Proof: (i) Here $\overline{O_{C_1+C_2}}(A)$ and $\overline{O_{C_1+C_2}}(A \cap B)$ are U or not U together and $\overline{O_{C_1+C_2}}(B)$ and $\overline{O_{C_1+C_2}}(A \cap B)$ are

U or not U together. Being common, we get $\overline{O_{C_1+C_2}}(A)$ and $\overline{O_{C_1+C_2}}(B)$ are U or not U together. So, A $\overset{C_1+C_2}{\approx}$ B.

(ii) Continuing with example 2 by taking A={x_1, x_2, x_3, x_4, x_6, x_7} and B={ x_2, x_3, x_4, x_6, x_7,x_8}, we have

THE IIOAB JOURNAL ISSN: 0976-3104

$\overline{O_{C_1+C_2}}(A)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}=U$ *and* $\overline{O_{C_1+C_2}}(B)=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}=U \Rightarrow A \overset{C_1+C_2}{\approx} B.$

But $A\cap B=\{x_2,x_3,x_4,x_6,x_7\}$. *Then* $\overline{O_{C_1+C_2}}(A\cap B)=\{x_2,x_3,x_4,x_5,x_6,x_7\}\neq U$
$\Rightarrow A\cap B$ *not* $\overset{C_1+C_2}{\approx}$ *A and B both.*

(3.7.2) (i) if $A\cup B \overset{\approx}{C_1+C_2}$ A and $A\cap B \overset{\approx}{C_1+C_2}$ B then A $\overset{\approx}{C_1+C_2}$ B.

(ii) The converse of (i) is not necessarily true.

Proof: (i) Here $\underline{O_{C_1+C_2}}(A)$ and $\underline{O_{C_1+C_2}}(A\cup B)$ are ϕ or not ϕ together and $\underline{O_{C_1+C_2}}(B)$ and $\underline{O_{C_1+C_2}}(A\cup B)$ are

ϕ or not ϕ together. Being common, we get and $\underline{O_{C_1+C_2}}(B)$ are ϕ or not ϕ together. So, A $\overset{C_1+C_2}{\approx}$ B.

(ii) Continuing with example 2 by taking A={x_6} and B={ x_7}we have

$\underline{O_{c_1+c_2}}(A)=\phi$ *and* $\underline{O_{c_1+c_2}}(B)=\phi \Rightarrow A \overset{\approx}{C_1+C_2} B.$ But $A\cup B=\{x_6,x_7\}$. Then

$\underline{O_{C_1+C_2}}(A\cup B)=\{x_6,x_7\}\neq\phi \Rightarrow A\cup B$ *not* $\overset{\approx}{C_1+C_2}$ *A and B both.*

(3.7.3) A $\overset{\approx}{C_1+C_2}$ A' and B $\overset{\approx}{C_1+C_2}$ B' may not necessarily imply that $A\cup B \overset{\approx}{C_1+C_2} A'\cup B'$.

Proof: When $\underline{O_{C_1+C_2}}(A), \underline{O_{C_1+C_2}}(B), \underline{O_{C_1+C_2}}(A'), \underline{O_{C_1+C_2}}(B')$ are all ϕ, one of $\underline{O_{C_1+C_2}}(A\cup B)$ *and* $\underline{O_{C_1+C_2}}(A'\cup B')$ is

ϕ but the other one is not ϕ the result fails to be true. Continuing example2 by taking A={x_6}, A'= {x_8},B={x_2}and B'={x_1},
we have

$\underline{O_{C_1+C_2}}(A)=\phi, \underline{O_{C_1+C_2}}(A')=\phi, \underline{O_{C_1+C_2}}(B)=\phi,$ *and* $\underline{O_{C_1+C_2}}(B')=\phi \cdot\cdot$

But $A\cup B=\{x_2,x_6\}$ *and* $A'\cup B'=\{x_1,x_8\}$

$\overline{O_{C_1+C_2}}(A\cup B)=\phi$ *and* $\overline{O_{C_1+C_2}}(A'\cup B')=\{x_1,x_8\}\neq\phi.$

$\Rightarrow A\cup B$ *not* $\overset{C_1+C_2}{\approx}$ $A'\cup B'$.

(3.7.4) A $\overset{C_1+C_2}{\approx}$ A' and B $\overset{C_1+C_2}{\approx}$ B' may not necessarily imply that $A\cap B \overset{C_1+C_2}{\approx} A'\cap B'$.

Proof: When $\overline{O_{C_1+C_2}}(A), \overline{O_{C_1+C_2}}(B), \overline{O_{C_1+C_2}}(A'), \overline{O_{C_1+C_2}}(B')$ are all U, one of $\overline{O_{C_1+C_2}}(A\cap B)$ *and* $\overline{O_{C_1+C_2}}(A'\cap B')$

is U but the other one is not U the result fails to be true.

Continuing with example 2 by taking A={x_1, x_2, x_3, x_4, x_5, x_7}, A'= { x_1, x_2, x_3, x_4, x_6, x_7, x_8},B={ x_2, x_3, x_4, x_5, x_7, x_8},

and B'={ x_2, x_3, x_4, x_6, x_7, x_8}, we have

$\underline{O_{C_1+C_2}}(A)=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}=U, \underline{O_{C_1+C_2}}(A')=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}=U,$

$\underline{O_{C_1+C_2}}(B)=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}=U,$ *and* $\underline{O_{C_1+C_2}}(B')=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}=U.$

But $A\cap B=\{x_2, x_3, x_4, x_5, x_7\}$ *and* $A'\cap B'=\{x_2, x_3, x_4, x_6, x_7, x_8\}$ $\underline{O_{C_1+C_2}}(A\cap B)=\{x_2, x_3, x_4, x_5, x$

and $\underline{O_{C_1+C_2}}(A'\cap B')=\{x_1, x_2, x_3, x_4, x_5, x_7, x_8\}=U \Rightarrow A\cap B$ *not* $\overset{\approx}{C_1+C_2}$ $A'\cap B'$.

(3.7.5) A $\overset{\approx}{C_1+C_2}$ B may or may not imply that $A\cup \sim B \overset{\approx}{C_1+C_2}$ U.

Proof: Continuing with example 2 by taking A={x_1, x_3, x_4, x_7, x_8} and B={x_1}, we have

$\overline{O_{C_1+C_2}}(A)=\{x_1,x_3,x_4,x_7,x_8\}$ *and* $\overline{O_{C_1+C_2}}(B)=\{x_1,x_8\}.$

~B={x_2, x_3, x_4, x_5, x_6, x_7, x_8} and $A\cup\sim B=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_7\}=U$

THE IIOAB JOURNAL ISSN: 0976-3104

$\underline{O_{C_1+C_2}}(A \cup \sim B) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} = U \Rightarrow A \cup \sim B \overset{C_1+C_2}{\approx} U$.

(3.7.6) A $\overset{C_1+C_2}{\approx}$ B may or may not imply that $A \cap \sim B \overset{C_1+C_2}{\approx} \phi$.

Proof: Continuing with example 2 by taking A=$\{x_5\}$ and B=$\{x_5, x_6, x_7\}$, we have
$\underline{O_{C_1+C_2}}(A) = \phi$ *and* $\underline{O_{C_1+C_2}}(B) = \{x_6, x_7\} \Rightarrow$ A and B is not top rough equal.
~B=$\{x_1, x_2, x_3, x_4, x_8\}$ and $A \cap \sim B = \phi$
$\underline{O_{C_1+C_2}}(A \cap \sim B) = \phi \Rightarrow A \cap \sim B \overset{\approx}{C_1+C_2} \phi$.
We would like to make the following comments in connection with the following properties from (3.7.7) to (3.7.11).

(i) We know that $\underline{O_{C_1+C_2}}(U) = U$. So, bottom $C_1 + C_2$ -equivalent to U can be considered under the case that
$\underline{O_{C_1+C_2}}(U) \neq \phi$.

(ii) We know that $\overline{O_{C_1+C_2}}(\phi) = \phi$. So, bottom $C_1 + C_2$ -equivalent to U can be considered under the case that
$\overline{O_{C_1+C_2}}(\phi) \neq U$.

The proofs of the properties from (3.7.7) to (3.7.11) are trivial and we omit them.

(3.7.7) If $A \subseteq B$ and B $\overset{C_1+C_2}{\approx} \phi$ then A $\overset{C_1+C_2}{\approx} \phi$.

(3.7.8) If $A \subseteq B$ and B $\overset{\approx}{C_1+C_2} U$ then A $\overset{\approx}{C_1+C_2} U$.

(3.7.9) A $\overset{\approx}{C_1+C_2}$ B iff ~A $\overset{\approx}{C_1+C_2}$ ~B.

(3.7.10) A $\overset{C_1+C_2}{\approx} \phi$, B $\overset{C_1+C_2}{\approx} \phi \Rightarrow A \cap B \overset{C_1+C_2}{\approx} \phi$.

(3.7.11) A $\overset{\approx}{C_1+C_2} U$, B $\overset{\approx}{C_1+C_2} U \Rightarrow A \cup B \overset{C_1+C_2}{\approx} U$.

CONCLUSION

The equality of sets in mathematics is too stringent and is mostly not applicable in real life situations. The problem in this definition is that although in real life situations we use our knowledge about the universe to decide about the equality of sets, which is mostly approximate in nature, we do not do so for set equality. As an attempt to incorporate user knowledge in equality, Novotny and Pawlak introduced the concept of rough equality and Tripathy et al introduced the concept of rough equivalence. The unigranular rough set concept introduced by Pawlak has been extended to define multigranular rough sets by Qian et al. Also, instead of using partitions, covers have been used to define covering based rough sets recently. In this paper we define and study the rough equality and rough equivalence in the context of covering based optimistic multigranular rough sets and to establish their properties in the general form as well as in the replacement form. We take the help of a real life example to illustrate the concepts and also to provide counter examples is establishing the properties.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

No financial support was received to carry out this project.

.

THE IIOAB JOURNAL ISSN: 0976-3104

REFERENCES

[1] SA Ade, Lin GP, Qian YH, Li J. [2011], A covering-based pessimistic multi-granulation rough set, in: Proceedings of International Conference on Intelligent Computing, August 11–14, Zhengzhon, China.

[2] Liu CH. Miao DQ. [2011], Covering rough set model based on multi-granulations, in: Proceedings of Thirteenth International Conference on Rough Sets, Fuzzy Set, Data Mining and Granular Computing, LNCS(LNAI) 6743 : 87–90.

[3] Liu CH, Wang MZ. [2011], Covering fuzzy rough set based on multi-granulation, in: Proceedings of *International Conference on Uncertainty Reasoning and Knowledge Engineering*, 2: 146–149.

[4] Liu Caihui Liu, Miao Duoqian, Quain Jin. [2012],On multi-granulation covering rough sets, International Journal of Approximate Reasoning, November.

[5] Novotny M and Pawlak Z [1985], Characterization of Rough Top Equalities and Rough Bottom Equalities, Bull. Polish Acad. *Sci Math*., 33: 91–97.

[6] Novotny M and Pawlak Z [1985], On Rough Equalities, Bull. Polish Acad. *Sci Math*, 33: 99–104.

[7] Pawlak Z. [1982], Rough sets, *Int. jour. of Computer and Information Sciences*,11: 341–356.

[8] Pawlak Z. [1991], Rough sets: Theoretical aspects of reasoning about data, Kluwer academic publishers (London).

[9] Qian YH, Liang JY. [2006], Rough set method based on Multi-granulations, Proceedings of the 5th *IEEE* Conference on Cognitive Informatics, 1:.297 – 304.

[10] Qian YH, Liang JY, Dang CY. [2007], MGRS in Incomplete Information Systems, *IEEE Conference on Granular Computin*:163 –168.

[11] Qian YH, Liang JY, Dang CY. [2010] Incomplete Multi-granulation Rough set, IEEE Transactions on Systems, Man and Cybernetics-Part A: *Systems and Humans*,.40(2): 420 – 431.

[12] Qian YH, Liang JY, Dang CY. [2010] Pessimistic rough decision, proceedings of RST 2010, Zhoushan, *China*: 440-449.

[13] Qian YH, Liang JY, Dang CY. [2010] MGRS: A multi-granulation rough set, *Information Sciences* 180 : 949–970

[14] Tripathi, Alka, and Kanchan Tyagi[2014], Approximate equalities using topological space, International Journal of Granular Computing Rough Sets and Intelligent Systems.

[15] Tripathy BK, Rashmi Rawat, Divya Vani Y, and Sudam Charan Parida [2014], Approximate Rough Equalities, *International Journal of Intelligent Systems and Applications* 6: 69–76.

[16] Tripathy BK and Anirban Mitra [2013], On Approximate Equivalences of Multigranular Rough Sets and Approximate Reasoning, *International Journal of Information Technology and Computer Science* 10: 103–113.

[17] Tripathy BK and Panda GK [2012] Approximate Equalities on Rough Intuitionistic Fuzzy Sets and an Analysis of Approximate Equalities, *International Journal of Computer Science Issues (IJCSI)* 9:371–380.

[18] Tripathy BK and M.Nagaraju [2012], On Some Topological Properties of Pessimistic Multigranular Rough Sets, *International Journal of Intelligent Systems and Applications* 8:10–17.

[19] Tripathy BK and M.Nagaraju [2012], A Comparative Analysis of Multigranular approaches and on Topological Properties of Incomplete Pessimistic Multigranular Rough Fuzzy sets, *International Journal of Intelligent Systems and Applications* 11:99–109.

[20] Tripathy BK and R.Raghavan [2011], On Some Topological Properties of Multigranular Rough Sets, *Journal of Advances in Applied science Research,* 2(3): 536–543.

[21] Tripathy BK and Mitra A. [2010],Topological Properties of Rough Sets and their Applications, *International Journal of Granular Computing, Rough Sets and Intelligent Systems (IJGCRSIS)*, (Switzerland),1:.4:355–369.

[22] Tripathy BK [2009] Rough sets on Fuzzy approximation spaces and Intuitionistic Fuzzy approximation spaces, Springer International studies in computational intelligence, vol.174, Rough Set Theory: A True landmark in Data Analysis, Ed: A. Abraham, R.Falcon and R.Bello: 3 – 44.

[23] Tripathy BK and G.K.Panda [2009], On Covering Based Approximations of Classifications of Sets, IEA/AIE 2009, LNAI 5579:777-786.

[24] Tripathy BK [2009] On Approximation of Classifications, Rough Equalities and Rough Equivalences, Springer International studies in computational intelligence, vol.174, Rough Set Theory: A True landmark in Data Analysis, Ed: A. Abraham, *R.Falcon and R.Bello*: 85 – 133

[25] Tripathy BK, H.K Tripathy. [2009] Covering Based Rough Equivalence of Sets and Comparison of Knowledge, Proceedings of the IACSIT Spring Conference 2009, Singapore, 17-20 April 2009:303-307.

[26] Tripathy BK, A. Mitra and J.Ojha[2008] On Rough Equalities and Rough Equivalence of Sets, , RSCTC 2008-Akron, U.S.A., Springer-Verlag Berlin Heidelberg 2008, LNAI 5306: 92–102.

[27] Yao YY. [2005], Perspectives of Granular Computing, Proceedings of 2005 *IEEE International Conference on Granular Computing*, I: 85–90.

[28] Yao YY, Yao B. [2012],Covering based rough set approximations, *Information Sciences* 200: 91–107.

[29] Zakowski W. [1993], Approximations in the space (U II), *Demonstration Mathematics* 16: 761–769.

THE IIOAB JOURNAL
ISSN: 0976-3104

ABOUT AUTHORS

M. Nagaraju *is a A.P.(SG), SCSE, VIT University at Vellore, India. He is doing his Ph.D. Degree in CSE under the supervision of Dr.B.K.Tripathy. His is working on topics like Rough sets, Granular Computing, Computational Intelligence Concepts, Soft Computing, Knowledge Engineering and Data Mining. He is a life member in ISTE and CSI.*

Dr. B.K. Tripathy *is a senior professor in the school of computing sciences and engineering, VIT University, Vellore, since 2007. He has produced 18 PhDs, 13 M.Phils and 02 M.S students so far. He has published around 260 papers in different international journals, conference proceedings and edited research volumes. He has edited two research volumes for the IGI publications and has written a book on Soft Computing. He is in the editorial board or review panel of over 60 international journals including Springer, Science Direct, IEEE and World Scientific publications. He is a life member/ senior member/member of 20 international forums including ACM, IEEE, ACEEE and CSI. His current interest includes Fuzzy Sets and Systems, Rough sets and Knowledge Engineering, Multiset Theory, List Theory, Data clustering and Database Anonymization, Content Based Learning, Remote Laboratories, Soft Set Analysis, Image Processing, Cloud Computing, content based learning and Social Network Analysis.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Priyadharsini et al

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A HYBRID ELM-WAVELET TECHNIQUE FOR THE CLASSIFICATION AND DIAGNOSIS OF NEUROMUSCULAR DISORDER USING EMG SIGNAL

Suja Priyadharsini[1*], Bala Sonia[1], Dejey[2]

[1]*Dept of Electronics and Communication,Regional Centre, Anna University, Tirunelveli, TN, INDIA*
[3]*Dept of Computer Science and Engineering,Regional Centre, Anna University, Tirunelveli, TN, INDIA*

ABSTRACT

Electromyogram (EMG) signal classification plays a major role in the diagnosis of neuromuscular disorder. The Motor Unit Action Potentials (MUAPs) in an electromyographic signal is one, which offers a significant source of information to evaluate the neuromuscular disorders. Neuromuscular diseases are the one that affect the control of muscular and nervous system. The proposed method employs a technique called Extreme Learning Machine (ELM) for the classification of EMG signal into healthy, myopathy or neuropathy. Discrete Wavelet Transform (DWT) / Wavelet Packet Transform (WPT) are the methods used for feature extraction individually. The performance of ELM together with DWT (ELM-DWT) and ELM with WPT (ELM-WPT) are compared with each other and it is found that the time complexity and number of feature vectors in ELM-WPT is reduced. The number of features is minimal in ELM-WPT compared with the ELM-DWT. The performance of ELM is evaluated using the confusion matrix and in terms of specificity, sensitivity, computational time and classification accuracy. The learning phase of ELM is completed in less than a second. The classification accuracy of ELM-WPT is 100%. The obtained result indicates that proposed ELM is very effective in the diagnosis of neuromuscular disorders.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th– June-2015
Published on: 20th –Aug-2015

KEY WORDS

Electromyogram (EMG); Motor Unit Action Potentials (MUAPs); Extreme Learning Machine (ELM); Discrete Wavelet Transform (DWT); Wavelet Packet Transform (WPT).

***Corresponding author:** **Email:** sujapriya_moni@yahoo.co.in

INTRODUCTION

The human skeletal muscular system consists of the nervous system and the muscular system, which together form the neuromuscular system. The disorders which originate in the nervous system, in the neuromuscular junctions, and in the muscle fibers are known as neuromuscular disorders. It has different degrees of severity ranging from minor loss of strength to amputation due to neuron or muscle death [1].Proper diagnosis of the disorder is of vital importance so that more focused treatment can be administered in the early stage [2]. Electromyography (EMG) signal is used for diagnosing patients with neuromuscular disorders. Pathological changes in the structure of motor units cause neuromuscular disorders, which can be classified into muscular (myopathy) and neuronal disorders (neuropathy) [3]. In myopathic disorders, the duration and the area to amplitude ratio of the action potential is reduced whereas in case of neuropathic disorder the duration and the area to amplitude ratio of the action potential is increased. In order to compensate low amplitude in myopathic disorder, a larger number of motor units are hired at lower than the normal levels of muscular contraction. In neurogenic disorder, the excited motor neurons are decreased in number and in order to keep a certain force of contraction, the available motor neuron must fire at higher rate than normal to balance the motor neuron loss. For an effective computerized EMG signal classification, an efficient treatment of EMG signals must be carried out [4]. The principle of this diagnostic system involves extraction of features from the acquired raw EMG signal which in turn helps in the diagnosis of neuromuscular disorder [1]. In this study, statistical features of Discrete Wavelet Transform (DWT) and Wavelet Packet Transform (WPT) have been used as a comparative study to characterize the EMG signal pattern in the diagnosis of neuromuscular disorder. These statistical features give major differences between healthy, myopathic and neurogenic signals and are useful for disease classification. The extracted statistical features are used as inputs to the ELM classifier. ELM is an efficient learning algorithm which performs well in classification application. This algorithm provides excellent performance at extremely fast learning speed [5]. ELM is an extensively used learning method that is capable of directly approximating nonlinear mappings by input data and provides models for a number of natural and artificial problems [6].

THE IIOAB JOURNAL
ISSN: 0976-3104

Englehart et al. [7] used the feature sets based on Short-Time Fourier Transform (STFT), Wavelet Transform (WT), and Wavelet Packet Transform (WPT) as an effective representation for EMG classification. The best performance is exhibited when using a combination of WPT, Principal Component Analysis (PCA), Linear Discrinminant Analysis (LDA), yielding an average classification error of 6.25%. Christodoulou and Pattichis [8] applied two different pattern recognition techniques for the classification of MUAPs. They compared the performance of two algorithms namely i) Artificial Neural Network (ANN) technique based on unsupervised learning, by means of a revised version of the Self-Organizing Feature Maps (SOFM) algorithm and Learning Vector Quantization (LVQ) ii) a statistical pattern recognition technique based on the Euclidean distance. It was found that the performance of ANN technique was better with the success rate of 97.6%. The advantage of this technique is that the learning is achieved in one epoch for both SOFM and LVQ algorithms. A key issue in LVQ is the choice of an appropriate measure of distance or similarity for training and classification. Subasi et al. [9] described the use of Autoregressive (AR) model with Wavelet Neural Network (WNN) to classify the EMG signals. They compared the performance of Feed forward Error Back Propagation Artificial Neural Network (FEBANN) and WNN based classifiers. The WNN performs better than FEBANN with a classification accuracy of 90.7%. The computational speed of WNN is faster when compared with Back Propagation Neural Network (BPNN). But the issue in this method is due to the unavailability of structured method to determine the optimum level of WNN factors, it is set by trial and error [10]. Katsis et al [11] proposed a novel method for the classification of MUAP's from the intramuscular EMG signals and obtained MUAPs classification accuracy of 86%. Abdulhmit Subasi [4] introduced the fuzzy based technique, ANFIS (Adaptive Neuro Fuzzy Inference System) for EMG signal classification and compared its performance with Multi Layer Perceptron Neural Network (MLPNN) and Dynamic Fuzzy Neural Network (DFNN). Among the three methods, it was reported that ANFIS performed better with an accuracy of 95%. In [1] Abdulhamit Subasi introduced PSO-SVM to improve the EMG signal classification and succeeded with the classification accuracy of 97.41%. Huang et al. [5] illustrated a new learning algorithm called Extreme Learning Machine (ELM) for Single-hidden Layer Feed forward Neural networks (SLFNs) which randomly chooses hidden nodes and analytically determines the output weights of SLFNs. Zong et al. [12] used ELM classifier to recognize the face and compared the result with that of SVM. It was reported that the recognition accuracy and training time of ELM was better than SVM with less optimization constraint. Liang et al. [13] used ELM to classify five mental tasks from different subjects using electroencephalogram (EEG) signals and compared the results with BPNN and SVM. It was found that ELM performed better than the other two techniques.

Neuromuscular disorders are those that affect the brain, spinal cord, nerves and muscles causing muscular weakness or muscle tissue wasting. During initial stages of the disease, pathological changes in the EMG signals are not much predominant which causes difficulty in the diagnosis of neuromuscular disorder. In such cases wavelet transforms can be used to characterize the localized frequency content of each MUP [4, 14, 15].

MATERIALS AND METHODS

Feature extraction methods

In the proposed method, ELM is used to classify the EMG signals as shown in Fig.1 and consists of three steps:

- The EMG signal is decomposed either using Discrete Wavelet Transform (DWT) or Wavelet Packet Transform (WPT) into different frequency bands.
- Statistical features are extracted from these sub-bands to represent each EMG signal.
- An unknown EMG signal is classified as Healthy or Myopathic or Neurogenic using the soft computing technique (ELM).

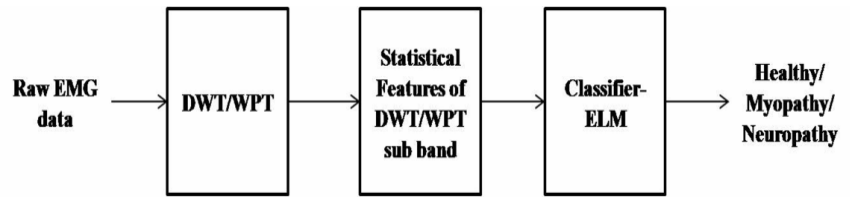


Fig: 1. Block diagram of the proposed method

THE IIOAB JOURNAL
ISSN: 0976-3104

This research work makes use of 30 EMG data collected from various subjects for analysis. Fig. 2 shows a sample of healthy, neuropathy EMG data.

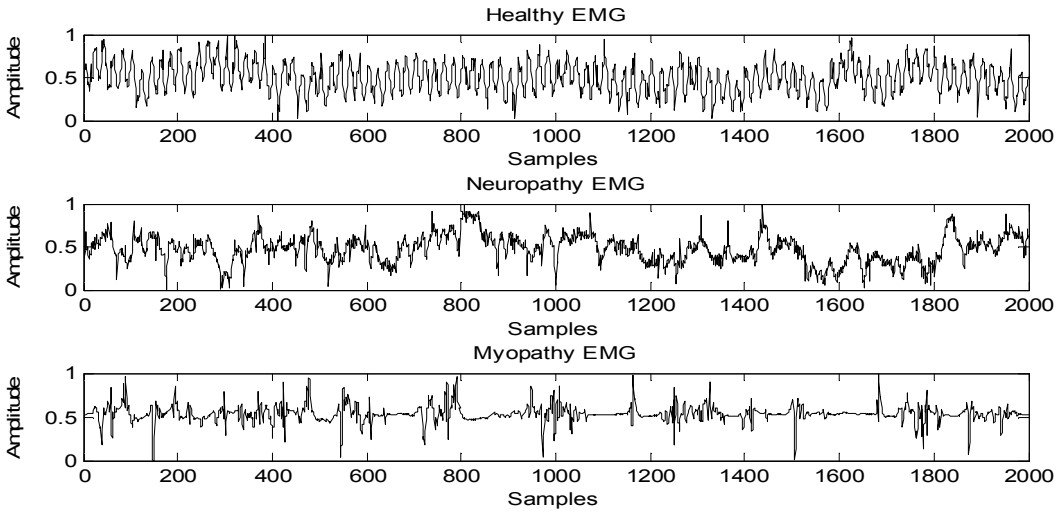


Fig: 2. Healthy, neuropathy and myopathy EMG data.

The choice of features of raw signals is very important for the accomplishment of any signal classification system. WT is one of the most capable methods to extract features from the EMG signals [4]. In this research work, WPT and DWT are used for feature extraction individually, from the EMG signal and their performances are compared with each other.

A) Feature extraction using discrete wavelet transform

The Wavelet transform (WT) is the one which provides both the time and frequency information at the same time. The wavelet transform decomposes a signal into a set of basic functions called wavelets that are obtained by dilations, contractions and shifts of an exclusive function called wavelet prototype translations [16, 17, 18, 19]. The wavelet transform in which the wavelets are sampled at discrete intervals is known as DWT [20]. As the time domain signal is passed through various high pass and low pass filters, the output of either is taken and the process is repeated. This process is known as decomposition. This continues until the signal decomposed to a pre-defined level. Thus, cluster of signals correspond to the same signal, but all related to different frequency bands. Lower and higher frequencies are better resolved in frequency and time respectively [4].

Repeated low-pass and high-pass filtering of the time domain signal results in the decomposition of the signal into different frequency bands. The down-sampled outputs of first high-pass and low-pass filters provide the detail, D1 and the approximation, A1 information respectively. The outputs of the high-pass and low-pass filters, are sub-sampled by 2. As the result of decomposition, the time resolution is halved and the frequency resolution is doubled, because the frequency band of the signal now spans only half the previous frequency band, successfully reducing the uncertainty in the frequency by half. The above procedure is also known as the sub-band coding which can be repeated for further decomposition. At each level, the filtering and sub-sampling results in half the number of samples and half the frequency band spanned. The first approximation A1 is further decomposed and this process is continued. In this research work, Daubechies 4 (DB4) wavelet filter is used for decomposition and reconstruction [4].

To represent the time-frequency distribution of the EMG signals, the following statistical features are used:

Mean values of the coefficients in each sub-band: The mean indicates the average value of a signal and is given by Equation (1)

$$\mu_{si} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

where x_i represents the corresponding coefficient; N denotes the total number of coefficients.

Variance of the wavelet coefficients in each sub-band: Variance indicates how far/close the data points are to the mean value of the coefficients and is computed as in Equation(2)

$$\mathrm{Var}(x_i) = E[x_i^2] - [E[x_i]]^2 \qquad (2)$$

Skewness of the coefficients in each sub-band: Skewness is a measure of extent to which the probability distribution of a real valued random variable leans to one side of the mean. The skewness value may be positive or negative and is given in Equation (3).

THE IIOAB JOURNAL
ISSN: 0976-3104

$$Skewness = \mu_3 / \sigma^3 \quad (3)$$

where μ_3 is third moment about mean; σ is the standard deviation.

Entropy of the coefficients in each sub-band. Entropy is a numerical measure of the randomness of a signal. Then entropy is given by Equation (4)

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (4)$$

where X is the random variable with n outcomes $\{x_1, x_2, x_3, \ldots, x_n\}$.

Standard Deviation of the coefficients in each sub-band: The standard deviation (σ) shows how much variation or dispersion from the average exists. A large value of σ indicates that the data points are far from the mean and a small value of σ indicates that they are clustered closely around the mean. σ is calculated using the Equation (5)

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \quad (5)$$

where μ is the mean, x_i denotes the corresponding coefficient.

Ratio of absolute mean values of adjacent sub-band: It denotes the ratio of the mean values of the adjacent sub-bands which is given by the Equation (6)

$$Ratio = \mu_{ai} / \mu_{ai+1} \quad (6)$$

where μ_{ai} represents the absolute mean of a coefficient, μ_{ai+1} is the absolute mean value of its adjacent coefficient.

Two sets of features, feature set 1(F1) and feature set 2 (F2) are used as input to ELM-DWT individually. F1 is the one, which is used in the existing method [1] .To compare the performance of the proposed method with the existing method; F1 is included in the present work. F1 comprises of features such as mean of absolute values of coefficient in each sub-band, standard deviation of the wavelet coefficient in each sub-band, ratio of absolute mean values of adjacent sub-bands and F2 comprises mean, variance, skewness and entropy of the wavelet coefficient in each sub-band.

Thus by extracting 6 different values for mean and standard deviation, 5 different values for ratio of absolute mean values of adjacent sub-band, a total of 17 features are extracted for F1. Similarly, by extracting 6 different values for mean, variance, skewness and entropy of the coefficients in each sub-band, a sum of 24 features are extracted in case of F2. These features are calculated from the frequency bands D1, D2, D3, D4, D5 and A5 of DWT and then used as input to ELM classifier.

B) Feature extraction using wavelet packet transform (WPT)

In DWT decomposition, a signal is decomposed into two frequency bands such as lower frequency band (approximation coefficients) and higher frequency band (detail coefficients) and the low frequency band is used for further decomposition, thus it gives a left recursive binary tree structure. In case of Wavelet packet Transform (WPT), a balanced binary tree structure is generated since both lower and higher frequency bands are decomposed into two sub-bands. It helps to divide the high frequency side into smaller bands which cannot be achieved by using DWT [21]. In our analysis, Daubechies 2 (db2) family of wavelet packets is implemented as the mother wavelet. As a result of 3 level decompositions, 8(2^3) feature vectors are extracted from each signal frame and are used as input to ELM classifier for the classification of the EMG signals. The feature extracted from the wavelet packet is energy feature and it is given by Equation (7)

$$E = \sum_i |x_i|^2 \quad (7)$$

where x_i denotes the wavelet packet coefficient.

Classifier- extreme learning machine

Guang-Bin Huang introduced ELM which is a Single-hidden Layer Feed forward Neural network (SLFN) with at most L hidden nodes and with almost any nonlinear activation function can exactly learn L distinct observations. SLFNs can be considered as a linear system after the input weights and the hidden layer biases are chosen randomly. The output weights (linking the hidden layer to the output layer) of SLFNs can be analytically determined through simple generalized inverse operation of the hidden layer output matrices. ELM can approximate any target continuous function and classify any disjoint regions [5]

THE IIOAB JOURNAL
ISSN: 0976-3104

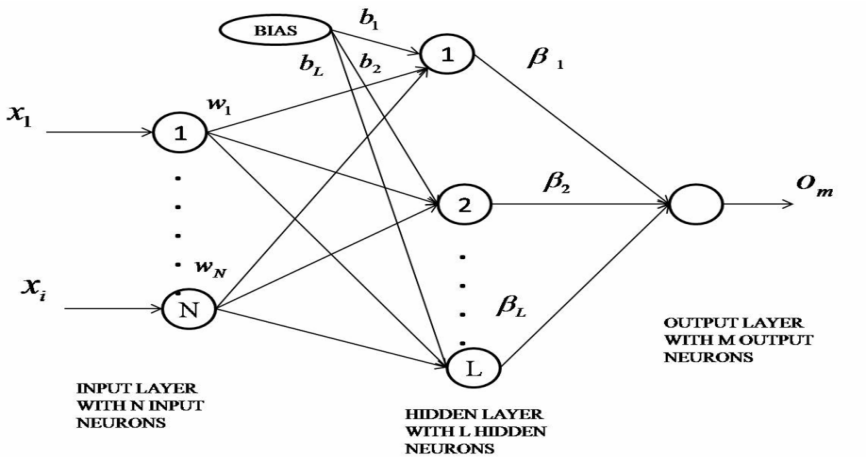


Fig: 3. ELM Architecture.

Fig: 3 shows the ELM architecture which consists of single hidden layer. The number of input neuron corresponds to the number of input features. The number of hidden neurons is equal to or greater than the input neurons. The number of output neuron is equal to the number of classes. For N arbitrary distinct samples (x_i, t_i) where $x_i = [x_{i1}, x_{i2} \ldots\ldots x_{in}]^T$ and $t_i = [t_{i1}, t_{i2} \ldots\ldots t_{in}]^T$, standard SLFNs with L hidden nodes and activation function $g(x)$ are mathematically modeled as in Equation (8)

$$\sum_{j=1}^{L} \beta_j g_j(x_i) = \sum_{j=1}^{L} \beta_j g_j(w_j x_i + b_j) = o_i \quad \text{where i=1,2, ...N} \qquad (8)$$

Where $w_j = [w_{j1}, w_{j2} \ldots\ldots w_{jn}]^T$ is the weight vector connecting the j^{th} hidden node and the input nodes, $\beta_j = [\beta_{j1}, \beta_{j2} \ldots\ldots \beta_{jm}]^T$ is the weight vector connecting the j^{th} hidden node and the output nodes, and b_j is the threshold of the j^{th} hidden node. $wj.xi$ denotes the inner product of w_j and X_i. That standard SLFNs with L hidden nodes with activation function g(x) can approximate these N samples with zero error means

i.e., $\sum_{j}^{L} = 1 \|o_j - t_j\| = 0$ and there exist w_j and β_j such that

$$\sum_{j=1}^{L} \beta_j g(w_j x_i + b_j) = t_j \quad j = 1,2,3,\ldots L \qquad (9)$$

The Equation (9) can be written compactly as shown in Equation (10)

$$T = \beta H \qquad (10)$$

Where H, β, T are given in Equation(11), Equation (15) and Equation (16)

$$H(w_1, \ldots\ldots, w_L, b_1, \ldots\ldots, b_L, x_1, \ldots, x_N) = \begin{bmatrix} g(w_1 x_1 + b_1) & \cdots & g(w_L x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 x_N + b_1) & \cdots & g(w_L x_N + b_L) \end{bmatrix} \qquad (11)$$

$g(w_j, x_i, b_j)$ is the activation function and it may be Unipolar sigmoid, Bipolar sigmoid or Gaussian . H is called the hidden layer output matrix of the neural network; the j^{th} column of H is the j^{th} hidden node output with respect to inputs $x_1, x_2 \ldots\ldots x_N$ [5].

Unipolar sigmoid activation function is calculated using the formula as given in Equation (12)

$$g(w_j, x_i, b_j) = \frac{1}{1 + e^{(-([w_j x_i] + b_j) - x)}} \qquad (12)$$

Bipolar sigmoid activation function is represented as in Equation (13)

$$g(w_j, x_i, b_j) = \frac{1 - e^{(-([w_j x_i] + b_j) - x)}}{1 + e^{(-([w_j x_i] + b_j) - x)}} \qquad (13)$$

Gaussian function is calculated using the Equation (14)

$$g(w_j, x_i, b_j) = e^{(-b_j * \|x_i - w_j\|)} \qquad (14)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

where λ is the learning rate whose range is 0-1.

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m} \quad (15)$$

$$T = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad (16)$$

RESULT AND DISCUSSION

Data acquisition

EMG data are collected from 15 healthy subjects (fourteen females and one male) with ages ranging from 22 to 42 years, 1 myopathy subject (male) with age 75 and 11 neuropathy subjects (seven males and four females) with ages ranging from 55 to 65. The signal was acquired by placing 3 electrodes on the biceps using Labview software and DAQ with a sampling rate of 1 KHz and 16 bit resolution. In addition, 1 healthy, 1 myopathy and 1 neuropathy EMG data are obtained from physionet [22].

Selection of Wavelet and level of decomposition plays a very important role in the extraction of features. In the existing method mentioned [1], four statistical features namely Mean of the absolute values of the coefficients, Average power of the wavelet coefficients in each sub-band, Standard deviation of the coefficients from each sub-band and Ratio of the absolute mean values of adjacent sub-bands were the features extracted for classification using PSO-SVM classifier and it resulted in a classification accuracy of 97.41%.

In the proposed method, ELM uses the features extracted through DWT and WPT separately for the classification of EMG signal. The performance of the methods ELM-DWT and ELM-WPT are compared with each other.

ELM-DWT

The performance of ELM-DWT is evaluated for two sets of features F1 and F2 as mentioned in section 2.1. In order to compare the proposed method with the existing method, the first feature set F1(17 features) is classified using ELM which yielded a classification accuracy of 100%.The number of feature vectors used in the proposed method is 17, whereas 23 feature vectors were used in the work mentioned in [1] . As the number of features used for classification reduces, the computational time reduces. Hence classification with limited number of features is advantageous. The performance of the proposed method is also evaluated with second feature set F2 (24 features), which again yielded a classification rate of 100%.

ELM-WPT

In this method, EMG signal is decomposed into 3 levels using Wavelet Packet Transform (WPT) with Daubechies order 2 (db2) wavelet. Thus $2^3 = 8$ numbers of packets are obtained. Wavelet packet energy feature is computed for each packet. Thus a total of 8 feature vectors have been extracted using WPT and classified using ELM which again yielded a classification accuracy of 100%.Though both methods ELM-DWT and ELM-WPT yielded a result of 100%, the latter method is computationally efficient because of its ability to classify with a limited number of features.

Theoretically the number of nodes in the hidden layer of the classifier is equal to or greater than the number of input nodes. If the number is too small, ELM may not reflect the relationship between input data and output value. On the contrary, a large number may create such a complex network that might lead to a very large output error caused by

THE IIOAB JOURNAL
ISSN: 0976-3104

over-fitting of the training sample. The parameters used for ELM-DWT and ELM-WPT are shown in **Table– 1**. The activation function used in the classifier may be Unipolar sigmoid or Bipolar sigmoid or Gaussian. The activation function for which ELM yielded 100% classification accuracy is shown in **Table– 2**. From the collected EMG data, 5 data (2 healthy, 1 myopathic and 2 neurogenic) are used for training and 25 data (14 healthy, 1 myopathic and 10 neurogenic) are used for testing.

Table: 1. Parameters used in ELM

Parameter	ELM-WPT	ELM-DWT	
		Feature set 1	Feature Set 2
Number of Input node	8	17	24
Number of Hidden node	8	20	30
Number of Output node	3	3	3
Learning Rate	0.001	0.001	0.001

Table: 2. Various activation functions used in ELM

Method		Unipolar Sigmoid	Bipolar Sigmoid	Gaussian function
ELM-WPT		×	×	×
ELM-DWT	Feature Set1			×
	Feature Set2	×		

The classification accuracy of the ELM classifier is 100 % each. Irrespective of the number of training data sets, ELM performs well and guarantees an accuracy of 100%.Classification results of the ELM are visualized using confusion matrix [23]. The confusion matrix showing the classification results of ELM is shown in **Table–3**.

Table: 3. Confusion matrix of ELM classifier

Output/desired	Result (Healthy)	Result (Myopathy)	Result (Neuropathy)
Result (Healthy)	14	0	0
Result (Myopathy)	0	1	0
Result (Neuropathy)	0	0	10

DISCUSSION

The computation of the following parameters showed the test performance of ELM:

Specificity: Number of correctly classified healthy subjects/ number of total healthy subjects

Sensitivity (myopathy): Number of correctly classified subjects suffering from myopathy/number of total subjects suffering from myopathy.

Sensitivity (neuropathy): Number of correctly classified subjects suffering from neuropathy disorder/number of total subjects suffering from neuropathy disorder.

Classification accuracy: number of correctly classified subjects/ number of total subjects.

$$\text{Classification accuracy} = \sum {}^{D_c}/_{N} \qquad (17)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

where, Dc is the diagonal elements of the confusion matrix, N is the total number of training/testing samples. The values of these parameters are tabulated in **Table–4**. The ELM classified healthy subjects, myopathy subjects and subjects suffering from neuropathy with the accuracy of 100%. Table .4 shows the classification success rate obtained for 30 EMG recordings.

Table: 4. Comparison of ELM-WPT and ELM-DWT models for EMG signal classification

Statistical Parameters	ELM-WPT	ELM-DWT	
		Feature set 1	Feature Set 2
Computational Time (sec)	12.678	14.522	17.427
Specificity	1	1	1
Sensitivity (neuropathy)	1	1	1
Sensitivity (myopathy)	1	1	1
Classification Accuracy (%)	100	100	100

CONCLUSION

Electromyography plays an important role in clinical neurological diagnosis, to indicate the location and type of abnormality or expose disorders that are clinically uncertain. The classification of neuromuscular disorders is essential for correct diagnosis. The proposed method, Extreme Learning Machine (ELM) is a simple and effective algorithm for single-hidden layer feed forward neural networks (SLFN) which automatically classifies the EMG signal into healthy, myopathic or neuropathic. For any classification application, feature extraction is necessary and in this research work DWT, WPT are the two used individually. The proposed techniques ELM-DWT and ELM-WPT are found to be best when compared with the existing techniques in terms of classification accuracy and number of feature vectors. Among the proposed techniques, ELM-WPT excels ELM-DWT in terms of computational time and number of features. The learning speed of ELM is extremely fast. From the simulation result, it can be seen that the learning phase of ELM is completed in less than a second. The ELM appears to be suitable in applications which require fast prediction and response capability. For any type of activation function, the classification accuracy of ELM-WPT is 100%.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors are grateful to Dr.V. Manakavalaperumal M.D, Sugam Health Centre, Tirunelveli, Tamil Nadu, India for providing EMG data and validating the results.

FINANCIAL DISCLOSURE

This research work was funded by Centre for Technology Development and Transfer (CTDT), Anna University, Chennai- 600025, Tamil Nadu, India under research support scheme.

REFERENCES

[1] Subasi A.[2013] Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. *Computers in Biology and Medicine*, 43(5): 576–586.

[2] Alkan A, Gunay M. [2012] Identification of EMG signals using discriminant analysis and SVM classifier. *Expert Systems with Applications,* 39(1): 44–47.

[3] SuitornsaneeS.[2011] Classification of EMG using Reccurance Quantification Analysis. *Procedia Computer science* 6;375–380.

[4] SubasiA.[2012]Classification of EMG signals using combined features and soft computing techniques. *Applied Soft Computing*, 12(8);2188–2198.

[5] HuangGB.[2006]Extreme learning machine: Theory and applications. *Neurocomputing* 70: 489–501.

[6] WangY.[2011] A study on effectiveness of extreme learning machine. *Neurocomputing*74 (16):2483–2490,

[7] EnglehartK, HudginsB, Parker PA, Stevenson M.[1999] Classification of the myoelectric signal using time-frequency based representations.*Medical Engineering & Physics* 21(6-7): 431–438.

[8] ChristodoulosI. Christodoulou and Constantinos S. Pattichis[1999] Unsupervised Pattern Recognition for the Classification of EMG Signals. *IEEE Transactions on Biomedical Engineering*, 46(2);169–178.

[9] SubasiA, YilmazM, OzcalikHR.[2006] Classification of EMG signals using wavelet neural network.*Journal of Neuroscience Methods* 156(1-2):360–367.

[10] OtokBW, Suhartono, Brodjol S. Ulama.S, EndhartaAJ [2011] Design of experiment to optimize the architecture of wavelet neural network for forecasting the tourist arrivals in Indonesia. *Communication in Computer and Information science*, 253;14–23.

[11] Katsis CD, GoletsisY, LikasA, FotiadisDI,SarmasI.[2006] A novel method for automated EMG decomposition and MUAP classification, *Artificial Intelligence in Medicine,* 37(1): 55–64.

[12] ZongW, HuangGB.[2011]Face recognition based on extreme learning machine, *Neurocomputing*, 74; 2541–2551.

[13] LiangNY, SaratchandranP, HuangGB ,SundarajanN.[2006]Classification of mental tasks from EEG signals using extreme learning machine, *International Journal of Neural Systems,* 16 (1);29–38.

[14] Abel EW, MengH, ForsterA,HolderD.[2006]Singularity charecteristics of needle EMG IP signals.*IEEE Transactions on Biomedical Engineering*, 53(2);219–225.

[15] PattichisCS, PattichisMS.[1999]Time-scale analysis of motor unit action potentials, IEEE transactions on Biomedical Engineering, 46(11): 1320–1329.

[16] SubasiA. [2005] Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients, Expert Systems with Applications, 28(4):701–711.

[17] CohenA, KovacevicJ.[1996] Wavelets: The mathematical background, *Proceedings of the IEEE*, 84(4) :514–522.

[18] DaubechiesI.[1996] Where do wavelets come from? A personal point of view, *Proceedings of the IEEE*, 84(4);510–513.

[19] RioulO, VetterliM.[1991] Wavelet and signal processing, *IEEE Signal Processing Magazine*, 8(4), 14–46.

[20] RinkyBP, MondalP, ManikantanK, RamachandranS.[2012]DWT based feature extraction using edge tracked scale normalization for enhanced face recognition, *Procedia Technology*, 6;344 – 353.

[21] PolikarR, The Wavelet Tutorial http:// users.rowan.edu/polikar/ Wavelets/WTtutorial.html.

[22] Electromyogram Database, http://physionet.org/ physiobank/ examples of Electromyogram.html.

[23] HariharanM, FookCY, SindhuR, IliasB, YaacobS. [2012]A comparative study of wavelet families for classification of wrist motions, *Computers and Electrical Engineering*, 38(6);1798–1807.

ABOUT AUTHORS

Prof. Mrs.S.Suja Priyadharsini *received her B.E degree from Manonmaniam Sundaranar University , Tirunelveli. She received her M.E degree in M.E Appplied Electronics from Anna University, Chennai, india. She is pursuing her Ph.D from Anna University, Chennai. She is working as a Assistant Professor in Electronics and Communication Engineering Department since 2008 in Regional Centre, Anna University Tirunelveli Region, Tirunelveli. India. Her main research interest includes Signal Processing, Bio-medical Signal Processing and Soft Computing*

Ms. R.Bala Sonia *received her B.E degree from Anna University, Chennai. India. She was the student of M.E Applied Electronics in Regional Centre, Anna University Tirunelveli Region, Tirunelveli. India. Her research interest includes Bio-medical Signal Processing and Soft Computing*

Dr. Dejey *received her B.E. and M.E. degrees in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, India, in 2003 and 2005, respectively. Later, she was with the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India, as a Junior Research Fellow under the UGC Research Grant. She completed her Ph.d in Computer Science and Engineering in 2011. She has been with the Department of Computer Science and Engineering, Regional Centre, Anna University: Tirunelveli Region as an Assistant Professor since 2010 and as the Head of the Department since 2011. Her research interests include image and signal processing, watermarking,information hiding and multimedia security.*

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Jayaraman and Velumani

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A HYBRID PSO-GA APPROACH FOR BIOMARKER DISCOVERY

Ramsingh Jayaraman and Bhuvaneswari Velumani*

Department of Computer Applications, Bharathiar University, Coimbatore, INDIA

ABSTRACT

The advancement in genomic and proteomic has paved way for identifying methodologies to identify gene involved in life threatening diseases. Biomarker refers to specific gene and products with biochemical features to measure the progress of the diseases. Various soft computing techniques are applied to identify biomarkers from large micro array chips. In this paper a hybridized approach for biomarker discovery using Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) is proposed. The proposed approach is tested with microarray lymphoma dataset. Based on the experimental run twelve gene were identified as significant gene and 3 gene were found in the disease pathway which can be verified with experts to name as biomarkers using the methodology proposed. The experimental results are validated using panther tool for biological significance and verified with related literature papers. From the results it is found the performance of the algorithm is good for identifying significant gene for bio marker discovery.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th-June-2015
Published on: 20th-Aug-2015

KEY WORDS

Biomarker, PSO, GA, Gene Pathway, Significant gene Gene Ontology (GO)

***Corresponding author:** **Email**: bhuvanes_v@yahoo.co.in, j.ramsingh@hotmail.com, **Tel:** 9500897389

INTRODUCTION

DNA microarray is a rapid growing technology used for the study of gene expression profiles and it is a classic probe hybridization method which provides access to thousands of gene at once. Generally, microarray data are images, which are transformed into gene expression matrixes in which rows represent gene and columns represent the expression values of gene under various experimental conditions. Microarray technologies enable the simultaneous interrogation of the expression level of thousands of gene to obtain a quantitative assessment of their differential activity in a given tissue or cell [1]. Recently the microarray technology has a gained interest of their use in clinical trials and disease diagnosis for identifying genomic factors that are prognostic for prediction or identification of Biomarker.

Biomarkers (short for biological markers) are biological measures of a biological state. By definition, a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention [2]. Biomarkers are the measures used to perform a clinical assessment such as blood pressure or cholesterol level and are used to monitor and predict health states in individuals or across populations so that appropriate therapeutic intervention can be planned [3-5]. Biomarker discovery is proved as one of the most broadly applicable and successful means of translating molecular and genomic data into clinical practice [6]. Evolutionary algorithm and soft computing techniques are applied in large for analysis of gene profiles for Biomarker prediction and identification. Genetic algorithms are hybridized with bio-inspired algorithms for microarray analysis to identify the significant gene. Identification of significant gene from large gene set becomes essential to identify the biomarkers of specific diseases [7]. The authors Zhang F [8], Hui-Huang Hsu [9] used data mining to identify biological markers for the diagnostic classification and prognostic assessment in the context of microarray and proteomic data A hybridized Particle Swarm Optimization and Genetic Algorithm PSO-GA is proposed in this paper to identify significant gene from DNA microarray for biomarker discovery. The proposed work is implemented using R-language for Lymphoma cancer dataset. The paper is organized as follows. Section II discusses the literature related to the work. Section III describes the hybridized approach for identifying biomarker using PSO with Genetic algorithm. In Section IV the experimental results are discussed followed by Conclusion in section V,

THE IIOAB JOURNAL
ISSN: 0976-3104

followed by Limitations and Future scope in section VI.

LITERATURE STUDY

This section discusses the literature related to the work. The literature are discussed in three sub sections namely Preprocessing techniques of microarray data for significant gene analysis, literature related to evolutionary algorithm and hybridized approaches.

Statistical techniques for microarray data

Statistical measures are widely used to identify the most and least selected attributes from large samples using various statistical tests as F-Test, T-Test, Chi-Square and Annova. In microarray dataset the gene are viewed as random samples and rank based statistical measures are widely used to identify the best expressed gene profiles from microarray data. The literature for ranking gene as part of pre-processing of microarray data is discussed in this section.

Statistical approaches like k Nearest Neighbor, Iterative regression imputation, Mean Squared Error, F-Test, T-test, chi-square, ANOVA and F-Test are used to pre-process the microarray dataset to impute missing values and rank gene. The authors Miguel Rocha and Isabel Rocha [4] suggested substituting of missing values with mean, median and mode using imputation method k nearest neighbor for microarray dataset. M. Templ et al., [10] used KNN methods for estimating missing values in compositional data. The authors also have discussed about T-test and ANOVA for identifying differentially expressed gene in microarray data. Matt Blackwell [11] analyzed Multiple Hypothesis Testing F-test in microarray data and S. N. Mukherjee, S. J. Roberts et al., [12] proposed a gene-ranking algorithm whose main novelty is the use of bootstrapped P-value for microarray datasets. The authors Guoqiang Yu, et.al, [13] used MSE (Mean Squared Error) to select candidate gene for classification. MSE is minimized to achieve the desired output (class target).Yoko Omura and Jun Sese [14], used Mean Square Error to select significant gene and generates pClusters from the selected gene. Mohd Sazli Saad [15], used MSE as objective function to calculate fitness value of each chromosome represented as gene in Genetic algorithm to pass the best chromosome for the consecutive generations.

PSO – GA

Particle swarm optimization is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality [16]. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity [17]. Each particle's movement is influenced by its local best known position but, is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles [18]. This is expected to move the swarm toward the best solutions. The algorithm is very simple but powerful.

The genetic algorithm is a model of machine learning which derives its behavior from a metaphor of the processes of evolution in nature Genetic algorithm are found to be widely used microarray data to improve the classification accuracy [19] . Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population [20]. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce [21]. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. Laetitia Jourdan et. al.[4] they have proposed a 2-phase approach using a specific genetic algorithm for the feature selection problem, they used a genetic algorithm (GA). Linyu Yang [19] and C.H. Ooi and Patrick Tan (2003) [22] used Genetic algorithm to design a gene-selection scheme predict the disease. J. Yu, et. al.,[23] and Edmundo Bonilla Huerta [6] applied Genetic Algorithm to find the gene subsets that produce high prognostic classification accuracy.

The authors Barnali Sahua et.al. [16] and Keisuke Kameyama [18] used Particle swarm Optimization (PSO) as a optimization technique for the classification of high dimensional cancer microarray data. Qinghai Bai [21],

THE IIOAB JOURNAL
ISSN: 0976-3104

concluded that performance of PSO is fast, easy and effective for optimization problem compared to other approaches and author strongly suggested that, PSO provides better result when combined with other techniques. Yuji Zhang et.al, [24] proposed a method of hybrid of genetic algorithm and particle swarm optimization (GA-PSO) to train the NN (Neural Network) models. The authors C.H. Ooi and Patrick Tan [8], Linyu Yang ,et.al., [19] applied genetic algorithm for multi-class prediction problem. The authors stated that the parallelized searching capability of GA helps to design a gene-selection scheme that determines the optimal set of gene in a multiclass dataset which classifies the samples within the dataset with minimal error and accurate prediction. Shital C. Shah [25] described that, GA based approach uniquely identified some gene/SNPs that could not be identified by the approaches like Information Gain and REG Regression approaches. Laetitia Jourdan et. al. [26] Proposed a 2-phase approach using a specific genetic algorithm with clustering algorithm for multi factorial disease prediction. The authors Ramsingh et.al. [27] Proposed genetic algorithm with K-Means clustering to identify significant gene from microarray dataset. The authors Vanitha et.al, [28] used evolutionary approaches such as optimization methods to generate best gene combinations to achieve higher level classification accuracy.

The biomarker gene are some specific gene which is highly significant causing gene mutation resulting in specific diseases. Indentifying biomarker gene can be viewed as an optimization problem to identify minimum number of gene which are highly significant bases on their gene expression profiles. From the literature study it is found that PSO, GA and other approaches are applied as standalone approaches for identifying significant gene from microarray dataset. These approaches when applied and iterated for more number of runs identify the best gene expression profiles which are highly significant. Biomarker identification requires accurate validation of the gene expression profiles to identify the gene in specific diseases. In this paper a novel methodology is proposed to filter the optimal gene using a two level optimization process using the hybrid PSO-GA approach to identify the significant gene to discover biomarkers for diseases.

BIOMARKER DISCOVERY USING PSO-GA

This section discusses the proposed methodology and framework to identify genomic biomarker for microarray dataset. The framework for Biomarker discovery using PSO-GA given in **Figure−1** is consists of three phases Pre Processing phase, phase II as PSO-GA hybrid approach and verification and validation phase as the third phase.

Phase I

Pre-processing phase is the first phase used for analyzing Gene expression features of microarray dataset using statistical techniques. Microarray data contains of noisy and inconsistent data. The Pre-processing techniques are used to reduce noisy and inconsistent data using imputation methods. k Nearest Neighbor (kNN) is a data mining technique to impute missing expression data in microarray in the proposed work to fill in the missing values. For each gene with missing values, it finds the k Nearest Neighbor using a distance metric, confined to the columns for which those gene are not missing. The microarray data is normalized using the statistical techniques.

Based on the detailed study and analysis of various statistical methods for microarray data through experimental result and comparison MSE is found to identify highly significant gene which are semantically relevant compared to other approaches from the literature study. MSE is used as fitness function to rank significant gene. The mean squared error is arguably the most important criterion used to evaluate the performance of a predictor or an estimator. (The subtle distinction between predictors and estimators is that random variables are predicted and constants are estimated.) The mean squared error is also useful to relay the concepts of bias, precision, and accuracy in statistical estimation.

$$MSE \quad \sum_{i=1}^{n}(x_i x)^2 \qquad \text{----Eq 1}$$

n = size (number of individuals)

$\mathbf{x}$ = mean value of individuals

$\mathbf{x}$ = value of individual

THE IIOAB JOURNAL
ISSN: 0976-3104

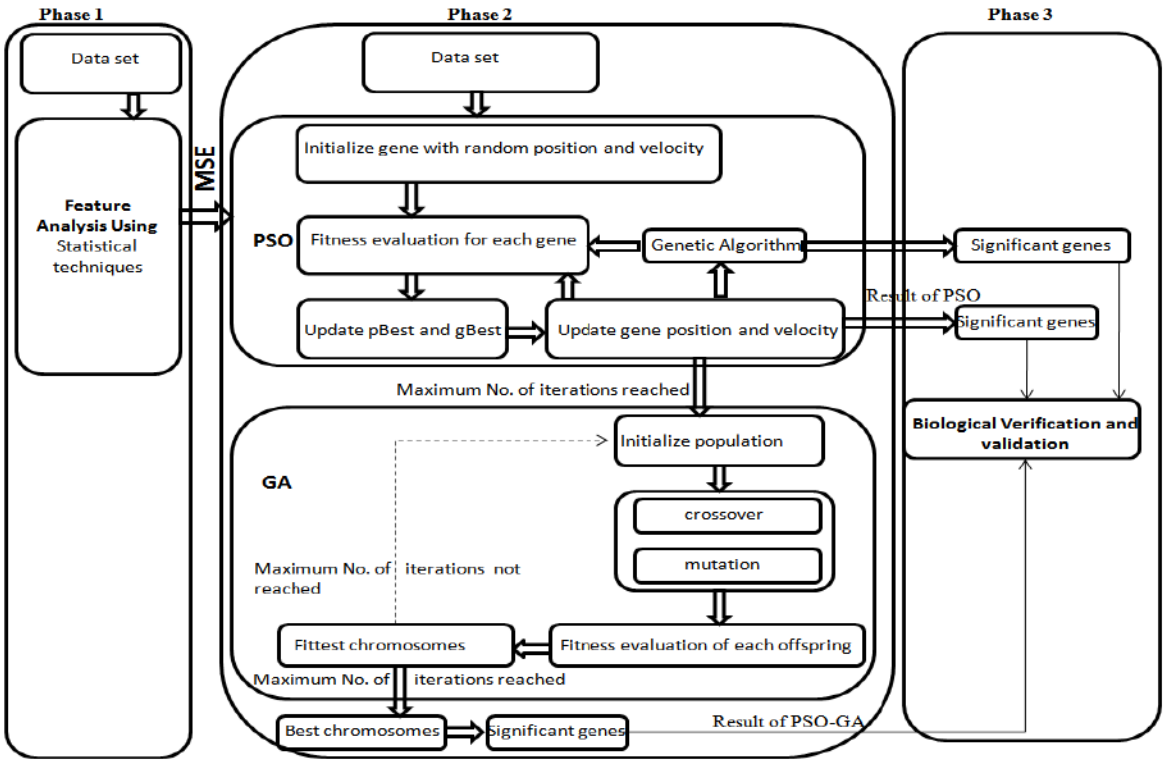


Fig: 1. Biomarker discovery using PSO-GA

Phase II: PSO – GA hybrid approach

The second phase of the framework is presented with the methodology to identify significant gene for biomarker discovery using hybridized PSO - GA algorithm. In this methodology the PSO algorithm is applied initially on the dataset and best populated gene are passed as input to GA process to identifying the best significant gene, as a two level optimization process the detailed methodology is presented in the below section.

PSO algorithm – parameters

In PSO algorithm the particles are initialized at random positions, to explore the search space to find better solutions. In each iteration each particle adjusts its velocity to follow two best solutions. The first is the cognitive part, where the particle follows its own best solution found so far. This is the solution that produces the lowest cost (has the highest fitness). This value is called pBest (particle best). The other best value is the current best solution of the swarm, i.e., the best solution by any particle in the swarm. This value is called gBest (global best).

In the proposed PSO algorithm gene expression values in microarray dataset are encoded as a particle. The fitness value of the particle is calculated using MSE given in Eq 1 gBest and pBest are calculated in each iteration based on the fitness value of gene, where gBest is the best fitness value of all gene in the entire population and pBest is the fitness value of gene in each population. the velocity and position of the particles are calculated using Eq 2 and Eq 3. The best ranked gene are passed to the GA process. The best ranked gene are removed from the PSO population for next iteration. This cycle is iterated until there is minimum number of gene in the population. The parameters of PSO algorithm is given in **Table–1**

$$\text{New velocity } v' = v + c1.r1.(pBest - x) + c2.r2.(gBest - x) \quad \text{----Eq 2}$$

$$\text{New position } x' = x + v' \quad \text{----Eq 3}$$

$$\text{New position } x' = x + v'$$

THE IIOAB JOURNAL
ISSN: 0976-3104

v = current velocity
x = current position
c = coefficient
$r = \text{random values}, 0 >= r1 <= 1 \text{ and } 0 >= r2 <= 1, 0 <= c1 <= 2, \text{and } 0 <= c2 <= 2$

The parameters for Hybridized PSO - GA algorithm for PSO Approach is given in **Table-1**.

Table: 1. Parameters of PSO

No. of particles	4026
Fitness function	MSE
Total no. of iterations	100

The below section describes the Genetic Algorithm parameters modeled for the work.

Genetic algorithm – parameters

Genetic Algorithms are efficient search methods based on the principles of natural selection and population genetics [29]. GA can search the solution space to find an optimal or near optimal solution by using evaluation and genetic operator functions to maintain the useful schemata of chromosomes in the population, in which chromosomes are evaluated using a fitness function to determine their fitness. According to the principle of survival of the fittest, Chromosome with a higher fitness will have higher probability of survival in each generation and thereby offspring with higher probability are generated. The genetic algorithm parameters modeled for identifying significant gene are given in **Table– 2**.

Table: 2. Parameters of GA

Chromosome - Size		33
Generations		100
Populations –Size		100
Selection		Random Selection
Fitness		MSE
GA operators	Crossover	Single point crossover
	Mutation	Substitution
	Mutation rate	0.8

The GA cycle stops the process if the genetic algorithm reaches local optimum or after the maximum number of solutions that already defined. The pseudo code for proposed work is given below:

Proposed pseudo code : Hybridized PSO and GA

```
Pop // input , population for PSO
// gene is a particle
Repeat
For each gene
   Initialize gene
End
do
   For each gene
      Calculate fitness value
      If the fitness value is better than the best fitness value (pBest) in history
         Update pBest = current best value
   End
   Choose the gene with the best fitness value of all the particles as the gBest
   For each gene
      Calculate gene velocity
Equation7  // equation to calculate velocity
      Update gene position
Equation 8  //equation to calculate position
   End
If (maximum iterations ==TRUE)
```

THE IIOAB JOURNAL
ISSN: 0976-3104

```
Population p=best 1000 gene from PSO
pop=(pop-best 1000 gene)
if(size(pop) < sufficient)
 break(PSO)
end if
//Genetic algorithm
Initialize population p
do
For i in 1:100   \\ Crossover
Randomly select two parents X_A and X_B from p
Generate X_C and X_D by one-point crossover to X_A and X_B
End for
For j in 1:size(p1)   \\ Fitness
Fitness value = sum_{j=1}^{m} (x_j - x̄)^2

End for

Select 100 chromosomes with minimum MSE value
 //Best chromosome selection
while there is no recurrence of chromosome in generation

If recurrence of chromosome = TRUE
For (k in 1:100)  //Mutation
Select duplicate chromosome Xk from p1
Mutate a gene of Xk
        If Xk is unfeasible
                Replace with best gene
        End if
End for
End if
Else
Stop // reaches local optimum or maximum no. of generation
Update p  // Update
p = best 100 fittest chromosomes
Returning best solution p //Return
PSO(pop)
```

Phase II: Verification and validation

The proposed work is validated biologically using PANTHER tool. PANTHER is an online biological tool used to analyze the biological significance of proteins and gene. PANTHER produces the output of list of gene which are involved in Gene Ontology functionalities such as Molecular function, biological process, cellular component and disease pathway. In this proposed work GO functionalities and pathway are analyzed to find out biological relevance and biological significance of 12 extracted gene from phase II. The experimental results of these approaches are discussed in result and discussion section. The results are also verified and compared with literature result.

Significant gene

The gene from the microarray dataset is identified as significant based on the biological validation using Gene Ontology. The gene is available in GO only when the gene has functionalities in Biological Process, Molecular Function and Cellular Component. The gene identified using the proposed approaches are termed as significant when the gene is available under GO taking part in all the three functionalities.

RESULTS AND DISCUSSION

Experimental results

The experimental results of proposed work are discussed in this section. Lymphoma microarray dataset is used for the proposed work. The dataset is downloaded from [http:/llmpp.nih.gov/lymphoma/data.shtml]. The microarray dataset consists of 4026 gene and 96 samples. In this section the experimental analysis of PSO-GA approaches are presented in detail. The experiments were conducted on the data set by executing the algorithms PSO and GA individually. The experimental analysis are presented below.

THE IIOAB JOURNAL
ISSN: 0976-3104

Normal PSO

The PSO algorithm is applied on the lymphoma dataset for identification of significant gene for biomarker discovery. On experimental run 33 gene are identified from 4026 gene using PSO approach in which 12 gene are found as significant based on the biological validation taking part in all the functionalities in this approach. The **Figure–1** presents the best significant gene based on fitness value using PSO approach. The fitness plot of the normal PSO approach for the dataset is given in **Figure–2**.

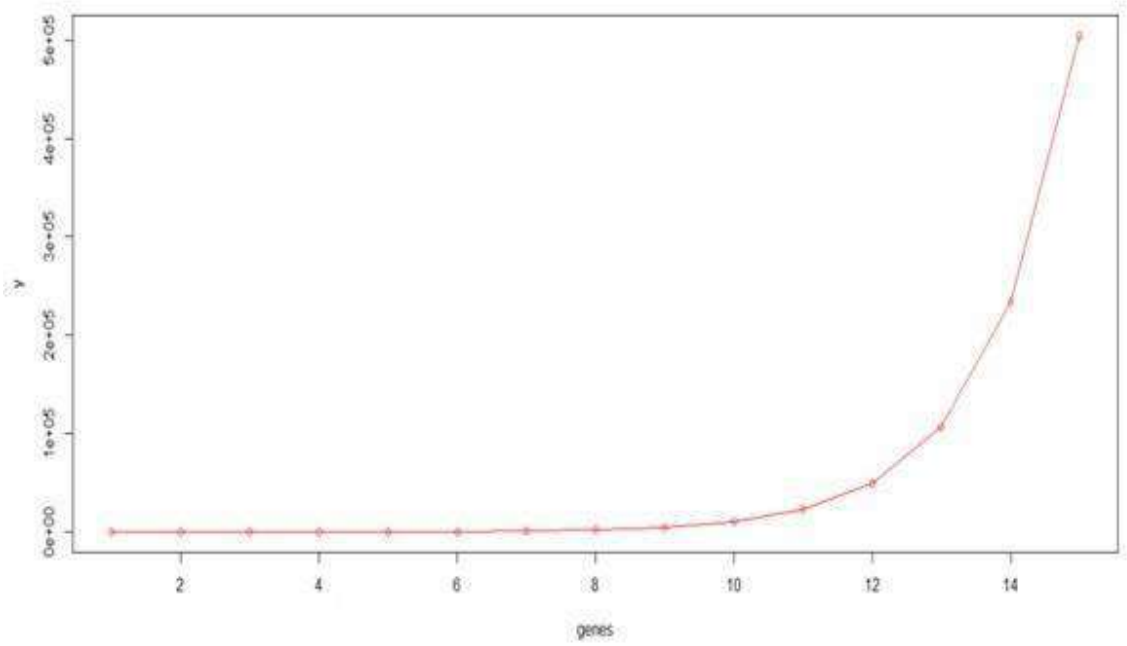


Fig: 2. significant gene from normal PSO

Genetic algorithm

The experimental result of Genetic Algorithm is presented below. The chromosome for GA approach is modeled with 33 gene of microarray data. The GA is run with the with a population size of 100 for 50 and 100 generations. 424 gene are identified using GA is given in **Table–3** with fitness score. **Figure–3** represents the significant gene for 50 and 100 generations. From the graph it is observed that the no of gene get optimum after 15 generations. 12 gene were found to be significant in GA approach based on the biological validation of GO.

Table: 3. Experimental Result - GA

Total No of Genes	4026
Significant Gene	424
Best Chromosome Fitness For first 10 generation	1230.21, 1046.75, 1009.72, 1140.30, 975.18, 975.72, 926.19, 911.82, 898.91, 911.82
Best Fit chromosome	688.31

THE IIOAB JOURNAL
ISSN: 0976-3104

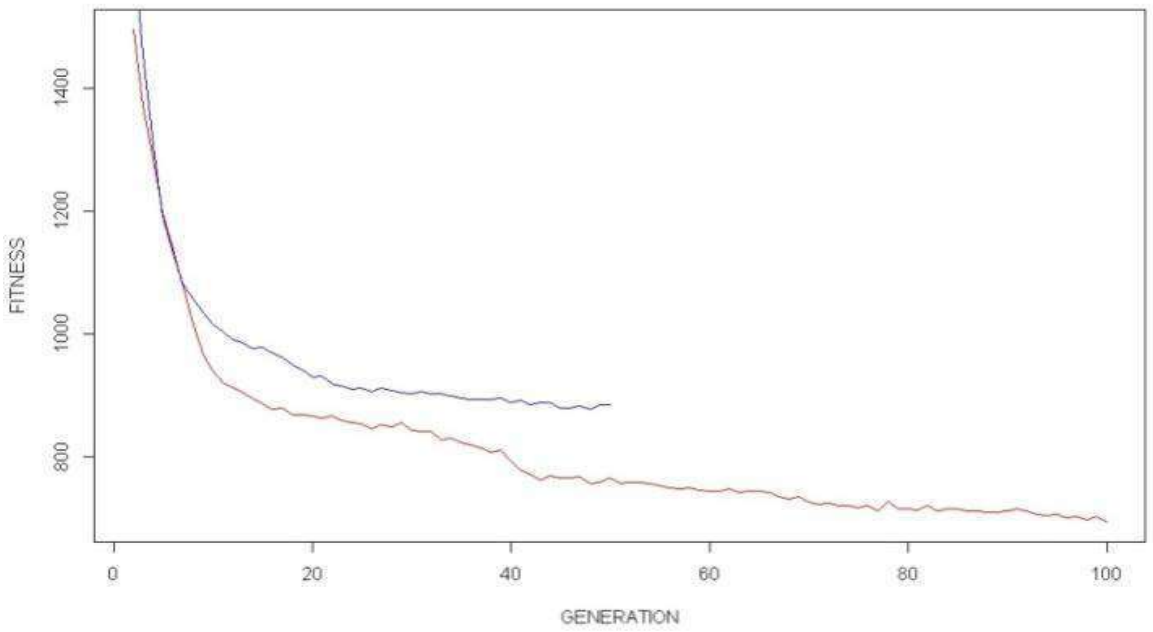


Fig: 3. Significant gene of 50, 100 generations

Hybrid PSO- GA

The experimental result of the proposed approach hybrid PSO- GA with comparison of Normal PSO and GA approach are discussed below. To identify the biomarker the proposed approach PSO-GA identifies the significant gene with a two level optimization process filtering the best gene in every population of the iterations.

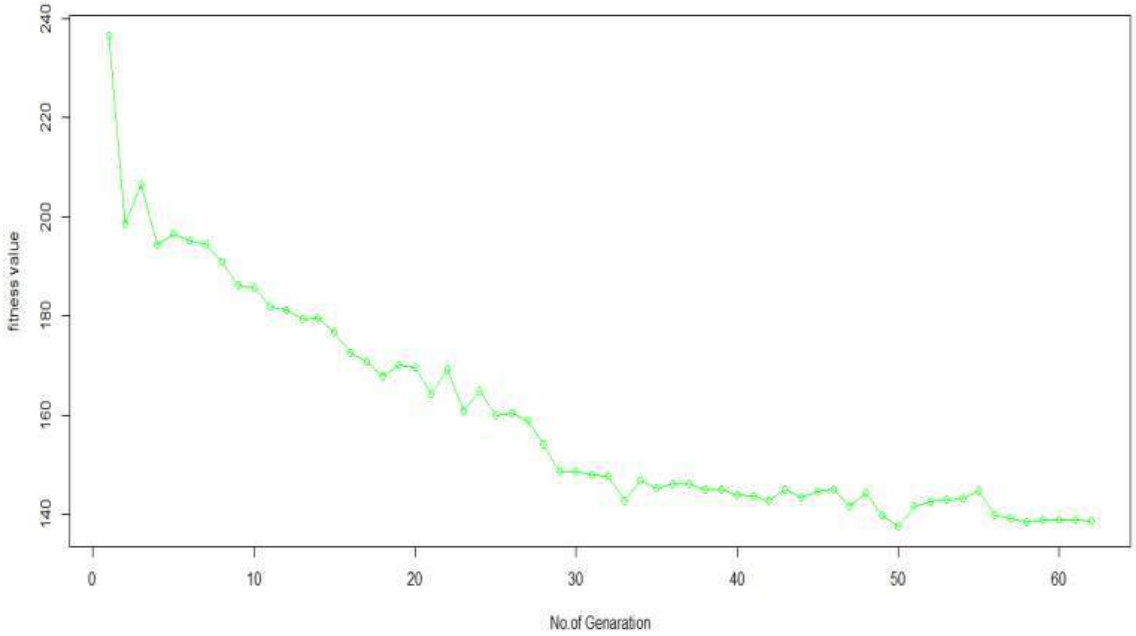


Fig: 4. Best Chromosomes from Hybrid PSO - GA

The **Figure–4** represents the value of 62 best chromosomes derived from hybrid PSO - GA approach. The comparison of gene for the proposed approach PSO-GA with PSO and GA is given in **Table–3** and **Figure–5**. On analysis of gene it is found that 12 common gene found in all the three approaches is termed as highly significant gene. 12 Gene identified from the dataset are:

THE IIOAB JOURNAL
ISSN: 0976-3104

GENE722X, GENE1634X, GENE699X, GENE3836X, GENE2485X, GENE707X, GENE1432X, GENE1806X, GENE185X, GENE2029X and GENE2551X.

Table: 4. Gene counts of PSO, GA & PSO-GA

	No. of iteration	No. of significant gene
Normal PSO	100	33
GA	100	424
Hybrid PSO - GA	100	62

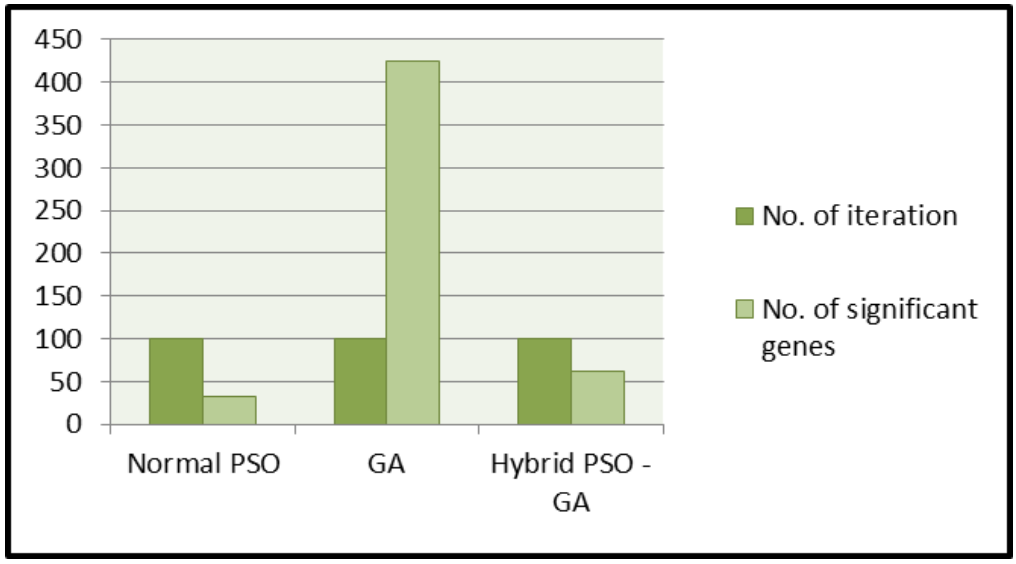


Fig: 5 Comparison of PSO approaches

Biological validation

The extracted gene using PSO-GA approaches are verified using PANTHER tool for validating biological significance. The biological validations of the gene are done using Gene Ontology (GO) functionalities and pathway.

GO functionalities

The identified gene are verified in Gene Ontology functionalities Biological Process, Molecular Function and Cellular Component. A biological process is a process of a living organism. Biological processes are made up of any number of chemical reactions or other events that result in a transformation. Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Cellular component refers to the unique, highly organized substances of which cells, and thus living organisms, are composed. Cells are the structural and functional units of life. The gene ontology found to be for the extracted twelve gene for BP,MF,CC is significant in **Figure– 6(a),6(b),6(c).**

ISSN: 0976-3104

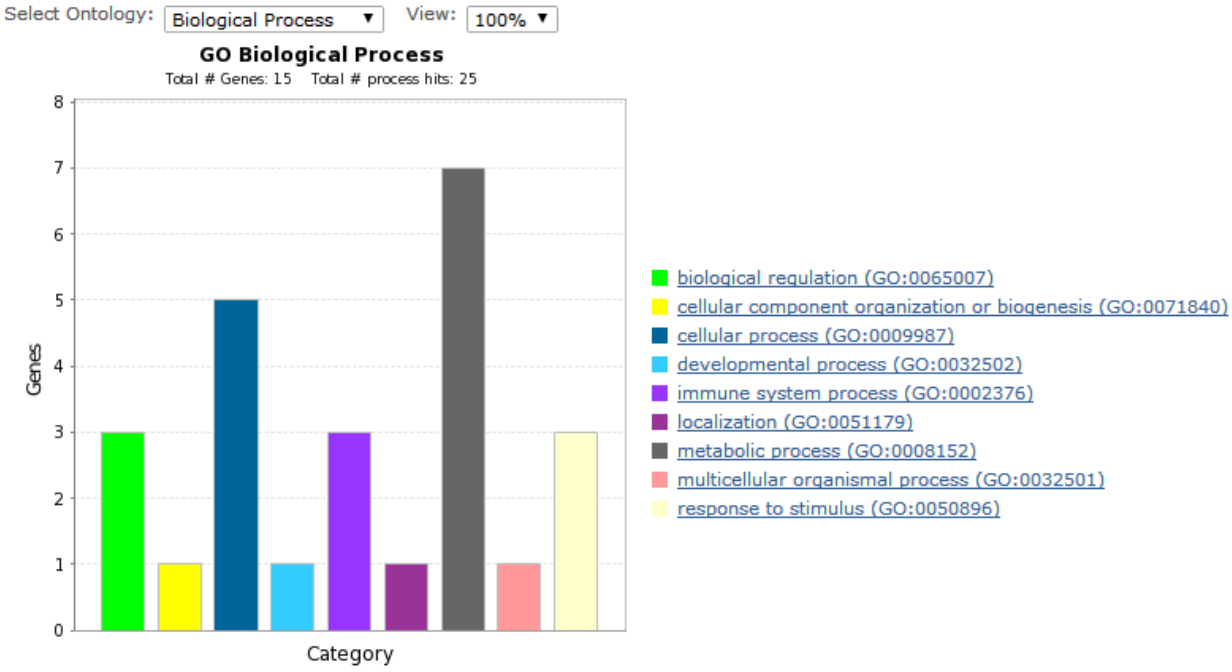


Fig: 6 (a) Biological processes

It can be inferred from **Figure–6(a)** identified 12 gene are involved in 9 biological process like biological regulation, biogeneis, cellular process, development process, immune system process, localisation, metabolic process, multi cellular organism process, response to stimulates and all the 12 gene are active in more than one process

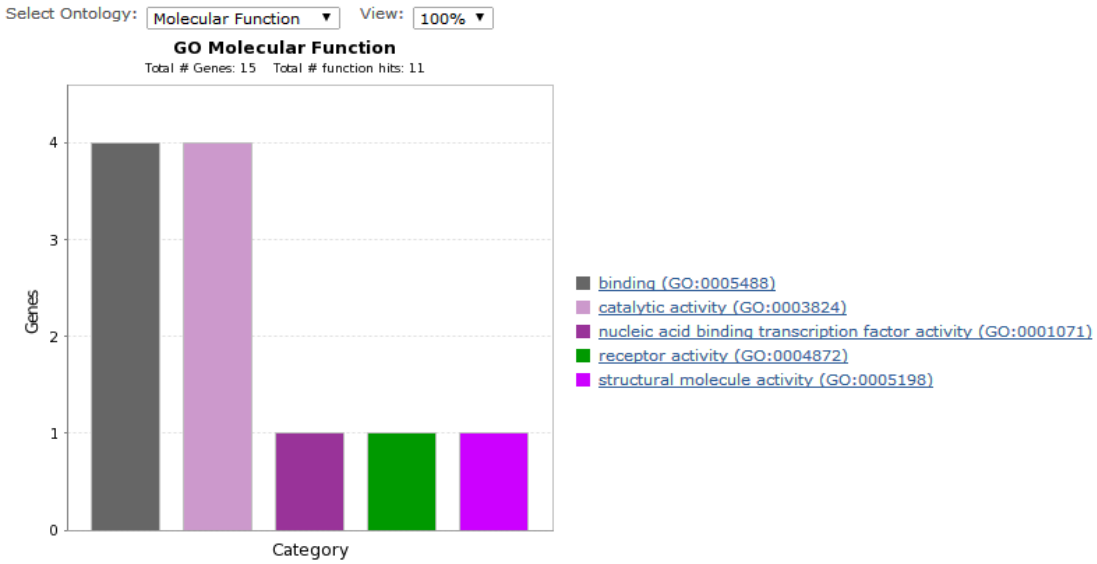


Fig: 6 (b). Molecular functions

ISSN: 0976-3104

From **Figure–6 (b)** it can be inferred that identified 12 gene are involved in 5 molecular function like binding, catalytic activity, nucleic acid binding, respirator activity, structural molecular activity and are active in more than one process. Each colour represent the molecular activity of the gene.

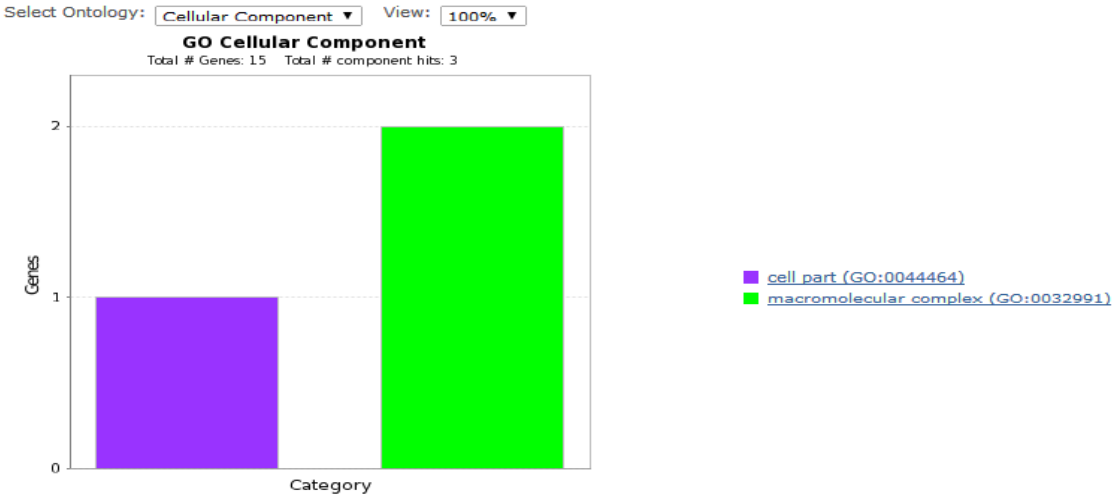


Fig: 6 (c). Cellular components

Figure–6 (c) It can be inferred that identified 12 gene are involved in 2 cellular components like cell part and macromolecular complex. Hence based on the result it is found that the gene are involved in the BP and there are chances to be marked as biomarker. where the five gene GENE1432X, GENE1806X, GENE185X, GENE2029X and GENE2551X identified in this work are also found in literatures [30][31][32].

Pathway validation

The experimental results are also verified with pathway database. The **Figure–7** shows gene involved in the diseases pathways using panther tool for the biological validation of the proposed approach.. The 12 gene identified from PSO-GA approach are found in more than 1 pathway. Out of 12 gene five gene GENE1432X, GENE1806X, GENE185X, GENE2029X and GENE2551X identified in this work is found already in literatures [30][31][32] .
From the experimental analysis the three gene GENE1432X [30], GENE1806X [31], GENE185X [32] are found in the diseases pathway. The three gene are identified in the proposed approach found in the diseases pathway which has to be further verified for biological significance by experts to mark as biomarkers.

In **[Figure–6]**, based on the forecast threshold, the number of data aggregation is larger when compared to the Kalman filter. Due to this data aggregation is reduced redundant transmission and communication consumption power.

CONCLUSION

Biomarker discovery has become important in many aspects of drug discovery and development. Biomarkers predict toxicity and help to save hundreds of millions of dollars with the early termination of costly clinical trials. Many biomarkers play important roles in diagnostic and prognostic applications, where they help to detect or predict diseases. In this work, a hybridized approach of Particle Swarm Optimization and Genetic Algorithm for retrieving highly significant gene is done as a two level optimization process. From the experimental results out of 4026 gene of Lymphoma microarray dataset 12 gene were found to be significant in all the three approaches PSO, GA, PSO-GA. Out of 12 gene 3 gene were found in the diseases pathway and GO functionalities using the biological validation approach of this work. 3 gene when verified with domain experts can be marked as biomarker.

The proposed approach is found to be better for identifying optimal gene using the two level process PSO-GA is verified based on biological validation. The proposed algorithm compared with the other approaches is found to be significant in performance.

THE IIOAB JOURNAL
ISSN: 0976-3104

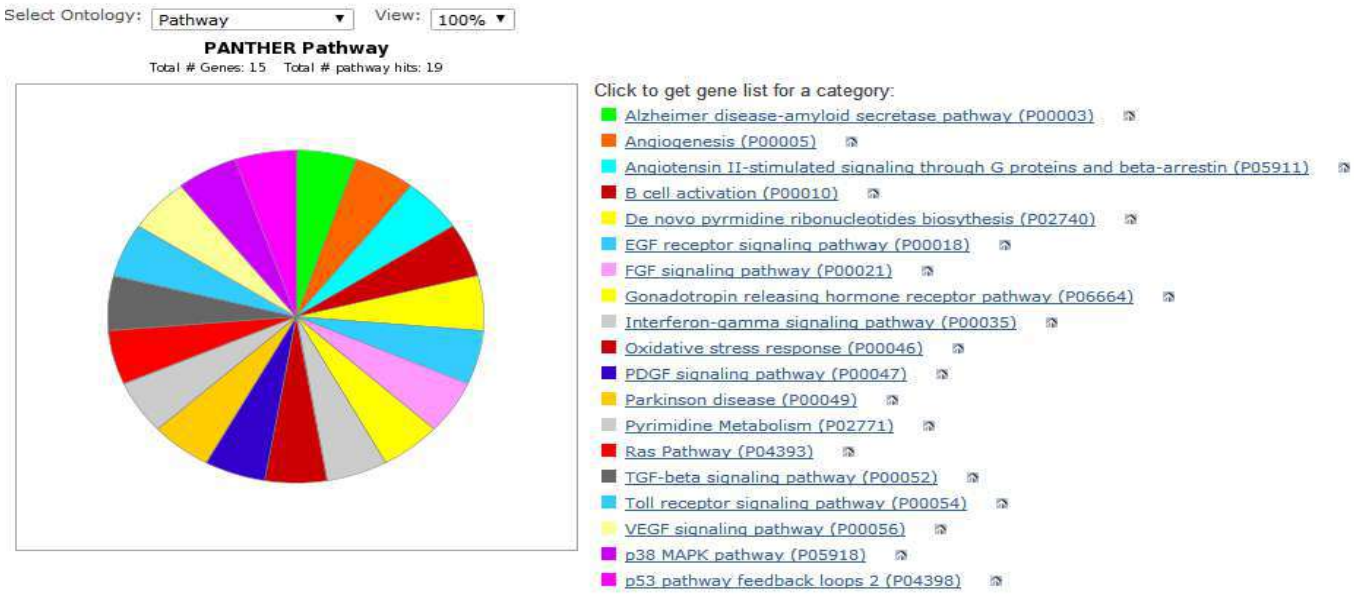


Fig: 7. Pathway of 12 Gene using PSO-GA

LIMITATION AND FUTURE SCOPE

The limitation of the proposed work is, that it is iterated only for little iteration due to the hardware constraints. The work is tested only on one dataset the proposed approach can be verified with other microarray dataset in future. The experimental run can be further extended by implementing using map reduce framework with the existing hardware. The approach will be further tried with other algorithmic approaches in future.

ACKNOWLEDGEMENT
None.

CONFLICT OF INTERESTS
Authors declare no conflict of interest.

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

REFERENCES

[1] Todd Walter [20.11] A multi-variance analysis in the time domain, 413−426.
[2] Ji-Hoon Cho, Dongkwon Lee et.al.[2004] Gene selection and classification from microarray data using kernel machine. *FEBS Letters* 571 :93–98.
[3] Laurent Gautier et,al. 2004] affy—analysis of Affymetrix GeneChip data at the probe level. *Oxford journals,* 20 (3): 307–315.
[4] Nicola Day et.al., [2012] Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. *Journal of Translational Medicine*, 10:217
[5] Edmundo Bonilla Huerta et.al., [2008] Gene Selection for Microarray Data by a LDA-Based Genetic Algorithm. *Springer Berlin Heidelberg*, ISSN 0302-9743, 5265: 250−261 .
[6] Ramachandran S Vasan. [2008] Biomarkers of Cardiovascular Disease : Molecular Basis and Practical Considerations. *American Heart Association* ,ISSN: 0009−7322.
[7] Margaret Sullivan Pepe et.al.,[2001] "Phases of Biomarker Development for Early Detection of Cancer", *Journal of the National Cancer Institute*, 93(14)
[8] Paul Yacci et.al. [2009] Feature Selection of Microarray Data Using Genetic Algorithms and Artificial Neural Networks ,RIT Scholar work.
[9] K Hron et al.[2010] Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis* 54(12):3095–3107.
[10] Matt Blackwell.[2008] Multiple Hypothesis Testing: *The F-test journals*, 2(3): 30–35.
[11] Shinn-Ying Hoa b et.al.[2006] Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Elsevier* 85(3):165−176. Epub.
[12] Guoqiang Yu, et.al.,[2010] Matched Gene Selection and Committee Classifier for Molecular Classification of

Heterogeneous Diseases. *Journal of Machine Learning Research* 11: 2141−2167.

[13] Shital C. Shah, et.al. [2004] "Data mining and genetic algorithm based gene/SNP selection", *Artificial Intelligence in Medicine* 31: 183—196, Elsevier B.V. doi:10.1016/j.artmed.2004.04.002.

[14] Mohd Sazli Saad,[2012] Implementation of PID controller tuning using differential evaluation and genetic algorithms, *ICIC international*, ISSN 1349−4198.

[15] Barnali Sahua et.al.[2012] A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data, International Conference on Modeling Optimization and Computing (ICMOC-2012)

[16] Javed khan et.al, [2001] Classification and diagnostic prediction of cancer using gene expression profiling and artificial neural network. *Nature Medicine* **7**: 673 − 679 doi:10.1038/89044.

[17] Keisuke Kameyama, [2009] Particle swarm optimization- a survey, IEICE TRANS.INF. & SYST., Vol-E92-D,No.7 .

[18] Linyu Yang, et.al., [2001] An Entropy-based Adaptive Genetic Algorithm for Learning Classification Rules . 2 ,*IEEE* published in **:**Evolutionary Computation, Proceedings of the 2001 Congress.

[19] Miguel Rocha and Isabel Rocha,[2011] Data Mining Techniques for DNA Microarray Data. *Nature Medicine* **7**: 673 − 679.

[20] Russell d Wolfinger et.al.,[2001] Assessing Gene Signi. cance from cDNA Microarray Expression Data via Mixed Models. Journal of computational biology ,Vol. 8, Pages 625–637.

[21] C.H. Ooiand Patrick Tan,[2003] Genetic algorithms applied to multi-class prediction for the analysis of gene expression data . *oxfordjournals*, 19(1)

[22] JY Yeh,[2008] Applying Data Mining Techniques for Cancer Classification on Gene Expression Data. *Cybernetics & Systems*, 39(6): 583−602. Doi: 10.1080/01969720802188292.

[23] Yongliang Yang et.al.[2008] Integrative Genomic Data Mining for Discovery of Potential Blood-Borne Biomarkers for Early Diagnosis of Cancer, 3(11):3661

[24] Sandrine dudoit, et.al, [2002] Statistical methods for identifying differentially expressed gene in replicated cDNA microarray experiments, *Statistica Sinica* 12: 111−139.

[25] Luciano S´anchez, et.al.,A Multiobjective,[2006] Genetic Fuzzy System with Imprecise Probability Fitness for Vage Data. International Symposium on Evolving Fuzzy Systems, 0-7803-9719-3/06/$20.00 ©2006 IEEE.

[26] K Dheenathayalan, J Ramsingh,V Bhuvaneswari.[2014] Identifying significant gene from DNA microarray using Genetic Algorithm. published in IEEE Xplore – 2014 International Conference on Intelligent Computing Applications - 978-1-4799-3966-4/14 $31.00 © 2014 IEEE DOI 10.1109/ICICA.20120

[27] Vathany Kulasingam, Eleftherios P Diamandis et.al, [2008] Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology*, 5(10)

[28] Alex Kotlarchyk, [2011] Identification of microRNA biomarkers for cancer by Combining multiple feature selection techniques. *Journal of Computational Methods in Sciences and Engineering*, 11(5−6):283−298.

[29] Barbarino Julia M, Staatz Christine E, et.al. [2008] PharmGKB summary: cyclosporine and tacrolimus pathways. *Pharmacogenetics and genomics* .

[30] Maitland Michael L, Lou Xing Jian, et.al..[2010] Vascular endothelial growth factor pathway. *Pharmacogenetics and genomics*

[31] M. Whirl-Carrillo, E.M. McDonagh, et.al. [2012] Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics.*

[32] Anuska M Glas et.al., [2005] Gene expression profiling in follicular lymphoma to access clinical aggressiveness and to guide the choice of treatment. *Blood Journal*;105(1):301−307.

[33] Ahmad m. Sarhan, [2009] Cancer Classification Based on Microarray Gene Expression Data Using DCT and ANN. JATIT.

[34] Christophe Ambroise, et.al,[2002] Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99(10).

ISSN: 0976-3104
SPECIAL ISSUE (ASCB)
Anantharaj and Thangavelu.

THE IIOAB JOURNAL

REVIEW ARTICLE OPEN ACCESS

A PREDICTIVE ANALYTICAL APPROACH TOWARDS IMPROVING THE CROP GROWTH YIELD USING FUZZY COGNITIVE MAPS – CROYAN

Gunasundari Anantharaj[1] and Arunkumar Thangavelu[2]
[1]*Department of Computer Science, Nerhu Arts and Science College, Coimbatore, TN, INDIA*
[2]*School of Computing Sciences & Engineering, VIT University, Vellore, TN, INDIA*

ABSTRACT

Improving the crop yield is always a major challenge for farming community as well for agricultural scientists. Though various computational approaches have been followed traditionally in practice, still a persistent decision making method to improve crop yield is not predicted and lies as a misnomer. Multiple rethegorical factors / parameters form the basis of improved crop yield. This paper proposes Fuzzy Cognitive Map approach to help in decisive making and suggesting an optimized solution of improving crop yield as well as compared with existing approaches such as Genetic Algorithm (GA) and Artificial Neural Network (ANN). The research work considers two major crops namely Sugarcane and Red Chilli for analysis. The results show that FCM provides a higher accuracy in supporting crop growth yield with higher similarity ratio of expected growth and actual growth.
.

Received on: 18th-March-2015
Revised on: 20th-May-2015
Accepted on: 26th-June-2015
Published on: 25th-Aug-2015

KEY WORDS

CROYAN, Fuzzy Cognitive Map, Crop Growth, Crop Yield, Yield Accuracy

***Corresponding author: Email:** gunaanantharaj@gmail.com, **Tel:** +40-9001010010; **Fax:** +40-9001010012

INTRODUCTION

Wireless Consistent monitor and control over crop growth [11] is the major variant of precision agriculture with support of greenhouses [22], which are considered as biophysical systems with inputs, outputs and control process mechanisms which adds to cost and complexity in maintenance. Hence designing a natural crop growth and yield system with support over understanding the growth variables and formulating the decision making is the primary objective of this work.

Decision making on improved crop yield helps to:

[a] Understand and identify the supportive metrics related to improving crop growth and yield.
[b] Support with deterministic decision making with consistent monitoring and control over crop yield at each growth phase
[c] Support crop growth by understanding and analyzing the environmental and demographic control parameters to estimate crop yield.

Understanding crop yield management [13] and its behavior in agricultural domain carries much importance since it relationally influences climate, region aspects or farming demographic factors. Computational approaches can be considered which support in predictive analysis of crop growth and yield. Approaches such as genetic algorithms [4][6], Artificial Neural Networks (ANN) [7], Bayesian Approach [12], fuzzy models [20] are well applicable, though such models can be adopted in complex situations, it does not achieve an optimal solution for highly dynamic and multi-variable solutions. Hence the need for an efficient knowledge-based system approach utilizing the Fuzzy Cognitive Maps (FCMs) [8] [10] approach for characterizing crop yield of behavior is discussed in this paper.

FCM is a modeling approach based on exploiting knowledge and experience. The novelty of the method is based on the use of the soft computing method of fuzzy cognitive maps to handle experts' knowledge and on the unsupervised learning algorithm for FCMs to assess measurement data and update initial knowledge. The advent

THE IIOAB JOURNAL
ISSN: 0976-3104

of precision farming generates data which, because of their type and complexity, are not efficiently analyzed by traditional methods. The FCM technique has been proved for its accuracy and efficiency from the literature survey as well flexible to handle experts' knowledge and through the appropriate learning algorithms can update the initial knowledge.

Two major crops are being considered for analysis such as Sugarcane (botanical name : Saccharum Officinarum), Chilli (botanical name : Capsicum annuum L.; Capsicum frutescene L.). These crops are considered for analysis and survey since all the three crops are well grown in India and commonly in most of the tropical regions.

The FCM model developed consists of nodes linked by directed edges, where the nodes represent the main factors in crop growth production such as soil texture, water composition as well organic matter of soil such as pH, K, P, Mg, N, Ca, Na. The directed edges show the cause-effect (weighted) relationships between the soil properties and crop grown field. The proposed method was evaluated for 300 cases measured for subsequent years such as 2007, 2008, 2009, 2010, and 2011 based upon data obtained from International Council of Agricultural Research (ICAR) and related agricultural datasets.

The proposed FCM model enhanced by the unsupervised nonlinear Bee Hive learning algorithm[23], can support a success rate on an average of greater than 75%, for years 2009, 2010 and 2011 dataset referred as estimating/predicting the yield between fuzzy categories (“worst”, "low", “average”, "high", “best”). The main advantage of this approach is the sufficient interpretability and transparency of the proposed FCM model, which make it a convenient consulting tool in describing crop yield behavior for Chilli and Sugarcane crops. The work describes crop yield analysis being obtained from an uncontrolled farm field and not from a precision or data not obtained from temperature and hygrometric conditions in greenhouses. Hence this work carries much importance in an uncontrolled, stochastic farm field procedure.

CROYAN method works on an association based mining algorithm which is binded to the moderated patterns obtained from FCM. Fuzzy Cognitive approach observes the variable metrics and factors contributing the improved yield of crop while predictive pattern mining model determines the accuracy of selecting the crop based on variable metrics. The paper is organized as follows, Section-2 focuses on detailed survey and analysis on mechanisms to improve crop yield and need for such methods, while Section-3 elaborates on FCM and its methods, dataset and algorithm. Section-4 discusses on CROYAN with its implementation and Section-5 summarizes the work with future works to be carried out.

SURVEY AND ANALYSIS

This section presents a detailed survey and analysis of existing works which support decisive making approaches [10, 15] for crop yield are discussed. Applying soft computational and statistical approaches for crop yield analysis is not much considered under major research challenges [20]. In order to derive an accurate crop yield, predictive mining approach on fuzzy cognitive method could prove elusive. Crop yield attributes to some common basic factors which are related to crop growth parameters and disorders parameters such as high rainfall, irrigation procedures [4], moderate sunlight intensity, rocky soil type which may not attribute to growth of crops in general. Such change in unexpected frequency may affect severity of crop yield as unexpected by farming community. Though this paper supports on decisive making and mechanisms to achieve crop yield [1,4], the detailed and in-depth analysis of the factors attributing to destruction of crop growth [2] is primarily not within the scope of this work.

Asefa Taa et al [2] discussed multiple factors which attribute to consistent crop growth needs to be automated which impose high accuracy and complexity [2]. The phenomenon of crop growth cannot be completed automated or sensed upon due to the need for supporting precise and periodical evaluation of biotic status of crops and their natural growth. This paper discusses on the underlying factors behind design and implementation of a crop yield decision making system which adopts growth and yield sensitive metrics on control and monitoring growth process of crops. This approach is compared and analyzed over similar crop growth trends which are achieved over similar climatic and regional factors phenomena.

NEED FOR FUZZY COGNITIVE MAPS

As artificial intelligence and heuristic approaches are not often sufficient to obtain high quality prediction, the proposal can be extended to supplement using knowledge base and adaptive soft computational approaches. Predicting crop yields at an early stage in the growing season can be of great importance. The correlation of the corn yield [7, 16] helped to permit the early yield prediction and the appropriate management of the farm land.

THE IIOAB JOURNAL
ISSN: 0976-3104

Crop growth models were developed in agriculture by using mean values of input and outputs [17]. Several crop growth models have been developed by Xirogiannis [19]. Papageorgiou [13] presents the soft computing technique of Fuzzy Cognitive Maps (FCM) to connect yield defining parameters with yield in cotton crop production in India as the basis for a decision support system for precision agriculture application. FCM was chosen because of the nature of the application as it is a complex process and FCMs have been proved suitable for this kind of problem. Due to consistent change in temperature and hygrometric conditions, rapid decisions are highly essential to predict and control yield of crop or manage diseases to avoid dissemination and permanent infestation of crop growth [4] [12] hence soft computational approaches would suffice.

METHODS

The methods and procedures being followed can be summarized.

(a) Data Set

The dataset collected over 300 record sets consisting of weather, soil, water, regional and crop growth data over the year 2007, 2009, 2010, 2011 catering to the region of Dharwad in Karnataka and Cuddalore in Tamilnadu for Sugarcane crop and Guntur in Andhra Pradesh and Madurai in Tamilnadu for Redchilli crop. The results are compared with bench marking genetic algorithm and ANN systems

(b) Metrics adopted

Few of the major metrics adopted in this research work are (i) Average yield achieved over per year (tons/hga), (ii)Average crop production per year each ha, (iii) Mean Crop Yield

(c) Soil - Water

Total Soil-Water Potential (SWP) can be used to determine the capacity of water available for crop transpiration from each soil layer. The capacity of the soil to hold enough water supports the growth of crop. Hence type of soil and water has a potential role in crop growth as well yield

(d) Role of NDVI

NDVI (Normalized Difference Vegetation Index) along with variable crop growth metrics suggests a useful way for crop yield assessment models whose approaches vary from simple integration to more complicated transformation. NDVI [9] has proved to improve vegetation greenness, which indicates the level of healthiness in the crop growth

(e) Selection of crop and growth region

Major Chilli growing states in India are Andhra Pradesh, Tamil Nadu, Maharashstra, and Karnataka which together constitute nearly 75 per cent of the total area. Hence dataset has been collected for Guntur (16.3008° N, 80.4428° E), and near by regions such as Khamman and Cuddappa in State of Andhra Pradesh. In Tamilnadu, Sugarcane dataset in Karur district (10.9338334 N, 78.0883645E) and near by Erode are collected for analysis. In both the regions Chilli variety K-1 and CO-1 are used for analysis while CO-265 variety for Sugarcane crop.

FUZZY COGNITIVE MAP

Fuzzy cognitive maps (FCM) [3] [8] along with SVM (Support Vector machines) [5] helps in understanding and modeling crop yield and representing crop expert knowledge, since fuzzy cognitive theory [6] supports on theory of fuzzy logic and cognitive map, which are capable of dealing with uncertainty issues.

Modeling approach using Fuzzy Cognitive Maps

Fuzzy Cognitive Map (FCM) methodology [14, 17] is a symbolic representation for the description and modeling of complex system. Fuzzy Cognitive Maps describe different aspects in the behavior of a complex system in terms of concepts. FCMs illustrate the whole system by a graph showing the cause and effect along concepts, and are a simple way to describe the system's model and behavior in a symbolic manner, exploiting the accumulated knowledge of the system. A Fuzzy Cognitive Map integrates the accumulated experience and knowledge on the operation of the system.

Fuzzy Cognitive Maps (FCM), Tsadiras [18], defines it as a qualitative alternative approach to formulate dynamic systems. This approach can be considered to represent formal method to represent predictions and taking decisions. This work investigates the crop yield and variability crop yield prediction in chilli and sugarcane crops.

THE IIOAB JOURNAL
ISSN: 0976-3104

Crop management in these crops are highly complex with interacting parameters such as water, climate, soil factors which play major role in improving the yield of crop.

The FCM model possess nodes which show the variable factors affecting crop yield using a directed graph G(V,E). The cause effect relationship between the factors and crop yield can be depicted and analyzed. As shown in Fig-1 each components or elements contributing to growth of crop can be considered as Node, 'Na', through 'Ne' while all linkages which may be related between multiple nodes can be termed as Edges as 'E1' through 'E6', generally as Ei.

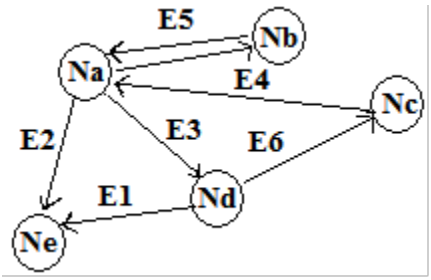


Fig:1. FCM representation using graph theory G (V,E)

FCM assigns a weight as 'Ei' for each edge, and each node 'Ni' binds the value to its weight. The causal relationship between two concepts Nj and Ni in fuzzy is explained with weights Ei, which considers the value in the range –1 to 1. Three possible types of causal relationships between concepts are followed:

(a) Ei > 0 which indicates positive causality between concepts Nj and Ni, which indicates that an increase or decrease in the value of Nj leads to an increase (decrease) in the value of Ni.

(b) Ei < 0 which indicates negative causality between concepts Nj and Ni, which indicates that an increase (decrease) in the value of Nj leads to a decrease (increase) in the value of Ni.

(c) Ei = 0 which indicates no relationship between Nj and Ni.

The FCM process can be inferred from the following mathematical formulation in equation 1.

$$g_i^{k+1} = f\left(g_i^{(k)} + \sum_{\substack{j \neq i \\ j=1}}^{N} g_j^{(k)} \cdot E_i \right) \longrightarrow \quad 1$$

Here, gi denotes the value of Ni node, while simulation is carried out over 'k', Ei is the linkage value between two nodes 'Ni' and 'f' is a sigmoid threshold function, which is 1/ (1+e-□x). □ lies between a value [0,1]. The fuzzy metric for crop yield lies between the ranges {low influence, moderate influence, acceptable influence, strong influence, best influence} over crop growth and production.

The fuzzy IF-THEN rules that experts use to describe the relationship among concepts assume the following form, where A and Bare linguistic variables:
IF value of concept INFLUENCE Ni is "Low", AND
value of concept INFLUENCE Nj is "Average" THEN linguistic weight INFLUENCE Ei is "Acceptable"

where Low, Average, Acceptable are linguistic INFLUENCE variables taking values in the range [0, 1].

Each interconnection described by a fuzzy linguistic variable from the determined set, associates the relationship between the two concepts and establishes the grade of causality between the two concepts.

IIOAB JOURNAL
ISSN: 0976-3104

Mapping the component values between the edges and variable nodes

The dataset collected over 300 cases consists of weather, soil, water and crop data during the years 2007, 2008, 2009, 2010 and 2011 for the region of Guntur in Andhra Pradesh. The Sugarcane data obtained for 250 record sets are obtained from Karur in Tamilnadu. The data is validated with base line data and maintained as repository. The results are compared with bench marking fuzzy and Artificial Neural Networks (ANN) systems.

RESULTS AND DISCUSSION

CROYAN: Crop yield analysis using FCM approach

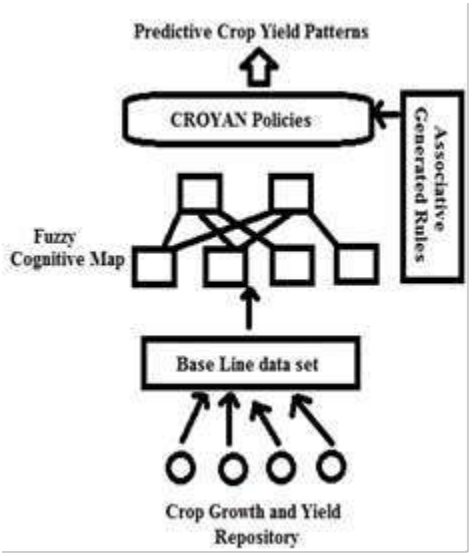


Fig: 2. CROYAN architecture and functionality

Figure−2 explains CROYAN architecture which extracts the features of crop growth from data set and determines the crop yield efficiently. The baseline dataset is validated based on historical datasets and verified. FCM is being generated on crop growth parameters using metrics **[Table−1]**. Polygon clusters are generated for variable growth parameters.

The coefficients of empirical equation have been obtained using a analytical breakpoint (m) based on optimization of a mean of three years of Sugarcane and Chilli crop yield. **[Table −2]** shows the breakpoint achieved when the actual maximum yield achieved is equivalent to the expected yield. CROYAN methodology uses polygonal analysis to mine related datasets, which formulates the following steps:

1. Collect/Generate polygonal clusters for multiple related datasets
2. Meta cluster polygonal clusters
3. Extract interesting patterns / create summaries from polygonal clusters

CROYAN methodology adopts a multi clustered polygon based architecture, which defines variable multiple clusters required to identify the fitness of crop yield. Polygonal clusters [14] defined using Hausdroff metric suggests each clusters based on crop type, region associated with crop growth and production as shown in **Figure−3**. Each polygonal clusters originating from different datasets typically overlap, which provide an option to restrict cluster overlap in final clustering. Each clusters usually assumes that inter dependent polygons do not overlap and most uses the Hausdroff distance [11] to assess polygon similarity for crop yield. **Figure-3** shows the set of polygonal clusters which depends on crop sow values and its relational yield values. This measure is vital in its definition due to its need for identifying the crop growth measure which is required for identifying the crop yield.

THE IIOAB JOURNAL
ISSN: 0976-3104

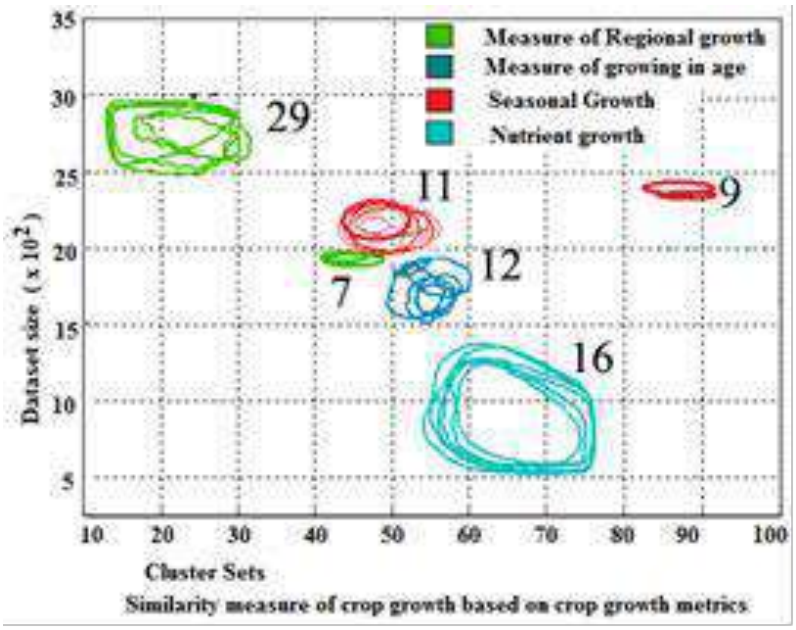


Fig: 3. Cluster sets generated as polygonal clusters

CROYAN inputs and definition

A clustering set S={X1, ..., Xn} —at most one metric is selected from each meta cluster Xi, (i=1,...n).

1. The crop growth provides individual cluster crop yield function FitU whose values are [0, ∞).
2. A fitness threshold value defined with θU
   U being set of lower unfit clusters are occupy not included in the final clustering.
3. A cluster distance threshold θcroyan_dist which determines the cluster overlap/coverage can be tolerated.
4. A cluster distance function Z is defined between clusters of croyan_dist ,

For $Xi \subset S$, where i = 1…n

Begin

Find $Z \subseteq X1 \cup ... \cup Xn$ that maximizes: $q(Z) > FitU(Xi)$

subject to:

a) $\forall X \in Z \; \forall X' \in Z \; (X \neq X' \Rightarrow croyan\_dist(X,X') > \theta croyan\_dist)$

b) $\forall X \in Z \; (FitU(X) > \theta U)$

c) $\forall X \in Z \; \forall X' \in Z \; ((X' \in Xi \in S \wedge X \neq X') \Rightarrow i \neq n)$

End

The goal is to maximize the number of fitness measures of crop yield among clusters which have been selected from polygon clusters. The constraint 'a' prevents any two clusters that are too close to each other but are both included in the final clustering. Constraint 'c' makes sure that at most one cluster from each cluster being selected. Constraint 'b' determines the policy for fitness measure where the fitness can satisfy step (2). Here, Z is the cluster distance function defined based on CROYAN distance threshold level, FitU is the fitness measure obtained to understand the optimal crop yield.

Model

The FCM model possess nodes which show the variable factors affecting crop yield using a directed graph G(V,E) shown in **Figure−1**. The cause effect relationship between the factors controlling the crop yield and aspects which support in crop yield can be analyzed.

THE IIOAB JOURNAL
ISSN: 0976-3104

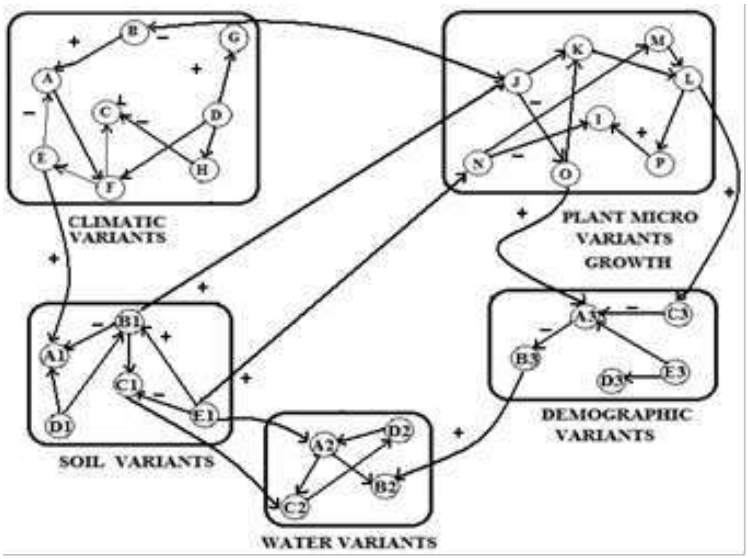


Fig: 4. Crop Yield analysis based on FCM variant metrics

Figure–4 shows the analysis of metrics required for crop growth, over variable FCM nodal components such as climatic variants, growth micro-variants, regional demographic variants, water and soil variants. Each variant possess multiple crop growth dependent elements which plays a vital role in crop production and detecting the yield.

Analysis and predictions

Early detection of crop yield growth symptoms or initial analysis of NDVI parameters related to the key-point in the crop yield index (CYI) Support in gathering the climatic data through meteorological stations, data on crop growth rendered by agricultural stations were the primary means of data collection procedure. This work adopts the following crop growth metrics such as temperature condition index (TCI) for mapping and monitoring of crop growth metric which also indicates the assessment of vegetation health and productivity. Variable parameters such as TCI, soil moisture, surface temperature and average rainfall as shown in Fig-5 are the valuable sources of information for the estimation and prediction of crop growth conditions.

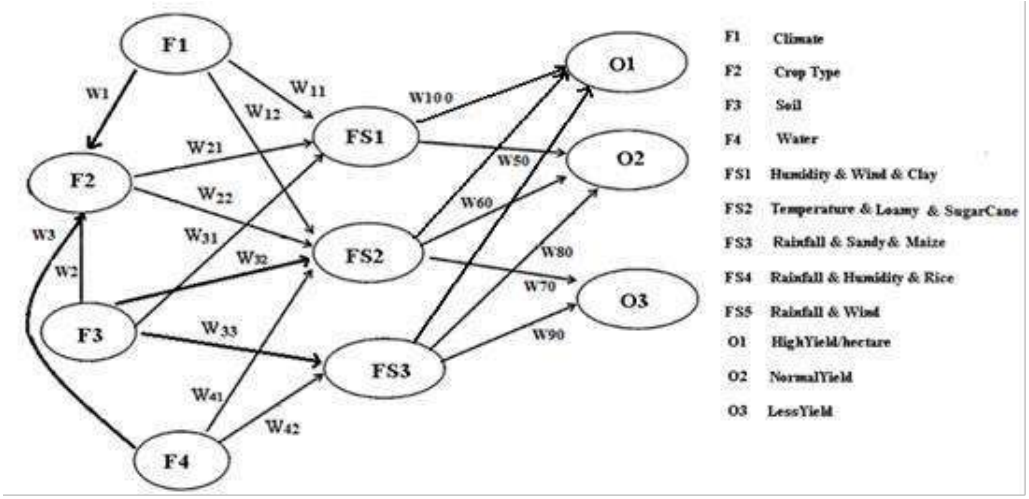


Fig: 5. FCM mapping of Chilli crop growth parameters and its corresponding outcome 'O'

THE IIOAB JOURNAL
ISSN: 0976-3104

Figure–5 shows components 'Fi' as metrics responsible for Chilli crop growth, which are interdependent on one other and hence map to 'FSi' components as primary variables that plays vital role in crop growth. 'Oi' depicts the final outcome of crop growth outcome, which may be "High in Yield" or "Normal Yield" or "Less Yield". Similar sugarcane crop growth can be obtained using FCM approach. 'Wi' indicates the weight assigned for components on priority, which gets updated based on crop yield based metrics. For any change in weight, the component required for yield also gets varied.

The actual crop yield achieved over Chilli and Sugarcane compared with higher predicted yield for crops were compared as shown in **[Table–2]**. The Residual crop yield becomes equal to expected change in crop yield, based on condition of growth issues which primarily depends on breakpoint applied. The model empirical equation for Sugarcane and Chilli crops, thus obtained with coefficients **[Table– 2]** is given as:

Table: 1. FCM nodal components adopted for CROYAN

A	Precipitation Value/ day
B	Maximum Cloud Cover/ day
C	Radiation Sunshine/ day
D	Vapour Pressure/ day
E	Wind Pressure/ day
F	Snow Depth/ day
G	Average Humidity/ day
H	Average Temperature/ day
I	Monthly Average Rainfall
J	Soil Ph concentration
K	Soil nutrient value
L	Water Ph value
M	Air Pollution Level
N	Climatic Metrics
O	Nutrient Manure metric
P	Water nutrient metric
A1	Soil Type
B1	Soil composition
C1	Soil Manure
D1	Percentage of Anti-oxidants
E1	Soil Humidity
A2	Water type
B2	Water Composition
C2	Oxidation Ratio
D2	Water mineral and pollutant ratio
A3	Soil Plough methods
B3	Crop watering interval
C3	Weed removal methods
D3	Manure laying methods
E3	Pesticides application methods

Table: 2. Achieving crop yield using FCM. Predicted and observed crop yield of Sugarcane for Karur, Tamilnadu and Chilli crop for Guntur, Andhra Pradesh using Fuzzy Cognitive Maps

	Sugar cane crop yield			Chilli crop yield		
Year	Predicted	Actual / Observed	Residual	Predicted	Actual /Observed	Residual
2007-J	120.8	119.5419	1.2581	36.5	34.9126	1.5874
2007–A	87.112	83.7547	3.3573	32.0	31.2125	1.7875
2008-J	126.13	120.6428	5.4872	31.5	32.8647	-1.3647
2008-A	139.130	138.4765	0.6535	38.0	36.8840	1.1160
2009-J	134.118	133.1821	0.9359	41.5	41.9016	-0.4016
2009-A	126 117	125.7217	1.3953	43.5	43.3660	0.1340
2010-J	99.462	98.5832	0.8788	35.0	34.2990	1.2990
2010-A	124.138	123.8340	0.3040	39.0	37.7836	1.2164
2011-J	138.145	137.5609	0.5841	41.9	41.5585	0.4585

THE IIOAB JOURNAL
ISSN: 0976-3104

The SSToolbox 3.61 [21] software supports methodology to store, represent, filter and analyze the acquired field data. All the collected data were interpolated in order to produce a map on a 10m x 10m grid size that corresponds to a reliable field management unit. The years are indicated in two formats as 'J' for January yield, 'M' for may yield, 'A' for august yield and 'O' for October yield. It can be observed from **[Table−2]** that the residual crop yield is 87.42% predicted similar both for sugarcane and Chilli crop.

Performance analysis

The performance of CROYAN is analyzed over crops Chilli and Sugarcane which are well grown in humid temperature climatic regions. Crop growth parameters required for crop growth, to improve the yield and improve production, are considered for analysis.

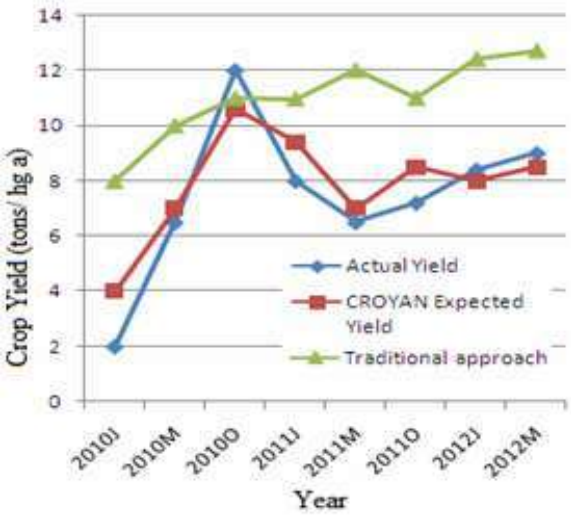


Fig: 6 Sugarcane Crop yield obtained for actual yield, expected yield using CROYAN

Figure-6 discusses on Sugarcane crop yield achieved using CROYAN analytical approach during the years 2010 to 2012 for the months Jan, May and October respectively. It is understood that the yield analysis using actual and expected is averaging over 5.93% to 7.27%. The accuracy of actual yield over the predicted yield confirms the performance of CROYAN for Sugarcane crop.

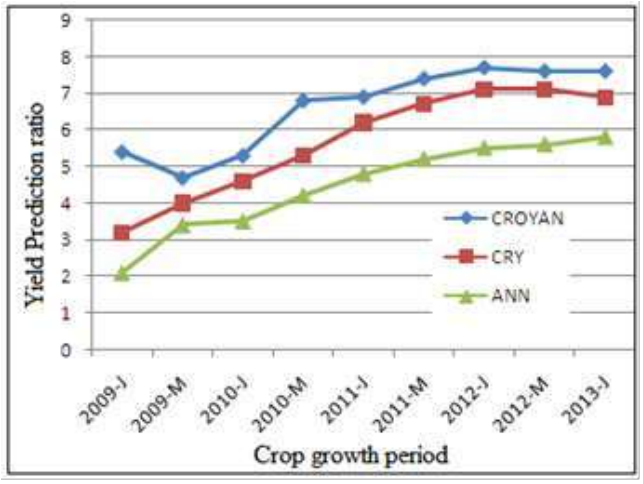


Fig: 7. Crop Yield analyzed over CROYAN and other approaches

IIOAB JOURNAL
ISSN: 0976-3104

Figure-7 compares the performance of crop yield analysis over CROYAN, CRY approach which uses Bee Hive computational approach and ANN approach which shows that the yield prediction ratio achieved over CROYAN is higher than other approaches. ANN approach show a lower performance for the dataset used with 250 records primarily due to memory occupation during runtime and achieve local optima.

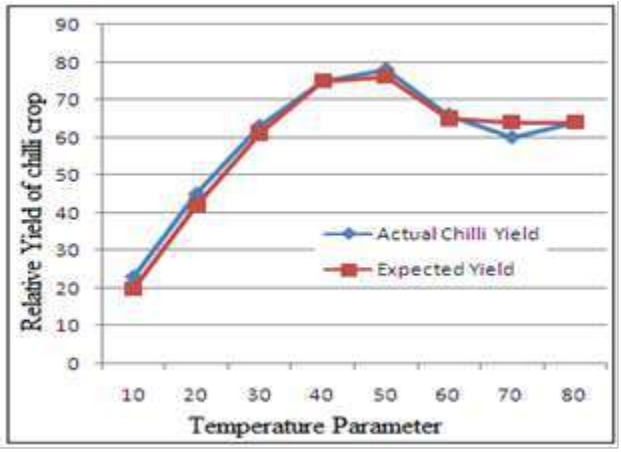


Fig: 8 Chilli Crop Yield achieved for Actual and Observed Yield

Figure–8 explains the performance of CROYAN over Chilli crop, where the observed yield and predicted yield are analyzed for variable 'H' **[Table–1]** for temperature as metric. The obtained yield is identified to be relative to temperature which aids in growth of Chilli crop for Guntur region. As Chilli crop is highly grown under humid climates as well in regions which require high sunlight, the field 'H' is considered for analysis.

CONCLUSION

This work investigates the crop yield and variability crop yield prediction in Chilli, Sugarcane. Fuzzy Cognitive Map is adopted in this research work to suggest and identify the yield among crops. The analysis had been carried out using NDVI parameters and its CYI metrics which determine the yield estimate and reference metrics. FCM approach is compared with other traditional approaches, such as CRY and ANN approaches, it can be noticed that CROYAN approach converges to every local optimum, hence the observed yield and actual yield datasets has much similarities. Further the work can be improved to find the suitability of variable climatic situations and challengeable crop growth analysis.

CONFLICT OF INTEREST
Authors declare no conflict of interest.

ACKNOWLEDGEMENT
None.

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

REFERENCES

[1] Akyildiz Anup K Prasad, Lim Chai, Ramesh P Singh, Menas Kafato.[2006] *Crop yield estimation model for Iowa using remote sensing and surface parameters, International Journal of Applied Earth Observation and Geoinformation* 8: 26–33.
[2] Asefa Taa, Douglas Tanner, Alan TP Bennie. [2004]Effects of stubble management, tillage and cropping sequence on wheat production in the south-eastern highlands of Ethiopia Soil & Tillage Research 76: 69–82.
[3] Bart Kosko, Fuzzy Cognitive Maps.[1986] *International Journal of Man Machine studies*, 24: 65–75.
[4] Inmaculada Pulido-Calvo, Juan Carlos Gutie´rrez-Estrada.[2009] Improved irrigation water demand forecasting using a soft-computing hybrid model biosystems engineering 10: 202–218.
[5] T Itoh, H Ishii and T *Nanseki* .[2003]A Model of Crop Planning under Uncertainty in Agricultural Management, *International Journal of Production Economics*, 81-82: 555–558.

[6] Lin CT and Lee CSG.[1996] Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems. Prentice Hall, Upper Saddle River, *NJ*, USA.
[7] Liu J, Goering CE, Tian L. [2001] A neural network for setting target corn yields, Transactions of the ASAE 44(3):705–713.
[8] Markinos AT, Papageorgiou EI, Stylios CD ,Gemtos TA. [2007] Introducing Fuzzy Cognitive Maps for decision making in precision agriculture. In: Proceedings of 6th European Conference on Precision Agriculture (6th ECPA), ed. JV Stafford, *Wageningen Academic Publishers*, Netherlands: 77–86.
[9] Moreno MA, Planells P, Ortega JF, Tarjuelo J. [2007] New methodology to evaluate flow rates in on-demand irrigation networks. *Journal of Irrigation and Drainage Engineering*, 133(4): 298–306.
[10] SA Mohaddes and MG Mohayidin.[2008] Application of the Fuzzy Approach for Agricultural Production Planning in a Watershed, a Case Study of the Atrak Watershed Iran, *American-Eurasian Journal of Agriculture & En-vironment Science*, 3(4) : 636–648.
[11] Miao Y, Mulla DJ and Robert PC. [2006] Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. *Precision Agriculture*, 7:117–135.
[12] Neapolitan RE. [2003] Learning Bayesian Networks. Prentice Hall Publishers, First edition, Upper Saddle River, *NJ*, USA
[13] Papageorgiou E, Stylios C and Groumpos P.[2003] An Integrated Two-Level Hierarchical Decision Making System based on Fuzzy Cognitive Maps (FCMs). *IEEE Trans Biomed Engin*, 50 (12):1326–1339.
[14] Papageorgiou EI. [2009] A novel approach on designing augmented Fuzzy Cognitive Maps using fuzzified decision trees, In: Proceedings at 4th international Symposium of Intelligence, Computations and Applications, ISICA 2009, Computers in Communication and Intelligent Systems-CCIS 51, pp. 266–275, by Z. Cai et al. (Eds.), Springer-Verlag Berlin Heidelberg 2009, ISICA 2009, 23-25 October, China.
[15] Rodrıguez-Dıaz JA, Camacho-Poyato Lopez-Luque R.[2004] Application of data envelopment analysis to studies of irrigation efficiency in Andalusia. *Journal of Irrigation and Drainage Engineering*, 130(3): 175–183.
[16] Stach W, Kurgan L and Pedrycz W. [2007] Parallel Learning of Large Fuzzy Cognitive Maps. Proceedings of *International Joint Conference on Neural Networks, Orlando, Florida, USA*, August 12–17.
[17] Sadiq R, Kleiner Y and Rajani B. [2006] Interpreting fuzzy cognitive maps (FCMs) using fuzzy measures to evaluate water quality failures in distribution networks. *National Research Council of Canada*, http://irc.nrc-cnrc.gc.ca
[18] Tsadiras A. [2008] Comparing the inference capabilities of binary, trivalent and sigmoid fuzzy cognitive maps. *Information Sciences*
[19] Xirogiannis G and Glykas M. [2007] Intelligent Modeling of e-Business Maturity, Expert Systems with Applications, 32/2:687–702
[20] Xirogiannis G, Stefanou J and Glykas M. (2004] A fuzzy cognitive map approach to support urban design. Expert Systems with Applications, 26 (2):257–268 http://sstoolbox.software.informer.com/
[21] Zhou *SL, McMahon TA, Walton A, Lewis J. [2002] Forecasting operational demand for an urban water supply zone, Journal* of *Hydrology*, 259: 189–202.
[22] Gunasundari Anantharaj, Arunkumar Thangavelu, Hemavathy Ramasubbian.[2013] CRY - An improved crop yield prediction model using bee hive clustering approach for agricultural data sets, pp-473-478, Proceedings of 2013 International Conference on Pattern Recognition, *Informatics, and Mobile Engineering*, 2013.

Ms. M.Gunasundari *is an established educator with experience in Education, Curriculum Planning and Development, and Educational Thought and Policy. Gunasundari has completed her Post Graduate degree at Nallamuthu Gounder Mahalingam College, Pollachi in the year 1998. In her extensive career, Ms.Gunasundari has been a Lecturer, Head of the Department in the MCA Department, Nallamuthu Gounder Mahalingam College, Pollachi. She was with SVS Institute of Computer Applications. Currently She is working as a Associate Professor and Head of the Department in the Department of Computer Science, Nehru Arts and Science College, Coimbatore. Ms.Gunasundari has played leading roles in various curriculum committees at Bharathiar University, Bharathidasan University, Periyar University etc., and presented conference and seminar papers on curriculum and pedagogy locally. She held memberships in various local and international bodies like ICTACT, CSI and IEEE etc.*

Dr. Arunkumar Thangavelu *is working as an Assistant Director at School of Computing Sciences, Vellore Institute of Technology.*