International Journal of Computational Intelligence and Informatics, Vol. 2: No. 1, April - June 2012

Effective Feature Selection via Featuristic Genetic on Heart Data

A. Pethalakshmi
Associate Professor & Head,
Department of Computer Science,
M.V.M. Government Arts College (W),
Dindigul-624 001, Tamil Nadu, India.
E-mail: pethalakshmi@yahoo.com

A.Anushya
Ph.D., Scholar,
Department of Computer Science,
Manonmaniam Sundaranar University,
Tirunelveli- 627 012, Tamil Nadu, India.
E-mail: anushya.alpho@gmail.com

***Abstract*- This work proposes a new algorithm namely compound featuristic genetic algorithm. We evaluate the problems while using genetic based feature selection, and propose more efficient technique for improving heart disease prediction. Genetic algorithm have been proposed and applied successfully to solve a wide variety of problems. Feature selection is the process of removing irrelevant features. It brings into play in reducing execution time and improving predictive accuracy of the classifier. In this paper, we examine the performance of four fuzzy classifiers using the proposed algorithm on heart data. The fusion of Fuzzy Logic with the classifiers Decision trees, K-means, Naive bayes and Neural network are used to evaluate the accuracy of occurrence of heart disease. The experiments are carried out on heart data set of UCI machine learning repository and it is implemented in MATLAB.**

Keywords- Naive bayes, K-means, Neural network, Decision tree, Fuzzy, Genetic, Memetic, Compound featuristic genetic, Heart disease.

I. INTRODUCTION

A. Data mining

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. The techniques in data mining are discovering new trends and patterns of behavior that previously went unnoticed. Once they've uncovered this vital intelligence, it can be used in a predictive manner for a variety of applications.

B. Fuzzy logic and Fuzzy sets

Fuzzy sets and fuzzy logic allows what is referred to as approximate reasoning. With fuzzy sets, an element belongs to a set to a certain degree of certainty. Fuzzy logic allows reasoning with these uncertain facts to infer new facts, with a degree of certainty associated with each fact. Fuzzy set is any set that allows its members to have different grades of membership in the interval [0, 1]. Fuzzy classification offers an alternative to crisp logic by evaluating data set based on their membership into each category. Fuzzy membership assumes that membership to a given category will range from complete membership (100%) to non-membership (0%), and that dataset may be classified as partial members into two or more categories.

C. Genetic algorithm

Genetic Algorithm incorporates natural evolution methodology. The genetic search starts with zero attributes, and an initial population with randomly generated rules. Based on the idea of survival of the fittest, new population is constructed to comply with fittest rules in the current population, as well as offspring of these rules. Offspring are generated by applying genetic operators crossover and mutation. The process of generation continues until it evolves a population P where every rule in P satisfies the fitness threshold.

D. Feature Selection

Feature selection, by identifying the most salient features for learning, focuses a learning algorithm on those aspects of the data which is most useful for analysis and future prediction. Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis.

ISSN: 2349 - 6363

E. Heart Disease

Heart disease is the world's leading killer, accounting for 29.2% of total global deaths in 2003. The World Health Organization in 2009 estimated that almost 20 million deaths occur annually from Heart disease and that by 2030 that figure could rise to almost 24 million. The World Health Organization estimated that 60% of the world's cardiac patients are Indian. Prediction of this disease will help to prevent it in its early stage. This work compares efficient to find the best Fuzzy logic rule based classifier, which is used as an effective tool to improve the classification accuracy. It can be implemented by the classifiers such as fuzzy naive bayes, fuzzy decision tree, fuzzy neural network and fuzzy k-means by using the compound featuristic genetic algorithm .

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3 presents data source. In section 4, we propose the new algorithm. Section 5 describes experimental analysis performed with dataset. Finally, Section 6 concludes the paper.

II. RELATED WORKS

Up to now, the extensive literatures have been reviewed that have paying attention on classification, fuzzy set and feature selection. These studies have related different approaches to the given problem and achieved high classification accuracies. Here we have few examples:

Wu,et al.[14] proposed that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome.

A. J.Myles [1], introduced a new method for the induction of fuzzy decision trees. It was introduced to determine the location and the associated uncertainty for each decision boundary during the construction process. The new fuzzy decision tree classifier was shown to compare favorably to both standard and bagged decision tree classifiers in prediction accuracy and model size.

Harleen et al. [6] examined the potential use of classification data mining technique like decision tree, rule induction and artificial neural network for diagnosis of diabetic patients.

Carlos Ordonez [5] implemented efficient search for diagnosis of heart disease comparing association rules with decision trees. Association rules were compared to predictive rules mined with decision trees, a well-known machine learning technique. In this work constrained association rules were used to predict multiple related target attributes, for heart disease diagnosis. The goal was to find association rules predicting healthy arteries or diseased arteries, given patient risk factors and medical measurements.

Latha Parthiban et al. [9] introduced a new approach based on coactive neuro-fuzzy inference system and was presented for prediction of heart disease. The proposed coactive neuro-fuzzy inference system model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach which was then integrated with genetic algorithm to diagnose the presence of the disease.

Sellapan Palaniappan et al. [12] developed a prototype Intelligent Heart Disease Prediction System using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees.

Weiguo Sheng et al. [13] proposed an approach for simultaneous clustering and feature selection using a niching memetic algorithm which made feature selection an integral part of the global clustering search procedure and attempts to overcome the problem of identifying less promising locally optimal solutions in both clustering and feature selection, without making any prior assumption about the number of clusters. Within the NMA_CFS procedure, a variable composite representation was devised to encode both feature selection and cluster centers with different numbers of clusters. In an experimental evaluation, demonstration the effectiveness of the proposed approach and compared it with other related approaches, using both synthetic and real data.

Asha Rajkumar et al. [4], projected the data classification and it was based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. This paper dealt with the results in the field of data classification obtained with Naive Bayes algorithm, Decision list algorithm and k-nearest neighbor‘s algorithm. Naive Bayes algorithm played a key role in shaping improved classification accuracy of a dataset.

Srinivas, K et al. [8], automated a system for medical diagnosis would enhance medical care and reduce costs. In this paper popular data mining techniques namely, Decision Trees, Naïve Bayes and Neural Network were used for prediction of heart disease.

M.Anbarasi et al. [10] exhibited that Decision Tree was used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. Naïve Bayes performed consistently

before and after reduction of attributes with the same model construction time. Classification via clustering performed poor compared to other two methods.

Ali.Adeli et al. [2] implemented a Fuzzy Expert System for heart disease diagnosis. Fuzzy Expert System for Heart Disease Diagnosis designed with follow membership functions, input variables, output variables and rule base. Designing of this system with fuzzy base in comparison with classic designed improve the results. This fuzzy expert system that dealt with diagnosis has been implemented.

K. Rajeswari et al. [7] discussed to reduce the number of features using Genetic algorithms is made. The system was a theoretical study which proposed implementation of Machine Intelligence algorithms. More important features are selected using Genetic Algorithm and a Risk factor can be made by summing the Risk score's of the various features. It was anticipated that data mining could help in the identification of risk subgroups of subjects for developing future events and it might be a decisive factor for the selection of therapy, i.e., angioplasty or surgery.

III. DATA SOURCE

Available dataset of Heart disease from UCI Machine Learning Repository has been considered for classification process. A total of 909 records with 13 medical attributes (factors) were obtained from the Heart Disease database. Figure 1 lists the attributes.

Input attributes

1. Sex (value 1: Male; value 0: Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0 :< 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in year

Figure 1. Attributes list and description

IV. PROPOSED ALGORITHM

In this section we compare the existing four classifiers namely, fuzzy Naive bayes, fuzzy K-means, fuzzy Neural network, fuzzy Decision tree which are considered under the proposed compound featuristic genetic algorithm. Before introducing the proposed algorithm, we present the information regarding the fuzzy classifiers.

A. Fuzzy Naive Bayes Algorithm:

The steps involved in Fuzzy Naive Bayes Algorithm are given below:

Step: 1 Initially split over all cases for the actual attribute values, assuming the class depends only on these

values.

Step: 2 Assume that the attribute values of example e are independent of each other and obtain

$$\sum_{v_1 \in V_{1,\dots} v_n \in V_n} P(c|v_1 \dots v_n) P(v_1|e) \dots P(v_n|e)$$

Step: 3 Now we can lift our reinterpretation of the fuzzy truth values and go back to membership degrees

and get $\sum_{v_1 \in V_{1,\dots} v_n \in V_n} P(c|v_1 \dots v_n) \mu^e_{v_1} \dots . \mu^e_{v_n}$

Step 4: Application of the Bayes rule to P(c | v1 . . . vn) yields:

$$\sum_{v_1 \in V_{1,\dots} v_n \in V_n} \frac{P(v_1 \dots v_n|c) P(c)}{P(v_1 \dots v_n)} \mu^e_{v_1} \dots . \mu^e_{v_n}$$

Step 5: Now we apply the same naive independence assumption as in the classical case.

$$\sum_{v_1 \in V_{1,\dots} v_n \in V_n} \frac{P((v_1 \mid c) \dots (v_n|c)}{P(v_1) \dots P(v_n)} \mu^e_{v_1} \dots . \mu^e_{v_{n1}}$$

Step 6: We move constant factors in front of the first sum and repeat this with the other sums. Finally, we find

using distributivity: $P(c|e) = P(c) \left(\sum_{v_1 \in V_{1,}} \frac{P(v_1|c)}{P(v_1)} \mu^e_{v_1} \right) \dots . \left(\sum_{v_1 \in V_n} \frac{P(v_n|c)}{P(v_n)} \mu^e_{v_n} \right)$

B. Fuzzy Decision Tree:

The input attributes are automatically discredited in linguistic terms, based on the distribution of pattern points in the feature space. In this technique, different forms of fuzzy entropy are computed at the node level, in terms of class membership, to take care of overlapping classes. Also, pruning is used to minimize noise, resulting in a smaller decision tree with more efficient classification.

Any input feature value is described in terms of some combination of overlapping membership values in the linguistic property sets low, medium and high. An n-dimensional pattern

Fi=[a1,a2,a3,…,an] is represented as a 3 -dimensional vector

Fi=[μlow(a1)(Fi),μmedium(a1)(Fi),μhigh(a1)(Fi),…,μhigh(an)(Fi)]

Where, μ values indicate the membership functions of the corresponding linguistic functions like low, medium and high along each feature axis. Each value is then discretized, using a threshold (generally 0.5), to enable a convenient mapping in the C4.5 framework.

This discretization to Boolean form speeds up computation by reducing the complexity of the search space. However the linguistic flavor of the attributes is retained, thereby enabling the extraction of more user-friendly natural rules that are then mapped to the fuzzy knowledge-based network. When the input feature is numerical, we divide it into three partitions (with range [0, 1]) using only two parameters. The formulae and concepts are detailed in.

C. Fuzzy K-Means Algorithm :

The fuzzy K-means algorithm generalizes the classic or hard K-means algorithm. The goal of k-means algorithm is to cluster n objects (here documents) in k clusters and find k mean vectors for clusters (here centroids). In the context of the vector space model for information retrieval we call these mean vectors concepts. The spherical K-means algorithm used in is just a variation of the hard K-means algorithm which uses the fact that document vectors (and concept vectors) are of the unit norm.

As opposed to the hard K-means algorithm which allows a document to belong to only one cluster, fuzzy K-means algorithm allows a document to partially belong to multiple clusters which generalizes the classic or hard K-means algorithm. The goal of k-means algorithm is to cluster n objects (here documents) in k clusters and find k mean vectors for clusters (here centroids). In the context of the vector space model for information retrieval we call these mean vectors concepts. The spherical K-means algorithm used in is just a variation of the hard K-means algorithm which uses the fact that document vectors (and concept vectors) are of the unit norm. As opposed to the hard K-means algorithm which allows a document to belong to only one cluster, fuzzy K-means algorithm allows a document to partially belong to multiple clusters [3].

D. Fuzzy Neural Network Algorithm:

The steps implicated in Fuzzy Neural Network Algorithm are given below:

Step 1: Read in the data file (the number of features N, the number of feature vectors Q, the dimension J of the labels, the number K of classes, all Q feature vectors and all Q labels).

Step 2: Find minimal distance Dmin over all feature vector pairs
Put F = Dmin/2
Put G = Q //Starting no. Gaussian centers

Step 3: Find two exemplar vectors of min. distance d with indices k1 and k2
If d < (½)Dmin //If vectors are close and
If label[k1] = label[k2] // have same label
Eliminate Gaussian center k2
G = G -1 //Reduce no. Gaussians
Goto Step 2

Step 4: Input next unknown x to FNN to be classified
For k = 1 to G do //For each Gaussian center
Compute g[k] = exp{-||x - x(k)||2/(2F2)}
Find maximum g[k*], over k = 1,...,G
Output x, label[k*] //label[k*] is class of x

Step 5: If all inputs for classifying are done, stop
Else, goto Step 3.

E. Genetic Algorithm:

The steps of the existing genetic algorithm are given below:

Step 1: Initialize the population and enter Step 2.

Step 2: Ranking the individuals using any ranking method and enter Step 3.

Step 3: Now the genetic algorithm in conjunction with the classification method is used to select the smallest subset of data from the above selected M values that gives maximum accuracy.

Step 4: Recombine individuals generating new ones and enter Step 4.

Step 5: Mutate the new individuals and enter step 5.

Step 6: If the stopping criterion is satisfied, STOP; otherwise, replace old individuals with the new ones restructure the population tree and return to Step 2.

Step 7: Finally presents a fitness function to Fitness (x) = A(x) + P/N (x) to maximize the accuracy where for chromosome x

F. Memetic Algorithm:

The existing genetic algorithms show that pure genetic algorithms are not well suited to fine tuning structures in complex search spaces and have its own disadvantages like pre matured convergence and population diversity. It can be hybridized with other techniques can greatly improve their efficiency. Genetic algorithms that have been hybridized with local searches are also known as memetic algorithms. Memetic algorithm which is a powerful combination of genetic algorithm and local search is the key technique that has been used here[10]. Even it is hybridized its search space time is more. For reducing time and space complexity, the niching technique is introduced to improve genetic. Niching and memetic both are simultaneously used for improving performance of genetic. Further, local search operations are introduced to refine feature selection and cluster centers encoded in the chromosomes. Finally, the niching method is integrated to preserve the population diversity and prevent premature convergence [12].

Here, we suggest the compound featuristic genetic algorithm. The proposed algorithm works with variable composite chromosomes, which are used to represent solutions. The operation of the algorithm consists of using a niching selection method for selecting pairing parents for reproduction, performing different genetic operators on different parts of the paired parents, applying local search operations to each offspring, and carrying out a niching competition replacement. The evolution is terminated when the fitness value of the best solution in the population has not changed for g generations. The output of the algorithm is the best solution encountered during the evolution. The steps of the compound featuristic algorithm are illustrated below:

G. Compound featuristic genetic algorithm:

```
Procedure Compound Featuristic Genetic Algorithm;
Begin
Initialize population;
for each individual to local-search individual;
    repeat
    for individual = 1 to #crossovers do
    select two parent individual1, individual2 in population  randomly;
    individual3:=crossover(individual1, individula2);
    individual3 := local-search (individual3);
    find smallest HD(child, individual3);
            of those find parent with worst fitness;
    calculate fitness (child);
    if better fitness: exchange (child, individual3);
    add individual i3 to population;
end for;
for individual=1 to #mutations do
    select an individual of population randomly;
    individual {m} := mutate (individual);
    individual {m} := local-search (individual{m});
    find smallest HD(child, individual{m});
    of those find parent with worst fitness;
    calculate fitness (child);
    if better fitness: exchange (child, individual{m});
    add individual {m} to population;
end for;
  population :=select  (population);
if population converged then
    for each individual of best populations do individual :=local-search (mutate(individual));
end if
    until terminate = true;
end
```

V. EXPERIMENTAL RESULTS

Experiments are carried out on heart data sets from UCI Machine Learning Repository. The dataset is evaluated using 10-fold cross validation and the results are compared the classification accuracy. Experiments are conducted with MATLAB. Data sets of 909 records with 13 attributes are used. To enhance the prediction of classifiers, Fuzzy logic is incorporated. The classifiers such as Decision tree, Naive Bayes, Neural network and K-means are integrated with fuzzy techniques and used for diagnosis of patients with heart disease. The four classifiers are studied under the proposed algorithm. Comparative analysis of various classifiers is shown in Table 1 and Figure 2. Observations exhibit that the Fuzzy K-means classification technique outperforms than other three classification techniques after incorporating fuzzy techniques.

Among the thirteen features, more important features are selected via the Compound featuristic genetic algorithm. Reduced attributes achieved for heart data set after affecting the Compound featuristic genetic

algorithm are: Type(Chest Pain Type), Rbp (Resting blood pressure), Eia (Exercise induced angina) and Vsl (No. of vessels colored).

VI. CONCLUSION AND FUTURE WORK

We have projected this scheme to reduce the attributes of the heart dataset. From the result, out of the four classifiers, fuzzy k-means inside proposed algorithm provides the best result. It is proved that the Compound featuristic genetic algorithm produced minimal trim down for the data sets. The proposed method picks a small subset of features that is used to predict heart disease. The method described in this paper has demonstrated that the approach is able to reduce the number of features selected as well to increase the classification rate. Among 13 attributes in the heart data set, 4 attributes only preferred for decision making. The compact 4 attributes are adequate for diagnosing heart patient or not. It achieves reasonable fair data in heart disease. It would be a promising algorithm for the heart disease all over the world in present scenario. This investigation assists in the complexity of processing time and space also requires much less computation. In near future, this work can extend the same by exploring other data mining techniques for the Intelligent Heart Disease Prediction System.

TABLE 1: COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS

Classifiers	Classification Accuracy (%)
Fuzzy Decision Tree	**90.06**
Fuzzy Naïve Bayes	**89.62**
Fuzzy Neural network	**91.09**
Fuzzy K-means	**99.49**

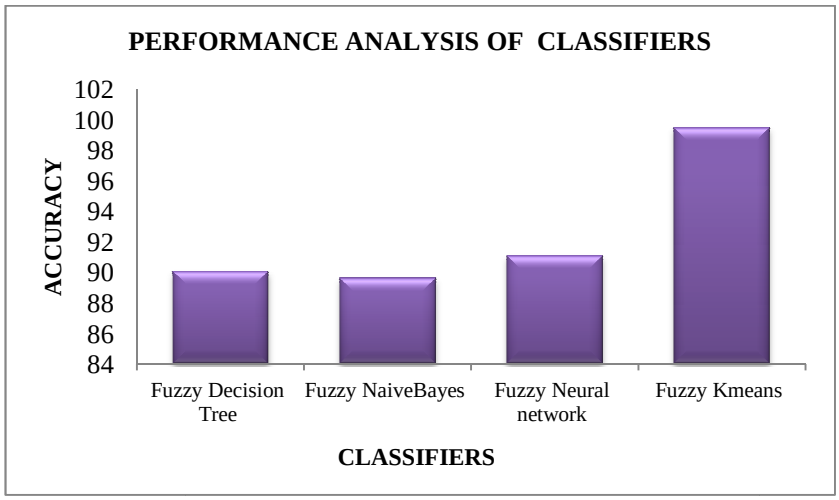

Figure 2: Performance Analysis of Classifiers

REFERENCES

[1] A. J.Myles and S.D.Brown, "Induction of decision trees using fuzzy partitions," Journal OF Chemometrics 17: , pp. 531–536,2003.

[2] Ali.Adeli and Mehdi.Neshat, "A Fuzzy Expert System for Heart DiseaseDiagnosis," Proceedings of International Multiconference of Engineers and computer scientist,2010,Vol I, IMECS,2010.

[3] Andreas Meier and Nicolas Werro,"A Fuzzy Classification Model for Online Customers," Informatica 31, pp. 175–182,2007.

[4] Asha Rajkumar and Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," Global Journal of Computer Science and Technology ,Vol. 10 Issue 10 Ver. 1.0 GJCST, pp. 17-24,2010.

[5] Carlos Ordonez , "Comparing Association Rules and Decision Trees for Disease Prediction," HIKM'06 pp. 38,20006.

[6] Harleen Kaur and Siri Krishan Wasan, "Empirical Study On Applications Of Data Mining Techniques In Healthcare," Journal Of Computer Science 2 (2): 194-200, ISSN pp.1549-3636,2006.

[7] K. Rajeswari, Dr. V. Vaithiyanathan , Dr.P. Amirtharaj," Prediction of Risk Score for Heart Disease in India Using Machine Intelligence," International Conference on Information and Network Technology(IPCSIT) vol.4,(2011).

[8] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan,"Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," (IJCSE) International Journal on Computer Science and Engineering,Vol. 02, No. 02, , pp. 250-255,2010.

[9] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences 3:3,2007.

[10] M.Anbarasi, E. Anupriya and N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," International Journal of Engineering Science and Technology Vol. 2(10), pp.5370-5376, 2010.

[11] Md. Kamrul Islam and Madhu Chetty,"Protein Structure Prediction:Clustering of Memetic Algorithm in Protein Structure Prediction," IEEE Science and Engineering Graduate Research Expo 2009,The University of Melbourne, Australia

[12] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, pp.343,2008.

[13] Weiguo Sheng, Xiaohui Liu, and Michael Fairhurst, "A Niching Memetic Algorithm for Simultaneous Clustering and Feature Selection," IEEE Transactions On Knowledge And Data Engineering, VOL. 20, NO. 7, July 2008.

[14] Wu, R., Peters, W., Morgan, M.W., "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice," Journal Healthcare Information Management.16 (4), pp 50-55, 2002.

[15] Y.G. Petalas and M.N. Vrahatis," Memetic Algorithms for Neural Network training On Medical data," In European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systemsm, (EUNITE2004), June 10-12, 2004, Aachen, Germany.

Pethalakshmi Annamalai, received the Master of Computer Science from Alagappa University, Karaikudi, Tamilnadu, India, in 1988 and received the Master of Philosophy in Computer Science from Mother Teresa Women's University, Kodaikanal, Tamilnadu, India, in 2000. She has received her Ph.D. Degree from Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India, in 2008. Currently she is working as Associate Professor and Head, Department of Computer Science, M.V.M. Govt. Arts College (w), Dindigul, Tamilnadu, India. Her areas of interests include fuzzy, rough set, neural network and grid computing.

A.Anushya was born in Nagercoil, Tamil Nadu (TN), India, in 1985. She received the Bachelor of Computer Science (B.Sc.,) degree from the Mother Theresa University, Kodaikanal, TN, India, in 2006 and the Master of Computer Applications (M.C.A.) degree from the Bharathidasan University, Tiruchirapally, TN, India, in 2009. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Manonmaniam Sundaranar University, Tirunelveli,TN, India. Her research interests include data mining, and artificial intelligence.