Verdict Accuracy of Quick Reduct Algorithm using Clustering and Classification Techniques for Gene Expression Data

T.Chandrasekhar[1], K.Thangavel[2] and E.N.Sathishkumar[3]

[1] Department of Computer Science, Periyar University, Salem, Tamilnadu-636 011, India

[2] Department of Computer Science, Periyar University, Salem, Tamilnadu-636 011, India

[3] Department of Computer Science, Periyar University, Salem, Tamilnadu-636 011, India

Abstract

In most gene expression data, the number of training samples is very small compared to the large number of genes involved in the experiments. However, among the large amount of genes, only a small fraction is effective for performing a certain task. Furthermore, a small subset of genes is desirable in developing gene expression based diagnostic tools for delivering reliable and understandable results. With the gene selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only the marker genes. An important application of gene expression data in functional genomics is to classify samples according to their gene expression profiles. Feature selection (FS) is a process which attempts to select more informative features. It is one of the important steps in knowledge discovery. Conventional supervised FS methods evaluate various feature subsets using an evaluation function or metric to select only those features which are related to the decision classes of the data under consideration. This paper studies a feature selection method based on rough set theory. Further K-Means, Fuzzy C-Means (FCM) algorithm have implemented for the reduced feature set without considering class labels. Then the obtained results are compared with the original class labels. Back Propagation Network (BPN) has also been used for classification. Then the performance of K-Means, FCM, and BPN are analyzed through the confusion matrix. It is found that the BPN is performing well comparatively.

Keywords: *Rough set theory, Feature Selection, Gene Expression, Quick Reduct, K-means, Fuzzy C means, BPN.*

1. Introduction

Feature selection is the process of choosing the most appropriate features when creating the model of the process. Most of the feature selection methods are applied across the entire data set. Once such genes are chosen, the creation of classifiers on the basis of the genes is another undertaking. If we survey the established investigations in this field, we will find that almost all the accurate classification results are obtained based on more than two genes. Rough sets have been used as a feature selection methods by many researchers among them Jensen and Schen, Zhong et al, Wang and Hu et al. The Rough set approach to feature selection consists in selecting a subset of features which can predict the classes as well as the original set of features. The optimal criterion for Rough set feature selection is to find shortest or minimal reducts while obtaining high quality classifiers based on the selected features. Here we propose a feature selection method based on rough set theory for reducing genes from large gene expression database [1, 4].

Discriminant analysis is now widely used in bioinformatics, such as distinguishing cancer tissues from normal tissues. A problem with gene expression analysis or with any large dimensional data set is often the selection of significant variables (feature selection) within the data set that would enable accurate classification of the data to some output classes. These variables may be potential diagnostic markers too. There are good reasons for reducing the large number of variables:
1) An opportunity to scrutinize individual genes for further medical treatment and drug development. 2) Dimension reduction to reduce the computational cost. 3) Reducing the number of redundant and unnecessary variables can improve inference and classification. 4) More interpretable features or characteristics that can help identify and monitor the target diseases or function types [5].

The rest of the paper is organized as follows: Section 2, briefs about the Rough set theory. Section 3 describes the clustering techniques. Section 4 briefs about classification techniques. Section 5 describes the confusion matrix for

two class problem. Section 6 explains briefly about experimental analysis and results. Section 7 presents a conclusion for this paper.

2. Rough Set Theory

Rough set theory (Pawlak, 1991) is a formal mathematical tool that can be applied to reducing the dimensionality of datasets. The rough set attribute reduction method removes redundant input attributes from datasets of discrete values, all the while making sure that no information is lost. The approach is fast and efficient, making use of standard operations from conventional set theory [3].

Definition: Let U be a universe of discourse, $X \subseteq U$, and R is an equivalence relation on U. U/R represents the set of the equivalence class of U induced by R. The *positive region* of X on R in U, is defined as $pos(R,X) = U\ \{Y \in U/R \mid Y \subseteq X\}$.

The partition of U, generated by $IND\ (P)$ is denoted U/P. If $(x, y) \in IND\ (P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]p$. The indiscernibility relation is the mathematical basis of rough set theory. Let $X \subseteq U$, the P-lower approximation $\underline{PX}$ and P-upper approximation $\overline{PX}$ of set X can be defined as:

$$\underline{P}X = \{ x \in U \mid [x]p \subseteq X \} \quad (1)$$

$$\overline{PX} = \{ x \in U \mid [x]p \cap X \neq \varphi \} \quad (2)$$

Let P, $Q \subseteq A$ be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$\mathrm{POS}_P(Q) = \bigcup_{X \in U/Q} \underline{P}X \quad (3)$$

$$\mathrm{NEG}_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}X \quad (4)$$

$$\mathrm{BND}_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X \quad (5)$$

An important issue in data analysis is discovering dependencies between attributes dependency can be defined in the following way. For P, $Q \subseteq A$, P depends totally on Q, if and only if $IND\ (P) \subseteq IND\ (Q)$. That means that the partition generated by P is finer than the partition generated by Q. We say that Q depends on P in a degree $0 \leq k \leq 1$ denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = \frac{\left|\mathrm{POS}_P(Q)\right|}{\left|U\right|} \quad (6)$$

If $k = 1$, Q depends totally on P, if $0 \leq k \leq 1$, Q depends partially on P, and if $k = 0$ then Q does not depend on P. In other words, Q depends totally (partially) on P, if all (some) objects of the universe U can be certainly classified to blocks of the partition U/Q, employing P. In a decision system the attribute set contains the condition attribute set C and decision attribute set D, i.e. $A = C \cup D$. The degree of dependency between condition and decision attributes, $\gamma_C(D)$, is called the quality of approximation of classification, induced by the set of decision attributes[6,10].

2.1 Quick Reduct Algorithm

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset R of the conditional attribute set C such that $\gamma R(D) = \gamma C(D)$. A given dataset may have many attribute reduct sets, so the setR of all reducts is defined as:

$$R_{all} = \{X \mid X \subseteq C, \gamma X(D) = \gamma C(D);\ \gamma X - \{a\}(D) \neq \gamma X(D),\ \forall a \in X\}. \quad (7)$$

The intersection of all the sets in Rall is called the core, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the representation of the dataset. For many tasks (for example, feature selection), a reduct of minimal cardinality is ideally searched for. That is, an attempt is to be made to locate a single element of the reduct set Rmin⊆Rall:

$$R_{min} = \{X \mid X \in R_{all},\ \forall Y \in R_{all}, |X| \leq |Y|\}. \quad (8)$$

The Quick Reduct algorithm shown below [8, 9], it searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset; attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until the search produces its maximum possible dependency value for that dataset ($\gamma c(D)$). Note that this type of search does not guarantee a minimal subset and may only discover a local minimum.

QUICKREDUCT(C, D)
C, the set of all conditional features;
D, the set of decision features.
(a) $R \leftarrow \{\}$
(b) ***Do***
(c) $T \leftarrow R$
(d) $\forall x \in (C-R)$
(e) ***if*** $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$

Where $\gamma_{R(D)} = card(POSR(D)) / card(U)$

(f) $T \leftarrow R \cup \{x\}$

(g) $R \leftarrow T$

(h) ***until*** $\gamma_{R(D)} = = \gamma_{C(D)}$

(i) ***return*** R

It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. Other such techniques may be found in [8, 9]

3. Clustering Techniques

Clustering is the process of grouping data into clusters, where objects within each cluster have high similarity, but are dissimilar to the objects in other clusters. Similarities are assessed based on the attributes values that best describes the objects. Often distance measures are used for the purpose. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. In this work K-Means, FCM and BPN algorithms which are used to classify the data.

3.1 K-Means Algorithm

K-Means algorithm (MacQueen, 1967) is one of a group of algorithms called partitioning methods[11]. The k-mean algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm.

Input:

Set of sample patterns $\{x1, x2, \ldots, x_m\}$, $x_i \in R^n$

Output:

Set of code vectors of quantization $z_1, z_2, \ldots, z_K$, which are centers of the clusters $\{C_1, C_2, \ldots, C_K\}$.

Step 1: Choose K initial cluster centers $z_1, z_2, \ldots, z_K$ randomly from the m patterns $\{x_1, x_2, \ldots, x_m\}$ where $K < m$.

Step 2: Assign pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$, if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$.

Ties are resolved arbitrarily. And compute cluster centers for each point x_i as follows,

$z_i = (1/n_i) \sum x_j$,

$i = 1, 2, \ldots, K$.

$xj \in Ci$

Where n_i is the number of elements belongs to cluster C_i.

Step 3: Assign each pattern x_i to cluster C_j,

where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$

if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$,

where $\| \bullet \|$ is an Euclidean metric norm.

Ties are resolved arbitrarily, without changing the cluster centers z_j, $j = 1, 2, \ldots, K$

Step 4: Stop.

The k-means algorithm is the most extensively studied clustering algorithm and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids [12].

3.2 Fuzzy C Means

Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition [15].

Algorithm Steps:

Step-1: Randomly initialize the membership matrix using this equation,

$$\sum_{j=1}^{C} \mu_j (x_i) = 1 \qquad i = 1,2,\ldots..k \qquad (9)$$

Step-2: Calculate the Centroid using equation,

$$Cj = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \qquad (10)$$

Step-3: Calculate dissimilarly between the data points and Centroid using the Euclidean distance.

Step-4: Update the New membership matrix using the equation,

$$\mu_j(x_i) = \frac{[\frac{1}{d_{ji}}]^{1/m-1}}{\sum_{k=1}^{c} [\frac{1}{d_{ki}}]^{1/m-1}} \qquad (11)$$

Here **m** is a fuzzification parameter, The range **m** is always {1.25, 2}

Step-5: Go back to Step 2, unless the centroids are not changing.

4. Classification Techniques

4.1 Back Propagation Networks (BPN)

Back Propagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value. The target value may be the known class label of the training tuple (for classification problems) or a continuous value (for prediction). For each training tuple, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual target value. These modifications are made in the "backwards" direction, that is, from the output layer, through each hidden layer down to the first hidden layer (hence the name backpropagation) [13,14].

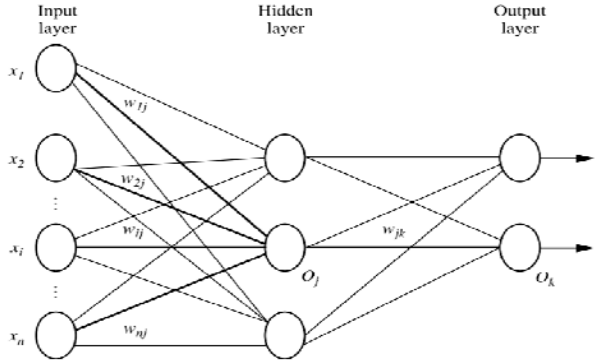

Fig. 1: A multilayer feed-forward neural network.

5. Confusion Matrix for a Two-Class Problem

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier [7].

Table 1: Confusion Matrix for a Two-Class Problem

	Positive prediction	**Negative prediction**
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

Table 1 shows a confusion matrix for a two-class problem with positive and negative class values. From such a matrix it is possible to extract a number of widely used metrics to measure the performance of a classifier, such as error rate, defined as in the following equation,

$$Err = \frac{FP + FN}{TP + FN + TN + FP} \quad (12)$$

and overall accuracy, defined as in the following equation,

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \quad (13)$$

It is possible to derive four performance metrics from Table 1 to measure the classification performance on the positive and negative classes independently:

True positive rate: TPrate is the percentage of positive samples correctly classified as belonging to the positive class.

$$TPrate = \frac{TP}{TP + FN}$$

True negative rate: TNrate is the percentage of negative samples correctly classified as belonging to the negative class.

$$TNrate = \frac{TN}{TN + FP}$$

False positive rate: FPrate is the percentage of negative samples misclassified as belonging to the positive class.

$$FPrate = \frac{FP}{FP + TN}$$

False negative rate: FNrate is the percentage of positive samples misclassified as belonging to the negative class.

$$FNrate = \frac{FN}{FN + TP}$$

The advantages of the four performance measures are of being independent of class costs and a priori probabilities. The aim of a classifier is to minimize false positive and negative rates, or similarly to maximize true negative and positive rates.

6. Experimental Results

6.1 Data Sets

We use four datasets: leukemia, breast cancer, lung cancer and prostate cancer which are available in the website: http://datam.i2r.a-star.edu.sg/datasets/krbd/, [2]. the gene number and class contained in four datasets are listed in Table 2.

Table2: Summary of the four gene expression datasets.

Dataset	**#Gene**	**Class**	**# Samples**
Leukemia	7129	ALL/AML	34 (20/14)
Prostate	12600	Tumor/Normal	21 (8/13)
Breast	24481	Relapse/Non Relapse	19 (12/7)
Lung	7129	Tumor/Normal	96 (86/10)

The data studied by rough sets are mainly organized in the form of decision tables. One decision table can be represented as $S = (U, A=C\ U\ D)$, where U is the set of samples, C the condition attribute set and D the decision attribute set. We can represent every gene expression data with the decision table like Table 3.

Table 3. Microarray data decision table.

Samples	Condition attributes(genes)				Decision attributes
	Gene 1	Gene 2	...	Gene q	Class label
1	g(1,1)	g(1,2)	…	g(1,q)	Class(1)
2	g(2,1)	g(2,2)	…	g(2,q)	Class(2)
…	…	…	…	…	…
p	g(p,1)	g(p,2)	…	g(p,q)	Class(p)

In the decision table, there are p samples and q genes. Every sample is assigned to one class label. Each gene is a condition attribute and each class is a decision attribute. $g(p, q)$ signifies the expression level of gene q in sample p. [2].

6.2 Data Pre-processing, Gene Selection

Before applying feature selection algorithm all the conditional attributes (samples) are normalized using Z-Score normalization and then discretized using K-Means discretization [16]. Let us considered U is the set of samples, C the condition attribute set and D the decision attribute set. By applying Quick Reduct Algorithm, In leukemia dataset gene 4 and gene 3252 are identified, where as in prostate gene dataset, gene 20 and gene 11154 are identified, in breast cancer dataset gene 3 and gene 22019 are identified, finally in lung cancer dataset gene 4817 as best attribute for finding appropriate decision.

Table 4: Features selected by Quick Reduct Algorithm

Gene Data	Identified Attributes (Genes)
Leukemia Cancer	Gene 4, Gene 3252
Prostate Cancer	Gene 20, Gene 11154
Breast Cancer	Gene 3, Gene 22019
Lung Cancer	Gene 4817

6.3 Classification Performance

In this section the selected data is clustered by the K-Means and FCM algorithm. The data presented in Table 5 and 6 shows the classification performance of True Positive (TP) rate, True Negative (TN) rate, False Positive (FP) rate, and False Negative (FN) rate as previously described. Table 7 shows classification performance of Back Propagation Network. Results are presented both in terms of classification accuracy and classification error [7].

Table 5: K-Means Classification Performance Rate

Gene Data	K-Means			
	TP	FP	TN	FN
Leukemia Cancer	0.9820	0.0180	0.2428	0.7571
Prostate Cancer	0.6750	0.3250	0.0769	0.9230
Breast Cancer	0.6667	0.3333	0.4286	0.5714
Lung Cancer	0.8883	0.1116	0.0180	0.9820

Table 6: FCMs Classification Performance Rate

Gene Data	FCM			
	TP	FP	TN	FN
Leukemia Cancer	0.9820	0.0180	0.2428	0.7571
Prostate Cancer	0.3250	0.7750	0.0769	0.9230
Breast Cancer	0.6667	0.3333	0.4286	0.5714
Lung Cancer	0.8883	0.1116	0.0180	0.9820

When comparing classification results, where the BPN method shows a high in classification accuracy, which is demonstrated in Fig. 2.

Table 7: K-Means, FCM and BPN Classification Accuracy

Gene Data	K-Means	FCM	BPN
Leukemia Cancer	0.9721	0.9721	1.0000
Prostate Cancer	0.7421	0.6370	0.8210
Breast Cancer	0.6521	0.6521	0.9160
Lung Cancer	0.9695	0.9695	1.0000

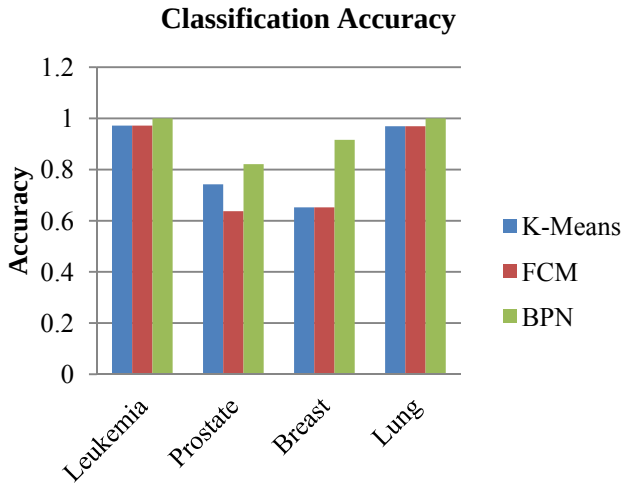

Fig 2: K-Means, FCM and BPN Classification Accuracy

Table 8: K-Means, FCM and BPN Classification Error

Gene Data	K-Means	FCM	BPN
Leukemia Cancer	0.0279	0.0279	0.0000
Prostate Cancer	0.2579	0.3630	0.1790
Breast Cancer	0.3479	0.3479	0.0840
Lung Cancer	0.0305	0.0305	0.0000

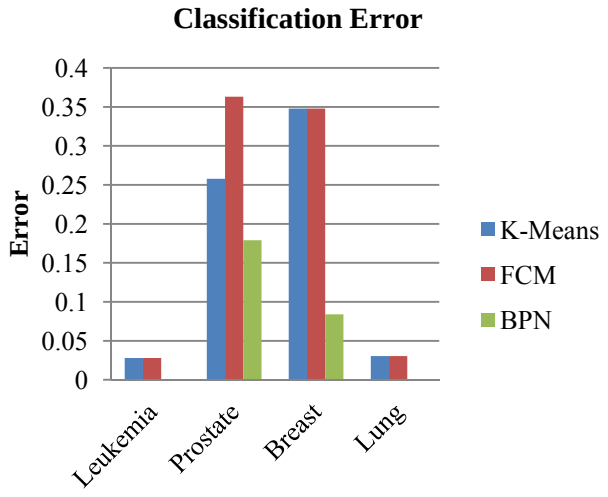

Fig 3: K-Means, FCM and BPN Classification Error

Fig. 2 and Fig.3 demonstrated the classification accuracy and error rate of Quick Reduct algorithm for gene expression data sets.

7. Conclusion

In this paper, Quick reduct algorithm based on rough set theory has been studied for gene expression datasets. The Quick reduct algorithm has been implemented and the obtained reduct features has been applied to find the clusters using K-Means and FCM clustering algorithm without class labels. The performance was evaluated using confusion matrix with positive and negative class values. Further, the selected features with class labels were classified using Back Propagation Network to find the accuracy of the algorithm. The BPN shows best performance for selected features over the K-Means and FCM algorithms.

References

[1] Jensen, R. and Shen, Q. (2003) 'Finding rough set reducts with ant colony optimization', Proceedings UK Workshop on Computational Intelligence, pp. 15–22.

[2] Xiaosheng Wang, Osamu Gotoh, "Cancer Classification Using Single Genes", pp 179-188.

[3] Pawlak, Z. (2002) 'Rough Sets and Intelligent Data Analysis', Information Sciences, Vol. 147, pp. 1–12.

[4] Changjing Shang and QiangShen, "Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study", International Journal of Computatonal Intelligence Research. ISSN 0973-1873 Vol.1, No.1 (2005), pp. 68–76

[5] Liang Goh, Qun Song, and Nikola Kasabov,"A Novel Feature Selection Method to Improve Classification of Gene Expression Data", Conferences in Research and Practice in Information Technology, Vol. 29.

[6] PradiptaMaji and Sankar K. Pal, "Fuzzy–rough sets for information measures and Selection of relevant genes from microarray data", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 40, no. 3, June 2010

[7] C.Velayutham, K.Thangavel, "Unsupervised Feature Selection Using Rough Set". Proceeding on International Conference, Emerging Trends in Computing(ICETC-2011), 17-18 Mar 2011.

[8] K.Thangavel, P. Jaganathan, A. Pethalakshmi, M.Karnan,"Effective Classification with Improved Quick Reduct For Medical Database Using Rough System", BIME Journal, Volume (05), Issue (1), Dec., 2005.

[9] K.Thangavel, A. Pethalakshmi, "Feature Selection for Medical Database Using Rough System", AIML Journal, Volume (6), Issue (1), January, 2006

[10] QiangShen, Alexios Chouchoulas, "A Rough Fuzzy Approach For Generating Classification Rules",ww.elsevier.com/locate/patcog, Pattern Recognition 35 (2002) 2425 – 2438

[11] Parvesh Kumar, SiriKrishanWasan,"Comparative Analysis of k-mean Based Algorithms". IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.4, April 2010.

[12] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm". Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[13] AshaGowdaKaregowda, A.S. Manjunath, M.A. Jayaram," Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes". International Journal on Soft Computing (IJSC), Vol.2, No.2, May 2011.

[14] Ping Chang and Jeng-Shong Shih," The Application of Back Propagation Neural Network of Multi-channel

Piezoelectric Quartz Crystal Sensor for Mixed Organic Vapours".Tamkang Journal of Science and Engineering, Vol. 5, No. 4, pp. 209-217 (2002).

[15] Binu Thomas, Raju G., and SonamWangmo, "A Modified Fuzzy C-Means Algorithm for Natural Data Exploration". World Academy of Science,Engineering and Technology 49 2009.

[16] SellappanPalaniappan, Tan Kim Hong, "Discretization of Continuous Valued Dimensions in OLAP Data Cubes".IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

T. Chandrasekhar was born in 1980 at Karur, Tamilnadu, India. He is received the Master of Science in information technology and management in 2003 and his M.Phil (Computer Science) Degree in 2004, from Bharathidasan University, Trichy, India.

Currently he is working as Guest lecturer, Department of Computer Science, Periyar University, Salem, Tamilnadu, India. He is pursuing his Ph.D., from Bharathiar University in Computer Science under the guidance of Dr. K.Thangavel. His area of interests includes Medical Data Mining, Rough Set and Bioinformatics.

K.Thangavel was born in 1964 at Namakkal, Tamilnadu, India. He received his Master of Science from the Department of Mathematics, Bharathidasan University in 1986, and Master of Computer Applications Degree from Madurai Kamaraj University, India in 2001. He obtained his Ph.D. Degree from the Department of Mathematics, Gandhigram Rural Institute-Deemed University, Gandhigram, India in 1999. Currently he is working as Professor and Head, Department of Computer Science, Periyar University, Salem. He is a recipient of Tamilnadu Scientist award for the year 2009. His area of interests includes Medical Image Processing, Artificial Intelligence, Neural Network, Fuzzy logic, Data Mining and Rough Set.

E.N. Sathishkumar was born in 1986 at Namakkal, Tamilnadu, India. He received his Master of Science in Information Technology from Anna University, Coimbatore in 2009. He obtained his M.Phil (Computer Science) Degree from Periyar University, Salem, India in 2011.

His area of interests includes Data Mining, Rough Set and Neural Network.