Simultaneous Clustering and Feature Selection Method for Gene Expression Data

T. Chandrasekhar, K. Thangavel, E. N. Sathishkumar

***Abstract*—**Microarrays are made it possible to simultaneously monitor the expression profiles of thousands of genes under various experimental conditions. It is used to identify the co-expressed genes in specific cells or tissues that are actively used to make proteins. This method is used to analysis the gene expression, an important task in bioinformatics research. Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes, biologically relevant groupings of genes and samples. In this work K-Means algorithms has been applied for clustering of Gene Expression Data. Further, rough set based Quick reduct algorithm has been applied for each cluster in order to select the most similar genes having high correlation. Then the ACV measure is used to evaluate the refined clusters and classification is used to evaluate the proposed method. They could identify compact clusters with feature selection method used to genes are selected.

***Keywords*—**Clustering, Feature selection, Gene expression data, Quick reduct.

I. Introduction

DISCRIMINANT analysis is now widely used in bioinformatics, such as distinguishing cancer tissues from normal tissues or one cancer subtype vs. another [1]. A critical issue in discriminant analysis is feature selection: instead of using all available variables (features or attributes) in the data, one selectively chooses a subset of features to be used in the discriminant system. There are a number of advantages of feature selection: (1) dimension reduction to reduce the computational cost; (2) reduction of noise to improve the classification accuracy; (3) more interpretable features or characteristics that can help identify and monitor the target diseases or function types. These advantages are typified in DNA microarray gene expression profiles. Of the tens of thousands of genes in experiments, only a smaller number of them show strong correlation with the targeted phenotypes [1].

Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes. DNA microarrays are emerged as the leading technology to measure gene expression levels primarily, because of their high throughput. Results from these experiments are usually presented in the form of a data matrix in which rows represent genes and columns represent conditions or samples [8]. Each entry in the matrix is a measure of the expression level of a particular gene under a specific condition. Analysis of these data sets reveals genes of unknown functions and the discovery of functional relationships between genes [15]. Co-expressed genes can be grouped into clusters based on their expression patterns of gene based clustering and Sample based clustering. In gene based clustering, the genes are treated as the objects, while the samples are the features. In sample based clustering, the samples can be partitioned into homogeneous groups where the genes are regarded as features and the samples as objects [14].

In this paper, gene or features will be select from a group, such that the gene in a group will be similar. Gene expression data set will divide k number of groups using clustering techniques. Clustering is a widely used technique for analysis of gene expression data. Most clustering methods group genes based on the distances, while few methods group according to the similarities of the distributions of the gene expression levels. Clustering is the process of finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns.

Rough sets have been used as feature selection methods. 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 reduces while obtaining high quality classifiers based on the selected features. Here we studied a feature selection method based on rough set theory for reducing genes from *k* number of similar object gene groups.

This paper is organized as follows. Section II presents and overview of Gene based clustering and pre processing Technique. Section III describes the Rough set algorithms. Section IV describes performance of Experimental analysis and discussion. Section V describes WEKA classification and Section VI presents conclusion and future work.

II. Gene Based Clustering Techniques

A. K- Means Clustering

The main objective in cluster analysis is to group objects that are similar in one cluster and separate objects that are

T.Chandrasekhar is with the Department of Computer Science, Periyar University, Salem, Tamilnadu-636 011, India (phone: +91-9942925467; e-mail: ch_ansekh80@rediffmail.com).

Dr. K. Thangavel is with the Computer Science Department, Periyar University, Salem, Tamilnadu-636 011, India (e-mail: drktvelu@yahoo.com).

E. N. Sathishkumar is with the Computer Science Department, Periyar University, Salem, Tamilnadu-636 011 (e-mail: en.sathishkumar@yahoo.in).

dissimilar by assigning them to different clusters. One of the most popular clustering methods is K-Means clustering algorithm [1], [3], [5]. It classifies object to a pre-defined number of clusters, which is given by the user (assume K clusters). The idea is to choose random cluster centres, one for each cluster. These centres are preferred to be as far as possible from each other. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids [6]. The Euclidean distance between two multi-dimensional data points X = (x1, x2, x3, ..., xm) and Y = (y1, y2, y3, ..., ym) is described as follows:

$$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2} \quad (1)$$

The K-Means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described below.

Algorithm 1: K-Means clustering algorithm [9]

```
Require: D = {d1, d2, d3, ..., dn } // Set of n data points.
         K - Number of desired clusters
Ensure:  A set of K clusters.
Steps:
1. Arbitrarily choose k data points from D as initial centroids;
2. Repeat
      Assign each point di to the cluster which has the closest centroid;
      Calculate the new mean for each cluster;
   Until convergence criteria is met.
```

Though the K-Means algorithm is simple, it has some drawbacks of quality of the final clustering, since it highly depends on the arbitrary selection of the initial centroids [2].

B. K-Means Discretization

Many data mining techniques often require that the attributes of the data sets are discrete. Given that most of the experimental data are continuous, not discrete, the discretization of the continuous attributes is an important issue. The goal of discretization is to reduce the number of possible values a continuous attribute takes by partitioning them into a number of intervals. K-means discretization method is used in this paper, in gene expression data set each gene attribute are clustered with K-means, replaces the attribute values with the cluster membership labels. These labels will be act as discrete values for gene expression data set.

III. Rough Set Theory

Rough set theory 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 [10].

Definition: Let *U* be a universe of discourse, *X*⊆*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*∈*U/R | Y*⊆*X}*.The partition of *U*, generated by *IND* (*P*) is denoted *U*/*P*. If (*x*, *y*) ∈*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*⊆*U,* the P-lower approximation $\underline{PX}$ and P-upper approximation *PX* of set *X* can be defined as:

$$\underline{P}X= \{ x\in U \mid [x]p \subseteq X \} \quad (2)$$

$$\overline{P}X= \{ x\in U \mid [x]p \cap X \neq \varphi \} \quad (3)$$

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

$$\mathrm{POS}_P(Q) = \bigcup_{X\in U/Q} \underline{PX} \quad (4)$$

$$\mathrm{NEG}_P(Q) = U - \bigcup_{X\in U/Q} \overline{PX} \quad (5)$$

$$\mathrm{BND}_P(Q) = \bigcup_{X\in U/Q} \overline{PX} - \bigcup_{X\in U/Q} \underline{PX} \quad (6)$$

An important issue in data analysis is discovering dependen cies between attributes dependency can be defined in the following way. For *P*, *Q* ⊆*A*, *P* depends totally on *Q*, if and only if *IND* (*P*) ⊆*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*≤ *k* ≤*1* denoted *P*⇒*k Q,* if

$$k = \gamma_P(Q) = \frac{|\mathrm{POS}_P(Q)|}{|U|} \quad (7)$$

If *k =1*, *Q* depends totally on *P*, if *0*≤ *k* ≤*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* ***U*** *D*. The degree of dependency between condition and decision attributes, ***γc***(*D*), is called the quality of approximation of classification, induced by the set of decision attributes [11], [12].

A. Quick Reduct Algorithm [16]

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 γR(D)=γC(D). A given dataset may have many attribute reduct sets, so the setR of all reducts is defined as:

$$\text{Rall} = \{X \mid X \subseteq C, \gamma X(D) = \gamma C(D); \gamma X-\{a\}(D) \neq \gamma X(D), \forall a \in X\}. \tag{8}$$

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:

$$\text{Rmin} = \{X \mid X \in \text{Rall}, \forall Y \in \text{Rall}, |X| \leq |Y|\}. \tag{9}$$

The Quick Reduct algorithm shown below [6], [7], 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 (γc(D)). Note that this type of search does not guarantee a minimal subset and may only discover a local minimum.

Algorithm 2: Quick Reduct(C, D)

```
C, the set of all conditional features;
D, the set of decision features.
(a) R ← {}
(b) Do
I T ← R
(d)   ∀ x ∈(C-R)
(e)ifγR∪{x} (D) >Γt (D)
          Where Γr(D)=card(POSR(D)) / card(U)
(f)         T ← R∪{x}
(g) R ← T
(h) untilγR(D) = = Γc(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 [6], [7].

B. Average Correlation Value (ACV)

It is used to evaluate the homogeneity of a cluster. Matrix A= (A_{ij}) has the ACV which is defined by the following function,

$$\text{ACV (A)} = max\left\{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}|Crow_{ij}|-m}{m^2-m}, \frac{\sum_{p=1}^{n}\sum_{q=1}^{n}|Ccol_{pq}|-n}{n^2-n}\right\} \tag{10}$$

where $Crow_{ij}$ _ is the correlation coefficient between rows i and j and $Ccol_{pq}$ is the correlation coefficient between columns p and q, ACV approaching 1 denote a significant cluster. Such technique may be found in [13].

IV. Experimental Results

A. Data Sets

We use three datasets: leukemia, lung cancer and prostate cancer which are available in the website: http://datam.i2r.a-star.edu.sg/datasets/krbd/, [18]. The gene number and class contained in three datasets are listed in Table I.

TABLE I
Summary of Gene Expression Datasets

Dataset	#Gene	Class	Dataset
Leukemia	7129	ALL/AML	34 (20/14)
Lung	7129	Tumor/Normal	96 (86/10)
Prostate	12600	Tumor/Normal	21 (8/13)

B. Cluster Analysis

Given Gene expression data set will divide K (K =5) number of groups using clustering techniques. Most clustering methods group genes based on the distances, while few methods group according to the similarities of the distributions of the gene expression levels. K-Means clustering is used to cluster the similar characteristics of genes. After K-Means grouped genes are listed in Table II.

TABLE II
Similar Genes after K-Means

Data	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Leukemia	39	873	6025	74	118
Lung	6011	188	29	101	800
Prostate	1240	915	10318	7	120

C. Feature Selection

Features or genes will be selected from a similar gene cluster that clusters shown in Table II. Rough sets have been used as feature selection methods. 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 in cluster K (K=1 to 5), C the condition attribute set and D the decision attribute set. We can represent every gene expression data with the decision table like Table III.

TABLE III
Microarray Data Decision Table

Samples	Cluster K (Condition attributes)				Decision attributes
	Gene1	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 in cluster K. 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[18]. Before applying feature selection algorithm all the conditional attributes (samples) are discretized using K-Means discretization. Table IV shows the selected genes from particular cluster by applying Quick Reduct Algorithm.

TABLE IV
SELECTED GENES BY QUICK REDUCT ALGORITHM

Cluster	Leukemia	Lung	Prostate
Cluster 1	19,44,930,1841	6044	637,3588,10960
Cluster 2	220,6855	989,1513	6849,8311
Cluster 3	4,3252	131,686,1116,1432	149,7028
Cluster 4	1674,3452,4017	1569,1906,6905	604,750,7185,10837, 10922
Cluster 5	543,1962,3361	4817	48,1507,7581

ACV is evaluating the homogeneity of a cluster. ACV approaching 1 denote a significant cluster. ACV will have been fined for each cluster of gene expression data and selected genes. Tables V-VII show the value of ACV for gene expression data set.

TABLE V
ACV FOR LEUKEMIA CANCER

Cluster	All Genes	ACV	Selected Genes	ACV
Cluster 1	39	0.3130	19,44,930,1841	0.6011
Cluster 2	873	0.2430	220,6855	1
Cluster 3	6025	0.6457	4,3252	1
Cluster 4	74	0.2666	1674,3452,4017	0.6767
Cluster 5	118	0.2524	543,1962,3361	0.6266

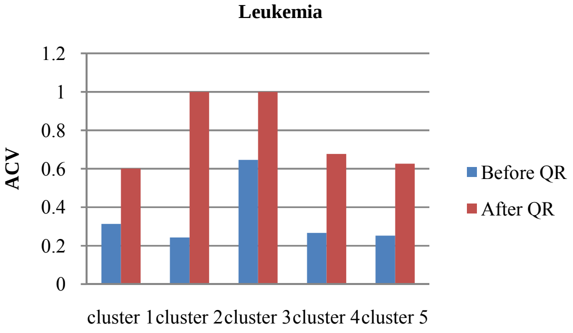

Fig. 1 ACV for Leukemia Cancer data set

Fig. 1 depicts the performance of the feature selected genes ACV values for Leukemia dataset. On the cluster1 and cluster2 genes exhibits highest ACV=1 and it shows best clusters for Leukemia data set.

TABLE VI
ACV FOR LUNG CANCER

Cluster	All Genes	ACV	Selected Genes	ACV
Cluster 1	6011	0.5729	6044	0.0104
Cluster 2	188	0.2067	989, 1513	1
Cluster 3	29	0.3834	131, 686, 1116, 1432	0.5181
Cluster 4	101	0.2877	1569, 1906,6905	0.6967
Cluster 5	800	0.2447	4817	0.0104

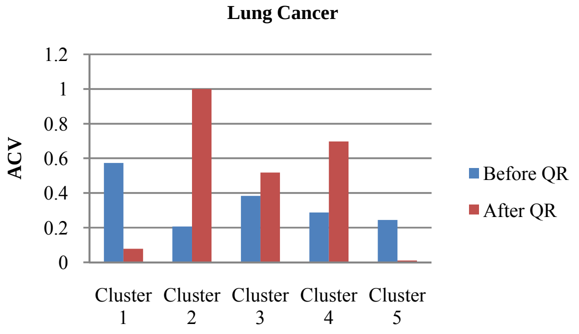

Fig. 2 ACV for Lung Cancer data set

Fig. 2 depicts the performance of the feature selected genes ACV values for Lung Cancer dataset. On the cluster2 genes exhibits highest ACV=1and it shows best cluster for Lung Cancer data set.

TABLE VII
ACV FOR PROSTATE CANCER

Cluster	All Genes	ACV	Selected Genes	ACV
Cluster 1	1240	0.8405	637, 3588, 10960	0.6199
Cluster 2	915	0.5186	6849, 8311	1
Cluster 3	10318	0.3834	149, 7028	0.0343
Cluster 4	7	0.6907	604,750,7185,10837,10922	0.7696
Cluster 5	120	0.4957	48, 1507, 7581	0.8611

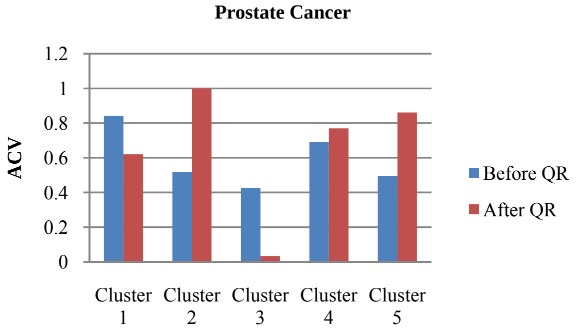

Fig. 3 ACV for Prostate Cancer data set

Fig. 3 depicts the performance of the feature selected genes ACV values for Prostate Cancer dataset. On the cluster 2 genes exhibits highest ACV=1 and it shows best clusters for Prostate Cancer data set.

V. WEKA CLASSIFICATION

The Waikato Environment for Knowledge Analysis (WEKA) is a comprehensive suite of Java class libraries that implement many state-of-the-art machine learning and data mining algorithms. WEKA is freely available on the World-Wide Web and accompanies a new text on data mining [17] which documents and fully explains all the algorithms it contains. Applications written using the WEKA class libraries can be run on any computer with a Web browsing capability; this allows users to apply machine learning techniques to their own data regardless of computer platform. Tools are provided for pre-processing data, feeding it into a variety of learning schemes, and analyzing the resulting classifiers and their performance [4].

The primary learning methods in WEKA are “classifiers”, and they induce a rule set or decision tree that models the data. WEKA also includes algorithms for learning association rules and clustering data. The core package contains classes that are accessed from almost every other class in WEKA. The most important classes in it are Attribute, Instance, and Instances. An object of class Attribute represents an attribute—it contains the attribute’s name, its type, and, in case of a nominal attribute, it’s possible values. An object of class Instance contains the attribute values of a particular instance; and an object of class Instances contains an ordered set of instances—in other words, a dataset [4].

In this paper we have taken the classifiers such as Naive Bayes, KStar, Decision Table, ZeroR, J48 and Simple Cart. Tables VIII-X shows the classification accuracy of original data set genes, selected genes from all clusters and selected genes from ACV = 1 clusters.

TABLE VIII
CLASSIFICATION ACCURACY FOR LUKEMIA CANCER

Classification	Original	QR selected gene	ACV = 1 cluster Gene
Naive	94.1176	97.0588	97.0588
K Star	58.8235	94.1176	97.0588
Decision Table	88.2353	94.1176	94.1176
ZeroR	58.8235	58.8235	58.8235
J48	91.1765	97.0588	97.0588
SimpleCart	88.2353	94.1176	94.1176

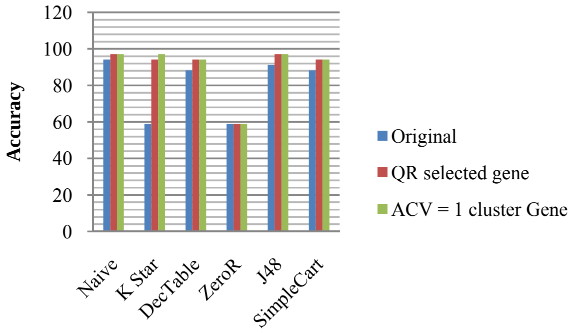

Fig. 4 Classification Accuracy for Lukemia Cancer

TABLE IX
CLASSIFICATION ACCURACY FOR LUNG CANCER

Classification	Original	QR selected gene	ACV = 1 cluster Gene
Naive	96.875	98.9583	97.9167
K Star	89.5833	98.9583	97.9167
Decision Table	96.875	97.9167	96.875
ZeroR	89.5833	89.5833	89.5833
J48	95.9583	97.9167	97.9167
SimpleCart	96.875	97.9167	96.875

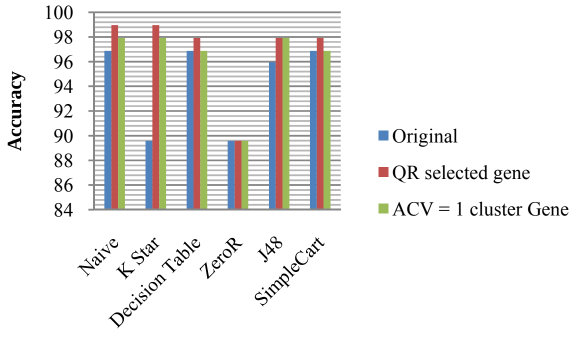

Fig. 5 Classification Accuracy for Lung Cancer

TABLE X
CLASSIFICATION ACCURACY FOR PROSTATE CANCER

Classification	Original	QR selected gene	ACV = 1 cluster Gene
Naive	71.4286	85.7143	71.4286
K Star	38.0952	52.381	57.1429
Decision Table	52.381	57.1429	57.1429
ZeroR	61.9048	61.9048	61.9048
J48	33.3333	57.1429	57.1429
SimpleCart	57.1429	42.8571	61.9048

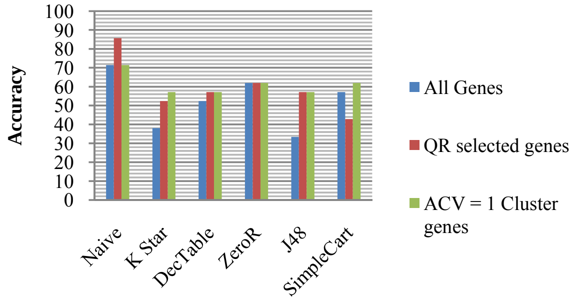

Fig. 6 Classification Accuracy for Prostate Cancer

VI. CONCLUSION

In this work, simultaneous clustering and feature selection method for gene selections is studied and apply to avoid too many redundant or missing values in Microarray gene expression data. In this gene selection method is based on Feature selection using Rough set methods. K-means clustering, Rough Quick Reduct and Average Correlation Value methods are studied and implemented successfully for gene selection. The proposed work gives sparse and interpretable classification accuracy, compared to other gene selection method on gene expression data set. The classification accuracy of selected genes has been done using WEKA classifier. In compare with other gene selection methods, our method is simple, effective and robust.

REFERENCES

[1] Chen Zhang and Shixiong Xia, 2009 “ K-Means Clustering Algorithm with Improved Initial center,” in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792.

[2] Chris Ding and Hanchuna Peng, “Minimum Redundancy Feature Selection from Microarray Gene Expression Data”, proceedings of the International Bioinformatic Conference, Date on 11-14, August – 2003.

[3] Dongxiao Zhu, Alfred O Hero, Hong Cheng, Ritu Khanna and Anand Swaroop, “Network constrained clustering for gene microarray Data”, doi:10.1093 /bioinformatics / bti 655, Vol. 21 no. 21, pp. 4014 – 4020, 2005.

[4] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, “Weka: Practical Machine Learning Tools and Techniques with Java Implementations”.

[5] Kohei Arai and Ali Ridho Barakbah, “ Hierarchical K-Means: an algorithm for centroids initialization for K-Means”, Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 25-31, 2007.

[6] 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.

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

[8] K.R De and A. Bhattacharya, “Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles,” bioinformatics, Vol. 24, pp.1359- 1366, 2008.

[9] Madhu Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa, 2010 “Enhancing K-Means Clustering Algorithm with Improved Initial Center”, Madhu Yedla et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2), pp121-125.

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

[11] Pradipta Maji 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.

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

[13] R.Rathipriya, Dr. K.Thangavel and J.Bagyamani, “Evolutionary Biclustering of Clickstream Data”, International Journal of Computer Science Issues, Vol. 8, Issue 3, No 1, May 2011. ISSN (Online): 1694-0814.

[14] Sauravjoyti Sarmah and Dhruba K. Bhattacharyya. “An Effective Technique for Clustering Incremental Gene Expression data”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May 2010.

[15] Sunnyvale, Schena M. “Microarray biochip technology”. CA: Eaton Publishing; 2000.

[16] T.Chandrasekhar, K.Thangavel and E.N.Sathishkumar, “Verdict Accuracy of Quick Reduct Algorithm using Clustering and Classification Techniques for Gene Expression Data”, International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012. ISSN (Online): 1694-0814.

[17] Witten, I. H., and Frank E. (1999) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco.

[18] Xiaosheng Wang, Osamu Gotoh, “Cancer Classification Using Single Genes”, pp 179-188.

T. Chandrasekhar was born in Karur at 1980, 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. He is pursuing his Ph.D in Bharathiar University in Computer Science under the guidance of Dr. K. Thangavel.

Currently he is working as Guest lecturer, Department of Computer Science, Periyar University, Salem, Tamilnadu, India. His area of interests includes Medical Data Mining, Rough Set and Bioinformatics.

Dr. K. **Thangavel** was born in Namakkal at 1964, 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, Bioinformatics, Artificial Intelligence, Neural Network, Fuzzy logic, Data Mining and Rough Set.

E. N. Sathishkumar was born in Namakkal at 1986, Tamilnadu, India. He received his Master of Science in Information Technology from Anna University, Coimbatore in 2009. He obtained his Master of Philosophy form the Department of Computer Science, Periyar University, Salem, India in 2011. He is pursuing his Ph.D in Computer Science at Periyar University under the guidance of Dr. K. Thangavel.

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