Journal of Computer Science 7 (7): 986-990, 2011
ISSN 1549-3636

A Rough Set based Gene Expression Clustering Algorithm

J. Jeba Emilyn and K. Ramar
Department of IT, Sona College of Technology, Salem,
SriVidhya College of Engineering and Technology, Virudhunagar, Tamilnadu, India

Abstract: Problem statement: Microarray technology helps in monitoring the expression levels of thousands of genes across collections of related samples. **Approach:** The main goal in the analysis of large and heterogeneous gene expression datasets was to identify groups of genes that get expressed in a set of experimental conditions. **Results:** Several clustering techniques have been proposed for identifying gene signatures and to understand their role and many of them have been applied to gene expression data, but with partial success. The main aim of this work was to develop a clustering algorithm that would successfully indentify gene patterns. The proposed novel clustering technique (RCGED) provides an efficient way of finding the hidden and unique gene expression patterns. It overcomes the restriction of one object being placed in only one cluster. **Conclusion/Recommendations:** The proposed algorithm is termed intelligent because it automatically determines the optimum number of clusters. The proposed algorithm was experimented with colon cancer dataset and the results were compared with Rough Fuzzy K Means algorithm.

Key words: Microarray technology, clustering algorithm, gene expression data, fuzzy membership, rough clustering, clustering technique, knowledge discovery, data mining, attribute clustering

INTRODUCTION

Biological data are being produced at a phenomenal rate. It is astonishing to see the repositories grow in an extraordinary way. On average, these databases double in size every 10 month. The enormous quantity and variety of information that is being produced cannot be handled that efficiently with the puny human brains. It would be easier if this data can be divided into a more comprehensible level by subdividing the genes into smaller categories and then analyze them. This is where clustering comes in.

Cluster Analysis plays a major role in Knowledge Discovery and Data mining (KDDM). The process of clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. It ultimately increases intra class similarity but decreases interclass similarity. Clustering of gene expression data helps to understand gene functions and gene regulations and assists in pattern recognition in gene expression profiles. Genes with similar expression patterns can be grouped together which would help us in further understanding the functionalities of unknown and abnormal patterns.

Related work: Hybrid fuzzy c-means clustering technique proposed by Valarmathie *et al*. (2009), combines Fuzzy C-Means with Expectation Maximization algorithm to determine the precise number of clusters and to interpret them efficiently. Noureen and Qadir (2009) have proposed a simple and efficient biclustering algorithm (BiSim) which proves to be very simple when compared the Bimax algorithm. It reduces the complexity and extra computation when compared to Bimax.

Thilagamani and Shanthi (2010) have done a survey stating that clustering algorithms designed based on rough sets are neither too restrictive as the Crisp clustering nor too descriptive as that of fuzzy clustering. Pavan *et al*. (2010) have proposed a Single Pass Seed Selection (SPSS) algorithm which is an extension of K-means++ which works well with high dimensional data sets. K-Biclusters Clustering (KBC Algorithm), proposed by Tsai and Chiu (2010), minimizes the dissimilarities between genes and bicluster centers. Additionally it tries to minimize the residue within the clusters and to involve as many conditions as possible. Venkatesh and Thangaraj (2008) have proposed a SOM based clustering and artificial intelligence technique to analyse patterns of soil distributed across a geographical area. Maji (2011) proposed a new clustering algorithm, termed as Fuzzy-Rough Supervised Attribute Clustering (FRSAC), to find groups of coregulated genes whose collective expression

Corresponding Author: J. Jeba Emilyn, Department of IT, Sona College of Technology, Salem. Tamilnadu, India

is strongly associated with sample categories. A new quantitative measure is introduced based on fuzzy-rough sets that incorporates the information of sample categories to measure the similarity among genes whereby redundancy among the genes are removed.

MATERIALS AND METHODS

Research background:

Rough set-definition: Rough set theory introduced by Pawlak (1982) deals with uncertainty and vagueness. It is a new mathematical approach to imperfect knowledge. Rough sets can be considered as sets with fuzzy boundaries i.e., sets that cannot be precisely characterized using the available set of attributes. Rough set theory has become popular among scientists around the world due to its fundamental importance in the field of artificial intelligence and cognitive sciences. Similar to fuzzy set theory it is not an alternative to classical set theory but it is embedded in it.

Suppose we are given a set of objects U called the universe and an indiscernibility relation R as U x U, representing our lack of knowledge about elements of U. For the sake of simplicity we assume that R is an equivalence relation. Let X be a subset of U. We want to characterize the set X with respect to R:

- The lower approximation of a set X with respect to R is the set of all objects, which can be for certain classified as X with respect to R (are certainly X with respect to R)
- The upper approximation of a set X with respect to R is the set of all objects which can be possibly classified as X with respect to R (are possibly X in view of R)
- The boundary region of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not-X with respect to R

Now we are ready to give the definition of rough sets:

- Set X is crisp (exact with respect to R), if the boundary region of X is empty
- Set X is rough (inexact with respect to R), if the boundary region of X is nonempty

Formal definitions of approximations are as follows:

R-lower approximation of X:

$$R_*(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}$$

R-upper approximation of X:

$$R^*(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \varnothing\}$$

Clustering gene expression data: Clustering is one of the first steps in gene expression analysis. One of the important characteristics of gene expression data is that it is meaningful to cluster both genes and samples. During cluster analysis, genes are clustered based on similarity. Proximity measurement measures the similarity (or distance) between two data objects. The proximity between two objects is measured by a proximity function of their corresponding vectors.

Euclidean distance is one of the most commonly used methods to measure the distance between two data objects. The main drawback is that Euclidean distance does not score well for scaled patterns or profiles of genes. The Manhattan distance is closely related to Euclidean distance. This finds out the sum of distances along each dimension while Euclidean distance finds the length of the shortest path between two points. Another measure is Pearson's correlation coefficient, which measures the similarity between the shapes of two expression patterns (profiles). Pearson's correlation coefficient is widely used and has proved to be efficient in many clustering algorithm for gene expression data (Jiang *et al*., 2004). The main drawback of this measures is that it is not more robust in handling outliers. In order to address the problems faced with pearson's correlation coefficient another measure named Spearsman correlation coefficient was introduced. It is more robust against outliers when compared to Pearson's correlation coefficient. A survey on Rough set based clustering and its preference over conventional methods was initially done and analyzed.

Rough fuzzy K means algorithm: K means is one of the traditional algorithms available for the clustering. However this algorithm is crisp as it allows an object to be placed exactly in only one cluster. To overcome the disadvantages of crisp clustering fuzzy based clustering was introduced. The distribution of member is fuzzy based methods can be improved by rough clustering. Based on the lower and upper approximations of rough set, the rough fuzzy k-means clustering algorithm makes the distribution of membership function become more reasonable (Shi *et al*., 2009).

The frame work of RFKM algorithm: Specific steps of the RFKM clustering algorithm are given as follows:

Step1: Determine the class number k (2<=k<=n), parameter m, initial matrix of member function, the upper approximate limit Ai of class, an appropriate number $\varepsilon > 0$ and s = 0.

Step 2: We can calculate centroids with the formula given below:

$$C_i = \sum_{j=1}^{n} U_{ij}^{m} X_j / \sum_{j=1}^{n} U_{ij}^{m}$$

Step3: If $X_j \notin$ the upper approximation, then $U_{ij} = 0$. Otherwise, update U_{ij} as shown below

$$U_{ij} = \frac{1}{\sum_{l=1}^{k}, x_j \in Rwi(\frac{d_{ij}^2}{d_{ij}^2})\frac{1}{m-1}}$$

Step4: If $\left\|U^{(S)} - U^{S+1}\right\| < \varepsilon$ then stop, else s = s+1, iterate to step 2.

Experimental results: The RFKM algorithm was experimented with yeast expression data set. The data set is 834 X 7 matrix. A total of 834 genes were clustered based on 7 experimental conditions into different no of clusters. Since RFKM requires the no of clusters to be given as input, 8 different clusters were generated. The result in Fig. 1 shows the membership matrix of the genes belonging to different clusters. A total of 8 clusters were generated with each graph representing their membership values of a particular cluster. The algorithm was implemented in matlab and was also experimented for variety of data sets.

The Proposed algorithm (RCGED): Our proposed new algorithm, Rough Clustering of Gene Expression Data (RCGED), clusters genes based on rough set theory. The main advantage of our method is that it does not restrict a gene to one cluster. Genes can get expressed in two are more clusters ie Overlapping of genes are possible. It also finds the lower and upper approximation of the clusters. Our algorithm is designed to be intelligent in the sense that it itself detects the optimum number of clusters. Our algorithm uses a similarity measure based on correlation coefficient.

The Frame work of the proposed algorithm:

Algorithm: RCGED
Input: Gene expression matrix
Output: No of clusters, membership matrix, similarity matrix.
Step1: For each gene g_i, compute the membership subset
Step2: Compute the similarity or distance matrix f_{sim}

K=1;
For each gene g_i
K++;
Ith gene is placed in cluster k;
For each gene j<>i
Compute the similarity of i^{th} gene with j^{th} gene $f_{sim}(i,j)$ using correlation coefficient metric;
If $f_{sim}(i,j)$ > threshold α place j in cluster k
End;
End;

Step3: Calculate mean m_i for the k clusters;
Step4: Assign each data object P to the lower approximation or the upper approximation by finding the difference in its distance from the cluster centroid pairs m_i and m_j:

$$[\, đ(P\text{-}m_i) - đ(P\text{-}m_j) \,]$$

Step5: If the distance is less than some threshold ∞, X_P is in the upper approximation and X_p is not in the lower approximation else X_p is in the lower approximation.
Step 6: Compute new mean for each cluster k and iterate until there are no more assignments.

The algorithm generates the membership matrix based on the rough set theory. Based on the similarity between the genes, the algorithm proceeds on to find out the possible number of clusters and the distance matrix for which it uses correlation coefficient as the metric. Genes that are more similar are put in the same cluster. Each object is either assigned to the upper or the lower approximation of each cluster. Then we also dynamically calculate the membership matrices for both upper and lower approximations as shown in the algorithm. The mean of each cluster(lower and upper) is then taken as the centroid (pair) of that cluster. The process iterates and dynamically updates the membership matrices and the similarity matrix until there is no more change in the cluster centroid.

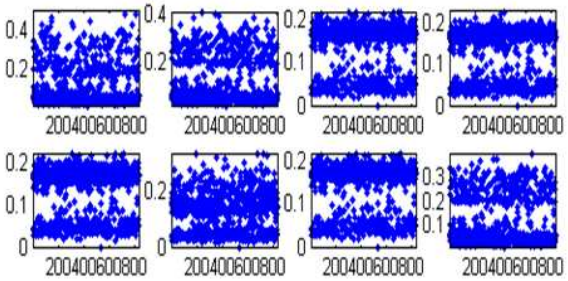

Fig. 1: Rough fuzzy clustering of yeast data set

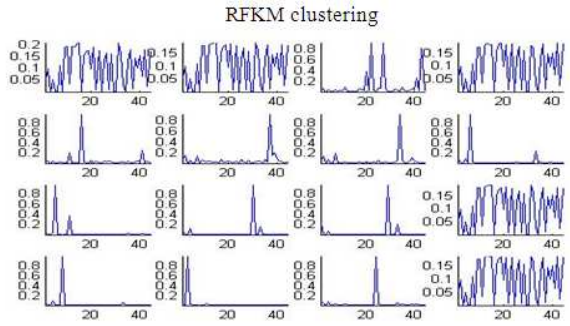

Fig. 2: Colon cancer data clusters generated using RFKM

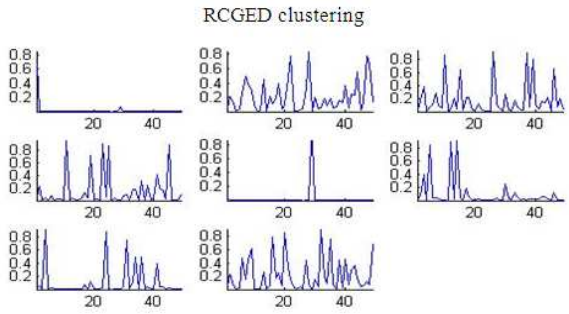

Fig. 3: Colon cancer data clusters generated using RCGED

RESULTS

Experimental results: The RFKM algorithm requires the user to specify the no of clusters prior to clustering. This does not suit all problems as the no of clusters specified by the user might be too small or too large. The result of RFKM on colon cancer data set is shown in Fig. 2. The colon cancer data set contains expression levels of 2000 genes taken in 62 different samples out of which 50 genes where chosen across all 62 samples. The proposed RCGED algorithm is designed to be intelligent. Unlike the RFKM, it finds out the optimum no of clusters on its own and proceeds with the clustering. The algorithm uses a method to tune the threshold and the relative importance of the upper and lower approximation of the rough sets is used in modeling the clusters. The RCGED algorithm was also experimented with colon cancer data set. The result is shown in Fig. 3.

Comparison of RCGED with RFKM: The effectiveness of the algorithm is shown as a comparative study between the performance of Rough Fuzzy K-Means and RCGED. Cluster validation of the clusters generated by these two algorithms is done. The procedure of qualitative evaluation of the clusters is referred to as cluster validation. Validation index is a real value that determines the quality of the clusters. Our algorithm is evaluated using Davis-Bouldin's measure as the validation index. This index is a function of the ratio of the sum of within-cluster scatter and between-cluster separation. Table 1 gives the sample results and the comparative study between RFKM and RCGED. The uncertainty that prevails in the overlapping clusters is eliminated in our proposed algorithm. We can observe that RCGED algorithm has minimum value for DB index when compared to RFKM.

In all rough clustering algorithms, the number of objects in the boundary region depends on the value of the threshold α. It has been noted for our algorithm that the number of genes in the boundary region decreases as the value of α becomes <0.1. When the threshold value becomes larger, the number of genes in the boundary region also increases.

DISCUSSION

There are dozens of clustering algorithms that have been applied to gene expression data. But there is no single-best solution or a fit-all solution to clustering because there is no clear criteria and definition of what and how a cluster is to be (Jain and Dubes, 1988). Clusters can be of any shape and size in the multidimensional pattern space. In Jain and Dubes words, "Each clustering criterion imposes certain structure on the data and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered".

Table 1: Performance comparison between RFKM and RCGED

Number of genes	Clustering algorithm	No of clusters (given as input)	No of clusters generated	Davies-bouldin index
1000	RFKM	5	-	1.148
	RCGED	-	8	1.9155
2000	RFKM	7	-	3.3312
	RCGED	-	15	3.1788
3000	RFKM	8	-	4.1495
	RCGED	-	18	3.2142
4000	RFKM	10	-	3.4999
	RCGED	-	19	3.2264
5000	RFKM	12	-	4.0122
	RCGED	-	24	3.2515

CONCLUSION

There are dozens of clustering algorithms that have been applied to gene expression data. But there is no single-best solution or a fit-all solution to clustering. In this study, we have proposed an intelligent clustering algorithm that is based on the frame work of rough sets. A more general rough fuzzy k means algorithm was implemented and experimented with different gene expression data sets. The proposed algorithm RCGED was also implemented and experimented with colon cancer gene expression datasets. A comparison of the algorithms and their results were studied. The importance of upper and lower approximations of the rough clusters is optimized using DB index value. This algorithm seems to prove better than the other rough set based clustering algorithms. As an extension of the current research work, a toolkit that integrates and visualizes the results of a few rough clustering algorithms for clustering gene expression data is being developed.

REFERENCES

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice Hall, New Jersey, USA., ISBN: 13: 978-0130222787, pp: 304.

Jiang, D., C. Tang and A. Zhang, 2004. Cluster analysis for gene expression data: A survey. IEEE Trans. Knowledge Data Eng., 16: 1370-1386. DOI: 10.1109/TKDE.2004.68

Maji, P., 2011. Fuzzy rough supervised attribute clustering algorithm and classification of microarray data. IEEE Trans. Syst., Man Cybern.-Part B: Cybern., 41: 222-233. DOI: 10.1109/TSMCB.2010.2050684

Noureen, N. and M.A. Qadir, 2009. BiSim: A simple and efficient biclustering algorithm. Proceedings of the International Conference of Soft Computing and Pattern Recognition, Dec. 4-7, Socpar, Malacca, pp: 1-6. DOI: 10.1109/SoCPaR.2009.14

Pavan, K.K., A.A. Rao, A.V.D. Rao and G.R. Sridhar, 2010. Single pass seed selection algorithm for k-means. J. Comput. Sci., 6: 60-66. DOI: 10.3844/jcssp.2010.60.66

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inform. Sci., 2: 341-356. DOI: 10.1007/BF01001956

Shi, P., 2009. Clustering fuzzy web transactions with rough K means AST 09. Proceedings of the 2009 International e-Conference on Advanced Science and Technology, Mar. 7-9, IEEE Computer Society Washington, DC, USA., pp: 48-51. DOI: 10.1109/AST.2009.23

Thilagamani, S. and N. Shanthi, 2010. Literature survey on enhancing cluster quality. Int. J. Comput. Sci. Eng.., 2: 1999-2002. http://www.enggjournals.com/ijcse/doc/IJCSE10-02-06-26.pdf

Tsai, C.Y. and C.C. Chiu, 2010. A novel microarray biclustering algorithm. World Acad. Sci., Eng. Technol., 65: 256-262. http://www.waset.org/journals/waset/v65/v65-39.pdf

Valarmathie, P., M.V. Srinath, T. Ravichandran and K. Dinakaran, 2009. Hybrid fuzzy c-means clustering technique for gene expression data. Int. J. Res. Rev. Applied Sci., 1: 2076-7366. http://www.arpapress.com/Volumes/Vol1/IJRRAS_1_04.pdf

Venkatesh, E.T. and D.P. Thangaraj, 2008. Self-organizing map and multi-layer perceptron neural network based data mining to envisage agriculture cultivation. J. Comput. Sci., 4: 494-502. DOI: 10.3844/jcssp.2008.494.502