Efficient Clustering for Gene Expression Data

Jacinth Salome J
Assistant Professor
Dept of Computer Science
L N Govt College, Ponneri-601204

R M Suresh, PhD.
Principal
Jerusalem College of Engineering
Pallikaranai, Chennai

ABSTRACT
In the past decade there have been advance in technologies, the amount of biological data such as DNA sequences and microarray data have been increased tremendously. To obtain knowledge from the data, explore relationships between genes, understanding severe diseases and development of drugs for patterns from the databases of large size and high dimensionality. Information retrieval and data mining are powerful tools to extract information from the databases and/or information repositories. The integrative cluster analysis of both clinical and gene expression data has shown to be an effective alternative to overcome the abovementioned problems. In this paper, we focus on how to improve the searching and the clustering performance in genomic data from commonly used clustering techniques. In the proposed gene clustering technique, firstly, the high dimensionality of the microarray gene data is reduced using LPP. The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. Secondly, through performance experiments on real data sets, the proposed method fuzzy C-means is shown to achieve higher efficiency, clustering quality and automation than other clustering method.

General Terms
Data Mining, information retrieval, Bio-informatics et al.

Keywords
Clustering, microarray, Locality Preserving Projection (LPP), Fuzzy C-Means (FCM), k-means.

1. INTRODUCTION
High-throughput techniques have become a primary approach to gathering biological data. [1] These data can be used to the actual clinical application of gene expression data analysis and guide development of drugs and other research. [2] Generally, the data mining tasks comprise Classification, Regression, Clustering, Rule generation, Discovering association rules, Summarization, Dependency modeling and Sequence analysis. [3] DNA microarray technology is also the field, which uses the data mining methods. Also, the molecular biologists face the challenges in discovering the essential knowledge from this kind of enormous volume of data. [4] However, the deluge of data contains an overwhelming amount of unknown information about the organism under study. Therefore, clustering is a common first step in the exploratory analysis of high-throughput biological data. [5] Although a number of clustering methods have been proposed, they incur problems such as Quality and Efficiency. In the aspect of efficiency, most clustering algorithms aim to produce the best clustering result based on the input parameters.

In this paper, we propose an efficient approach for information retrieval in DNA microarray technology. In the process of mining gene expressions under multi-conditions microarray experiments, gene clustering is relatively a tough task, because of the features of the data that have high dimensionality and small sample size. [6] A combination of the approaches is utilized continually in practice for clustering with microarray data. Such classification measures usually have the following stages: i) gene selection/dimension reduction, where a small amount of gene components are created from a vast number of genes; ii) clustering, where the samples are clustered into groups by applying standard models on the gene components. [7] Normally, microarray experiments create a large number of datasets with expression values for thousands of genes but still not more than a few dozens of samples, thus very accurate arrangement of tissue samples in such high dimensional problems is a complicated task. Also, there is a high redundancy in microarray data as well as several genes have irrelevant information for exact clustering of diseases or phenotypes.[8] Therefore, a robust clustering method is indispensable to retrieve the gene information from the microarray experimental data.

Microarray gene data clustering is challenging problem and researches are rising in this problem. Here we propose a work to group the microarray gene data with the aid of FCM. [9] Initially, the dimensionality reduced data is applied with the FCM for clustering. The proposed work is detailed in the upcoming sessions of this paper and the organization of the paper is as follows. Section 2 details the related researches and section 3 explains the proposed methodology with the appropriate equations and diagrams. The results are discussed in the section 4 with proper diagrams and tables. Section 5 concludes the paper.

2. RELATED WORK
Jian Wen [10] in his biomedical literatures on ontology is useful to improve the performance of information retrieval. The method of ontology-based solves synonym problems with a new frame for genomic information retrieval based on UMLS. In genomic information retrieval includes three processes: first, documents were indexed based UMLS, which means documents were represented by concepts, besides, the concept weight was re-calculated combined with similarity between concepts. Second, documents were clustered using fuzzy c-means method. At last cluster language model is utilized for information retrieval.

Yuen et al. [11] have improved the searching and the clustering performance in genomic sequence databases. A search based on index can efficiently search the homologous database sequences to a query sequence. It makes use of two novel hashing techniques to enhance the efficiency of indexing and retrieval. Besides, we measure the similarity of sequences instead of sequence alignment. Thus, our search

algorithm can run faster than the famous Blast algorithm. Our clustering algorithm is a hybrid of partitioning method and hierarchical method. It quickly clusters a group of nearest neighbors and finally merges the clusters.

Wai-Ho Au et al. [12] have proposed an attribute clustering method which was able to group genes based on their interdependence in order to mine meaningful patterns from the gene expression data. Their method grouped interdependent attributes into clusters by optimizing a criterion function obtained from an information measure that exhibited the interdependence between attributes. Meaningful clusters of genes were determined, by applying their algorithm to gene expression data. To analyze the performance of their approach, they applied it to two recognized gene expression data sets and compared their results with those acquired by other methods. Their experiments proved that their method was able to determine the meaningful clusters of genes.

3. THE PROPOSED GENE CLUSTERING TECHNIQUE BASED ON LPP AND FCM

The most commonly used computational method for analyzing microarray gene expression data is clustering. Based on a correlation measure between the row vectors, genes are partitioned into clusters, using clustering algorithm. The main problems associated with the traditional clustering algorithms are handling multidimensionality and scalability with rapid growth in size of data. The increase in size of data increases the computational complexities which have a devastating effect on the runtime and memory requirements for large applications.

The proposed technique is comprised of two stages, dimensionality reduction and FCM-based clustering. Let, the microarray gene expression data is $X_{jk}\ ;\ 0 \le j \le n_g\ ,\ 0 \le k \le n_s$, where, n_g represents the number of genes from which the data is taken and n_s represents the number of samples. The gene data is of higher dimension and so it is subjected to dimensionality reduction. In the dimensionality reduction, the high dimensional gene data X_{ij} is converted to low dimensional data. [13] The resultant low dimensional data is clustered using a well-trained FCM.

3.1 Dimensionality Reduction by LPP

Dimensionality reduction, one of the two stages of the proposed gene clustering technique is performed using LPP [14]. From the gene data of different classes Y_{ijk}, a concatenated matrix is obtained as given in the Eq. (1). In the concatenated matrix, the gene data of all the classes are combined and it is given as a single matrix. The matrix Y_{conc} is given as follows

$$Y_{conc_{jl}} = \sum_{i=0}^{n_c - 1} Y'_{ijl} \tag{1}$$

where,

$$Y'_{ijl} = \begin{cases} Y_{ijl} & ;\ if \quad l \in (i, n_s(i+1) - 1) \\ 0 & ;\ otherwise \end{cases} \tag{2}$$

The concatenated matrix Y_{conc} of dimension $n_g \times n'_s$; $n'_s = n_s.n_c$, $n'_s << cn_g$, which is highly dimensional and so the dimensionality of the matrix is reduced using LPP. The LPP is a linear dimensionality reduction algorithm that shares most of the properties of data representation of nonlinear techniques, namely, locally linear Embedding or Laplacian Eigenmaps. The LPP procedure for dimensionality reduction constitutes of three steps, namely, (1) generation of Distance matrix (2) determining adjacency matrix and (3) Calculating dimensionality reduced matrix.

3.1.1 Generation of Distance Matrix

For the concatenated matrix Y_{conc}, the distance matrix of size $n_g \times n_g$ is determined as follows

$$D_{xy} = \sqrt{\sum_{l=0}^{n'_s} \left(Y_{conc_{xl}} - Y_{conc_{yl}} \right)^2}\ ;\quad 0 \le x, y \le n_g \tag{3}$$

The determined distance matrix is based on the Euclidean distance calculated by considering each row of the Y_{conc} as a network node. The resultant D_{xy} is subjected to calculate adjacency matrix, which can be determined based on the relationship of an element with every neighbor elements.

3.1.2 Determination of Adjacency matrix

In the virtual network consisting of n_g nodes, the adjacency matrix is a $n_g \times n_g$ with binary entries representing if there is an edge between two nodes. Here, the adjacency matrix W is determined with the aid of the D_{xy} as follows

$$W_{xy} = \begin{cases} 1 & ;if \quad D_{xy} > 0 \\ 0 & ;otherwise \end{cases} \tag{4}$$

From the Eq. (4), it can be seen that the adjacency matrix W_{xy} is constituted of binary values depending upon the distance calculated in D_{xy}.

3.1.3 Calculation of Dimensionality Reduced matrix

From the adjacency matrix W, a diagonal matrix A is determined as follows

$$A_{xy} = \begin{cases} S_x ; \ if \ \ x = y \\ 0 \ \ ; \ otherwise \end{cases} \tag{5}$$

where,

$$S_x = \sum_{y=0}^{n_g - 1} W_{xy} \tag{6}$$

Based on the A, which is obtained from the Eq. (5), Z_1 and Z_2 are calculated as follows

$$Z_1 = \frac{1}{2}\left(A_p + A_p^T\right) \tag{7}$$

$$Z_2 = \frac{1}{2}\left(L_p + L_p^T\right) \tag{8}$$

In Eq. (7) and (8), A_p and L_p can be determined by $A_p = Y_{conc}.A.Y_{conc}^{'}$ and $L_p = A - W$, respectively. The obtained Z_1 and Z_2 are subjected to a generalized eigenvector problem [14] as follows

$$Z_2 E = \lambda Z_1 E \tag{9}$$

Once the eigenvectors are determined, the embedding is performed as

$$\hat{Y} = E^T Y_{conc} \tag{10}$$

The $\hat{Y}$ obtained from the above equation is the dimensionality reduced gene data with size $n_s^{'} \times n_s^{'}$. The $\hat{Y}$ is utilized to cluster the input microarray gene data using FCM.

3.2 Clustering of Gene Data by FCM

Any clustering method aims to produce a K × n partition matrix U($\hat{Y}$) of the given data set $\hat{Y}$, consisting of n objects, $\hat{Y}$ = {y1, y2, . . . , yn}, where K is the number of clusters. The partition matrix may be represented as U = [ukj], k = 1, . . .,K and j = 1, . . . , n, where ukj is the membership of pattern yj to the kth cluster. Greater value of ukj implies that the probability of belongingness of point yj to the kth cluster is more.

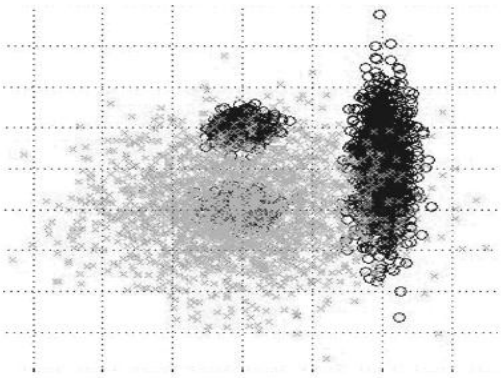

Fig 1: Shows cluster in DNA chromosomes

The well known fuzzy C-means uses cluster centers encoded in chromosomes have been used in designing as shown in figure 1. FCM clustering method optimizes Jm Eq. (11) andXB Eq. (12) cluster validity indices respectively.

$$J_m = \sum_{J=1}^{n} \sum_{k=1}^{k} \left(U_{kj}\right)^2 D^2 \left(Z_k, \hat{Y}_j\right) \tag{11}$$

$$XB\,(U, Z; X) = \frac{\sum_{i=1}^{k} \left(\sum_{k=1}^{n} (U_{ik})^2 D^2 (Z_i, X_n) \right)}{n \times \left(\min_{i \neq j} \{ \| Z_i - Z_j \|^2 \}\right)} \tag{12}$$

Here data set $\hat{Y}$ is partitioned into K clusters with centers Z = {z1, z2, . . . , zK} and m is the fuzzy exponent. ‖.‖ represents the Euclidean norm and D(zk, yj) denotes the distance of point yj from the center of the kth cluster. In this article, the Euclidean norm is taken as a measure of the distance between two points. The Jm index represents the global cluster variance, whereas XB index is expressed as a function of the ratio of total intra-cluster variance to the minimum inter-cluster separation. Hence both the measures are to be minimized.

4. RESULTS AND DISCUSSION

The proposed technique for microarray gene clustering has been implemented in the working platform of MATLAB (version 7.11). For evaluating the proposed technique, we have utilized the microarray gene samples of human acute leukemia and colon cancer data. [15] The high dimensional gene expression data has been subjected to dimensionality reduction and so a dimensionality reduced gene data with dimensions has been obtained. Thus LPP method is applied to identify informative genes and reduce gene dimensionality for clustering samples to detect their phenotypes.

Table I Microarray gene data dimension utilized for the evaluation process

Type of Gene Data	Number of Samples	Number of Genes	Dimensionality Reduced Data with the aid of LPP
ALL	38	7129	38 X 40
AML	34	7129	34 X 40
COLON	62	3000	62 X 42

A sample of microarray gene dataset of three classes that has been used for testing is given in the Table II. Clustering for microarray gene expression data whose amount is large can be fully calculated by determining the boundary of the clusters.

Table II: A sample of the microarray gene data to test the proposed technique

Class	**ALL**		**AML**		**COLON**	
Sample / Gene	**ALL 16125 TA-Norel**	**ALL 23368 TA-Norel**	**AML SH 5**	**AML SH 13**	**AFFX-MurIL2**	**AFFX-MurIL10**
AFFX-CreX-5_at (endogenous control)	-172A	-93A	-271A	-11A	20.6	-16
AFFX-CreX-3_at (endogenous control)	52A	10A	-12A	112A	-8.7	41.2
AFFX-BioB-5_st (endogenous control)	-134A	159A	-104A	-176A	4880	26.2

While testing, when a gene dataset is given, the proposed technique has to identify its belonging cluster. Clustering algorithms, such as K-means and Fuzzy C-means approaches are applied both to group genes, to partition samples in the early stage and have proven to be useful. The performance of each clustering algorithm may vary greatly with different data sets. Complete-link clustering method uses the smallest similarity within a cluster as the cluster similarity, and every data object within the cluster is related to every other with at least the similarity of the cluster. In order to test the performance of the data, N artificial m-dimensional feature vectors from a multivariate normal distribution having different parameters and densities were generated. Situations of large variability of cluster shapes, densities, and number of data points in each cluster were simulated.

Table III: Performance comparison in percentage between the proposed clustering technique (FCM) and the k-means techniques

Type of Gene Data	Accuracy		Correlation		Distance		Error Rate	
	FCM	k-means	FCM	k-means	FCM	k-means	FCM	k-means
ALL	83.9	71.1	0.342	0.235	0.00379	0.00476	0.21	0.38
AML	80.6	70.6	0.024	0.013	0.00364	0.00472	0.30	0.31
COLON	79.0	67.6	0.119	0.062	0.02029	0.02653	0.01	0.04

From the Table III, it can be seen that the proposed technique FCM has provided more accuracy, correlation and less distance and error rate rather than the k-means gene clustering techniques. More accuracy and less error rate leads to effective clustering of the given microarray gene data to the actual class of the gene.

5. CONCLUSION

In this paper, an effective microarray gene data clustering technique has been proposed with the aid of LPP and FCM. Initially, the dimensionality of the microarray data has been reduced with the aid of LPP mechanism. The technique has been tested by clustering the microarray gene expression data of human acute leukemia and colon cancer data. From the results, it can be noticed that our approach yields equally good results for the entire functional category. The comparative results have shown that the proposed technique possesses better accuracy, correlation and lesser distance, error rate than k-means gene clustering techniques. We have achieved improvement in the quality of the results by using FCM. Hence, this means of gene clustering have paved the way for effective information retrieval in the microarray gene expression data.

6. ACKNOWLEDGMENTS

I express my sincere thanks to Supervisor Dr. R.M Suresh for his continuous motivation, innovative ideas and assistance throughout this paper. My heart filled gratitude to doctoral committee members Dr. Raja and Dr. Palanivel for their timely support.

7. REFERENCES

[1] Satchidananda Dehuri and Sung-Bae Cho, "Multi-objective Classification Rule mining Using Gene Expression Programming," in proceedings of Third International Conference on convergence and Hybrid Information Technology, Vol. 2, pp. 754-760, 11-13 November, Busan, 2008. .

[2] Andrew K. Rider, Geoffrey Siwo, Scott J. Emrich, Michael T. Ferdig, Nitesh V, "A Supervised Learning Approach to the Ensemble Clustering of Genes",

[10] Jian Wen, "Ontology Based Clustering for Improving Genomic IR", Twentieth IEEE International Symposium International Journal of Data Mining and Bioinformatics, Vol. 3, No. 3, pp.229-259, 2009.

[3] Sushmita Mitra, Sankar K. Pal and Pabitra Mitra, "Data Mining in Soft Computing Framework: A Survey," IEEE Transactions On Neural Networks, Vol. 13, No. 1, 2002.

[4] Slavkov, I., Dzeroski, S., Struyf, J., Loskovska, S. "Constrained Clustering Of Gene Expression Profiles" in Proceedings of the Conference on Data Mining and Data Warehouses at the 7th International Multi-conference on Information Society, pp. 212-215, October 10-17, Slovenia, 2005.Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[5] Prabhjot Kaur, Anjana Gosain "A density oriented fuzzy C-means clustering algorithm for recognising original cluster shapes from noisy" International Journal of Innovative Computing and Applications 2011 - Vol. 3, No.2 pp. 77 - 87

[6] Y.Y. Leung and Y.S. Hung, "An Integrated Approach To Feature Selection And Classification For Microarray Data With Outlier Detection," in proceedings of 8th Annual International Conference on Computational Systems Bioinformatics, August 10-12, 2009

[7] Jian J. Dai, Linh Lieu, and David Rocke, "Dimension reduction for classification with gene expression microarray data," Statistical Applications in Genetics and Molecular Biology, Vol. 5, No. 1, pp. 1–21, 2006.

[8] D.Napoleon, S.Pavalakodi, "A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set", International Journal of Computer Applications Volume 13– No.7,pp. 41-46 January 2011

[9] P. Valarmathie, Dr MV Srinath, Dr T. Ravichandran. "Hybrid Fuzzy C-Means Clustering Technique for Gene Expression Data", International Journal of Research and Reviews in Apld Sci, Vol 1, No 1, pp. 33-37, October 09

on Computer-Based Medical Systems, pp.225 – 230, June 07

[11] Yuen, Man-chun, "Genomic sequence search and clustering using Q-gram", Bioinformatics thesis 2007.

[12] Wai-Ho Au, Keith C. C. Chan, Andrew K. C. Wong and Yang Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 2, pp. 83-101, 2005.

[13] Jacinth Salome and Suresh, "An Effective Classification Technique for Microarray Gene Expression by Blending

[14] of LPP and SVM", European Journal of Scientific Research, Vol.64, No.1, pp.34-43, 2011

[15] X. He and P. Niyogi, "Locality preserving projections," in Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press, 2003.

[16] [15] Microarray gene samples of human acute leukemia and colon cancer data http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi