Efficient Clustering for Gene Expression Data

K.sathishkumar[#1],M.Ramalingam[*2]

[#]*Assistant Professor of Computer Science, Gobi Arts & Science College (Autonomous)*

Gobichettipalayam, India.

[1]Sathishmsc.vlp@gmail.com

[*]*Assistant Professor of Computer Science, Gobi Arts & Science College (Autonomous)*

Gobichettipalayam, India.

[2]ramsgobi@gmail.com

***Abstract*---Many clustering techniques have been proposed for the analysis of gene expression data obtained from microarray experiments. However, choice of suitable method(s) for a given experimental dataset is not straightforward. The amount of biological data such as DNA sequences and microarray data have been increased tremendously. 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. 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 above mentioned problems. In this paper, to improve the searching and the clustering performance in genomic data proposed gene clustering technique. Firstly, the high dimensionality of the microarray gene data is reduced using an improved LPP method named Scatter-Difference Discriminant Locality Preserving Projection (SDDLPP). The SDDLPP 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 Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM) is shown to achieve higher efficiency, clustering quality and automation than other clustering method.**

Keywords*--- *Gene expression data, Locality Preserving Projection, Fuzzy Possibilistic C-Means Algorithm using EM Algorithm

I. INTRODUCTION

Gene expression (GE) is the fundamental link between genotype and pheno- type in a species, with microarray technologies facilitating the measurement of thousands of GE values under tightly controlled conditions, e.g. (i) from a particular point in the cell cycle, (ii) after an interval response to some environmental change, (iii) from RNA, isolated from a tissue exhibiting certain phenotypic characteristics and so on. Many excellent reviews of microarray data analysis using clustering techniques are available. Microarray gene expressions studies are routinely carried out to measure the transcription levels of an organism's genes. A common aim in the analysis of expressions measurements observed in a population is the identification of naturally occurring sub-populations. In cancer studies, for instance, the identification of sub-groups of tumours having distinct mRNA profiles can help discover molecular fingerprints that will define subtypes of disease (Smolkin and Ghosh, 2003) [1].

The goal of clustering is to determine a natural combination in a group of patterns, points, or objects, without knowledge of any class labels. Clustering is widespread in any discipline that involves analysis of multivariate data. It is, of course, impractical to exhaustively list the numerous uses of clustering techniques. In the background of the human genome development, new technologies were emerged, it facilitate the parallel execution of experiments on a large number of genes at the same time. Hence it is called as DNA microarrays, or DNA chips, constitute a prominent example. This technology aims at the measurement of mRNA levels in particular cells or tissues for many genes at once. To this end, single strands of balancing DNA for the genes of interest which can be immobilized on spots arranged in a grid on a support which will typically be a glass slide, a quartz wafer, or a nylon membrane. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample [2].

The parallelism in this kind of experiment lies in the hybridization of mRNA extracted from a single sample to many genes at once using clustering technique. The measured values are not obtained on an absolute scale. Because it depends on many factors such as the efficiencies of the various chemical reactions involved in the sample preparation, as well as on the amount of immobilized DNA available for hybridization. The class of transcripts that is probed by a spot may differ in different applications. Most commonly, each spot is meant to probe a particular gene. The representative sequence of DNA on the spot may be either a carefully selected fragment of cDNA, a more arbitrary PCR product amplified from a clone matching the gene. Another level of sophistication is reached when a spot represents, e.g., a particular transcript of a gene. In this case or for the distinction of mRNA abundances of genes from closely related gene families, careful design and selection is

made of the immobilized DNA are required. Similarly, the selection of samples to study and to compare to each other using DNA microarrays requires careful planning as will become clear upon consideration of the statistical questions arising from this technology [3] [4] [5].

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 [6]. Also, there is a high redundancy in microarray data as well as several genes have irrelevant information for exact clustering of diseases or phenotypes.[7] 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. [8] Initially, the dimensionality reduced data is applied with the EMFPCM for clustering.

II. LITERATURE SURVEY

Bradley et al., [9] put forth a technique for the purpose of refining initial points for clustering algorithms, in particular K-Means clustering algorithm. They presented a quick and competent algorithm in order to refine an initial starting point a normal class of clustering algorithms. The iterative techniques that are more sensitive to initial preliminary conditions were used in most of the clustering algorithms like K-Means, and EM normally converges to one local minima. They implemented this iterative technique for refining the initial condition which permits the algorithm to converge to an enhanced local minimum value. The refined initial point is used to evaluate the performance of K-Means algorithm in clustering the given data set.

Jiabin Deng et al., [10] proposed an improved fuzzy clustering-text clustering method based on the fuzzy C-Means clustering algorithm and the edit distance algorithm. The author used the feature evaluation to reduce the dimensionality of high-dimensional text vector. Because the clustering results of the traditional fuzzy C-Means clustering algorithm lack the stability, the author introduced the high-power sample point set, the field radius and weight. Due to the boundary value attribution of the traditional fuzzy C-Means clustering algorithm, the author recommended the edit distance algorithm.

Celikyilmaz et al., [11] proposed a new fuzzy system modeling approach based on improved fuzzy functions to model systems with continuous output variable. The new modeling approach introduces three features: i) an Improved Fuzzy Clustering (IFC) algorithm, ii) a new structure identification algorithm, and iii) a nonparametric inference engine. The IFC algorithm yields simultaneous estimates of parameters of c-regression models, together with fuzzy c-partitioning of the data, to calculate improved membership values with a new membership function. The structure identification of the new approach utilizes IFC, instead of standard fuzzy C-Means clustering algorithm, to fuzzy partition the data, and it uses improved membership values as additional input variables along with the original scalar input variables for two different choices of regression methods: least squares estimation or support vector regression, to determine ldquo fuzzy function srdquo for each cluster. With novel IFC, one could learn the system behavior more accurately compared to other FSM models. The nonparametric inference engine is a new approach, which uses the alike -nearest neighbor method for reasoning.

In 1997, Pal et al., [12] proposed the fuzzy-possibilistic C-Means (FPCM) technique and algorithm that generated both membership and typicality values when clustering unlabeled data. FPCM constrains the typicality values so that the sum over all data points of typicality's to a cluster is one. For large data sets the row sum constraint produces unrealistic typicality values. In this approach, a new model is presented called Possibilistic-Fuzzy C-Means (PFCM) model. PFCM produces memberships and possibilities concurrently, along with the usual point prototypes or cluster centers for each cluster. PFCM is a hybridization of FCM and Possibilistic C-Means (PCM) that often avoids various problems of PCM, FCM and FPCM. The noise sensitivity defect of FCM is resolved in PFCM, overcomes the coincident clusters problem of PCM and eliminates the row sum constraints of FPCM. The first-order essential conditions for extreme of the PFCM objective function is driven, and used them as the basis for a standard alternating optimization approach to find local minima of the PFCM objective function. PFCM prototypes are less sensitive to outliers and can avoid coincident clusters, PFCM is a strong candidate for fuzzy rule-based system identification

An EM (Expectation maximization) algorithm is very use full in statically model. The most common algorithm uses an iterative refinement technique. These algorithms are giving the best result in clustering method; it is also referred to as LoyardAlgo particularly in the computer science community. EM algorithms given an initial set of c-means $m_1^{(1)}, \ldots\ldots, m_1^{(k)}$, the algorithm proceeds by alternating between two steps [13].

FPCM constructs memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. Hybridization of PCM and FCM is the FPCM using EM Algorithm that often avoids various problems of PCM, FCM and FPCM using EM. FPCM using EM solves the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM. But the estimation of centroids is influenced by the noise data. Hence Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM) [14] is proposed.

III. THE PROPOSED GENE CLUSTERING TECHNIQUE BASED ON SDDLPP AND EMFPCM

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 EMFPCM-based clustering.

A. Dimensionality Reduction by SDDLPP

Firstly, define the objective function of the given SDDLPP method as follows:

$$J = S_B - kS_W$$

$$= \sum_{i,j=1}^{C} (m_i' - m_j')^2 w_{ij}' - k \sum_{i,j=1}^{N} (Y_i - Y_j)^2 w_{ij}$$

where, SB and SW is called the between-class scatter and within-class scatter, respectively. The SDDLPP subspace can be obtained by maximizing the objective function J with its purpose to seek efficient discrimination among the different classes while preserving the local structure of the data. The scatter-difference discriminant rule is consistent with maximizing the between-class scatter while minimizing the within-class scatter. k is a nonnegative constant to balance the contribution of SB and SW. When k = 0, the between-class scatter plays a key role, otherwise, the SDDLPP subspace is mainly decided by the within-class scatter while $k \to \infty$. In this low dimensional SDDLPP subspace, the projection data from the same class should be 'close' with each other while from the different classes should be 'far' with each other for considering the class-specified information.

For within-class scatter SW, yi = AT xi is the projection of xi onto the transformation matrix A and Tij is the within-class weight coefficient between the data xi and xj that are from the same class. The within-class scatter SW can be reduced to:

$$S_W = \sum_{i,j=1}^{N} (y_i - y_j)^2 w_{ij}$$

$$= \sum_{i,j=1}^{N} (A^T x_i - A^T x_j)^2 w_{ij}$$

$$= A^T \left[\sum_{i,j=1}^{N} (x_i x_i^T - x_i x_j^T - x_j x_i^T - x_j x_j^T) w_{ij} \right] A$$

$$= 2A^T \left(\sum_{i=1}^{C} x_i D_{ii} x_i^T - \sum_{i,j=1}^{C} x_i w_{ij} x_j^T \right) A$$

$$= 2A^T (XDX^T - XWX^T) A$$

$$= 2A^T X(D - W) X^T A$$

$$= 2A^T XLX^T A$$

In order to make the projection data from the same class be 'close' with each other, the within-class weight matrix w_{ij} is defined as $w_{ij} = \exp\left(-\| X_i - \frac{X_j\|^2}{t}\right)$. If data xi and xj belong to the same class, otherwise, $W_{ij} = 0$. $D_{ij} = \sum_j w_{ij}$ is a diagonal matrix and $L = D - W$ is called within class Laplacian matrix.

Similar to S_W, the between-class scatter S_B can be reduced to:

$$S_B = \sum_{i,j=1}^{C} (m_i' - m_j')^2 w_{ij}'$$

$$= \sum_{i,j=1}^{C} [A^T (m_i - m_j)]^2 w_{ij}'$$

$$= 2A^T \left(\sum_{i=1}^{C} m_i D_{ij}' m_i^T - \sum_{i,j=1}^{C} m_i w_{ij}' m_i^T \right) A$$

$$= 2A^T (MD'M^T - MW'M^T) A$$

$$= 2A^T M(D' - W')M^T A$$

$$= 2A^T ML'M^T A$$

Where, $m_i' = A^T m_i$ is the projection vector of m_i with $m_i = (\sum_{k=1}^{N_i} x_i)/N_i$ being the average data of the class i. w_{ij}' is the between-class weight coefficient between the class i and class j. If m_i and m_j are the k-nearest neighbors, w_{ij}'

can be defined as:

$$w_{ij}' = \exp\left(-\| \frac{m_i - m_j\|^2}{t'}\right)$$

w_{ij}' can also be simplified as $w_{ij}' = 1$, If m_i and m_j are the k-nearest neighbors. Parameter t' is a suitable constant that can be decided by experiment results. The purpose of

defining w_{ij}^{l} is to make the two classes which are 'close' in original high-dimensional space be 'far' in the low dimensional SDDLPP subspace. Then, between-class weight matrix W^{l} is constructed through the k-nearest neighbor's graph whose nodes are the class average data. $D_{ii}^{l} = \sum_{j} w_{ij}^{l}$ is a $C \times C$ dimensional diagonal matrix and $L^{l} = D^{l} - W^{l}$ is the between-class Laplacian matrix. Hence the high dimensional matrix obtained above is reduced to gene expression data size. Hence it is utilized to cluster the input microarray gene data using EMFPCM.

B. Clustering of Gene Data by EMFPCM

The proposed algorithm is called Expectation maximization fuzzy Possibilistic c-means clustering (EMFPCM). EMFPCM algorithms is very help full to increase the performance of machine learning, all data mining approaches, image processing, network security etc. Proposed algorithms used the clustering techniques and EM algorithms, which provided the sufficient result for the cluster analysis in maximum mean to calculate the fixed centroid and correct threshold value. Proposed algorithms have these steps.

Step1: In these step we find the membership matrix (U) initialize randomly in by the below equation

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, \ldots, n\}$$

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)} \right)^{2/(m-1)} \right], 1 \le i \le c, 1 \le j$$

This equation represent the membership matrix, it has taken the value equal to 1.

Step2: Calculate the centroids (V_i) in equation 5

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}, 1 \le i \le j \le n$$

Centroid is main point of the cluster analysis system, in clustering this value of v_i is depends on the member matrix function and related parameter of x_j.

Steps3: Using dissimilarity function to calculate the dissimilarities between centroid and data points in equation and check threshold value in equation 3.4 then stop if we find the correct threshold value.

$$J_{FPCM}(U, T, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij}^{m} + t_{ij}^{n}) d^{2}(x_j, v_i)$$

In this steps check the threshold value using membership matrix and Euclidian distance between ith centroid (V_i) and jth data point.

$$\text{if} \| U(k+1) - U(k) \| < \varepsilon$$

In this equation we have to check the value of present classes and the next classes of the membership function. Check the value of membership matrix in next membership matrix in correct threshold value, which have been done by dissimilarities between centroid and data points. If values are satisfied, then we forwarded the next steps.

Step4: If threshold value is not correct we find the new mean (m_i) using EM algorithm that has constraints in equation (14) using this equation we can find new mean, which is provided the correct threshold value for the dissimilarity function, So we can say these algorithms is deemed to have converged when the assignment no longer change. And it gave the best performance of initial means. EM algorithms commonly used initialization methods are random partition.

$$m_i^{(l+1)} = \frac{1}{|S_i^{(l)}|} \sum_{x_j \in S_i^{(l)}} X_j$$

Step5: In these steps assign each observation to the cluster with the closest means

$$S_i^{(l)} = \{x_p : \| x_p - m_i^{(1)} \| \le \| x_p - m_j^{(l)} \| \ \forall 1 \le j \le k$$

Hence find new mean for new cluster. So we can say these algorithms is deemed to have converged when the assignment no longer change.

The proposed algorithm using EM Algorithm is developed to obtain better quality of clustering results. The objective function is based by adding new weight of data points in relation to every cluster and modifying the exponent of the distance between a point and a class.

IV. 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.

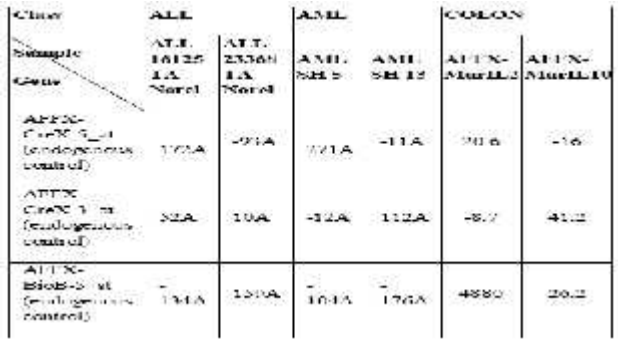
TABLE I
MICROARRAY GENE DATA DIMENSION UTILIZED FOR THE EVALUATION PROCESS

Types of Gene Data	Number of Samples	Number of Genes	Dimensionality Reduced Data with the aid of LPP
ALL	41	7139	41X41
AML	36	7128	36X40
COLON	68	3000	62X42

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



While testing, when a gene dataset is given, the proposed technique has to identify its belonging cluster. Clustering algorithms, such as Fuzzy C-means and Fuzzy Possibilistic C-Means Algorithm using EM Algorithm 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 EMFPCM CLUSTERING TECHNIQUE AND OTHER EXISTING TECHNIQUES

Type of Gene Data	Accuracy			Correlation			Distance			Error Rate		
	FCM	FPCM	EMFPCM	FCM	FPCM	EMFPCM	FCM	FPCM	EMFPCM	FCM	FPCM	EMFPCM
ALL	83.1	83.9	85.69	0.345	0.368	0.412	0.00379	0.00346	0.00263	0.21	0.20	0.18
AML	80.06	81.02	83.84	0.024	0.029	0.0315	0.00364	0.00331	0.00201	0.30	0.29	0.24
COLON	79.0	79.9	81.96	0.119	0.125	0.139	0.02029	0.02011	0.0126	0.04	0.03	0.01

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

V. CONCLUSION

Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. In this paper, an effective microarray gene data clustering technique has been proposed with the aid of SDDLPP and EMFPCM. Initially, the dimensionality of the microarray data has been reduced with the help of SDDLPP 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 FCM, FPCM gene clustering techniques. We have achieved better improvement in the quality of the results by using EMFPCM. Hence, this means of gene clustering have paved the way for effective information retrieval in the microarray gene expression data.

REFERENCES

[1] M. Smolkin and D. Ghosh (2003). Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics, 4(1), 36.

[2] Wolfgang Huber, Anja von Heydebreck and Martin Vingron, “Analysis of microarray gene expression data”, Max-Planck-Institute for Molecular Genetics 14195 Berlin April 2, 2003.

[3] M. Kathleen Kerr and Gary A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 77:123–128, 2001.

[4] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32 Suppl. 2:490–495, 2002.

[5] Yee Hwa Yang and Terence P. Speed. Design issues for cDNA microarray experiments. *Nat. Rev. Gen.*, 3:579–588, 2002.

[6] 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.

[7] 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

[8] 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

[9] P.S. Bradley and U.M. Fayyad, “Refining Initial Points for K-Means Clustering”, ACM, Proceedings of the 15th International Conference on Machine Learning, Pp. 91-99, 1998.

[10] Jiabin Deng, JuanLi Hu, Hehua Chi and Juebo Wu, “An Improved Fuzzy Clustering Method for Text Mining”, Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), Vol. 1, Pp. 65–69, 2010.

[11] Celikyilmaz A and Burhan Turksen I, “Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm”, IEEE Transactions on Fuzzy Systems, Vol. 16, No. 3, Pp. 779–794, 2008.

[12] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, “A Possibilistic Fuzzy c-Means Clustering Algorithm”, IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[13] Neal, Radford; Hinton, Geoffrey (1999). Michael I. Jordan. ed."A view of the EM algorithm that justifies incremental, sparse, and other variants". Learning in Graphical Models (Cambridge, MA: MIT Press): 355–368.

[14] R.Shanthi and R.Suganya, “Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM)”, International Journal of Computer Applications (0975 – 8887) Volume 61– No.12, January 2013

[15] 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.