Gene Expression Analysis Using Fuzzy K-Means Clustering

K.Sathishkumar[#1], Dr. V. Thiagarasu[*2], Dr. E. Balamurugan[#3], Dr. M. Ramalingam[#4]
[#1, 3] *Lecturer of IT/IS, Associate Professor of IT/IS,*
BlueCrest College, Accra, Ghana.
[1]*sathishmsc.vlp@gmail.com*
[*2]*Associate Professor of Computer Science, Gobi Arts & Science College (Autonomous), Gobi, TN.*
[*4]*Assistant Professor of Computer Science, Gobi Arts & Science College (Autonomous), Gobi, TN.*

Abstract — **A microarray measures the expression levels of thousands of genes at the same time. Clustering helps to analyse microarray gene expression data. In this paper, have implemented a biclustering algorithm to identify subgroups of data which shows correlated behaviour under specific experimental conditions. In the process of finding biclusters, Fuzzy K-means clustering is used to cluster the genes and samples with maximum membership function. Both dimensionality and reducing the gene shaving are done using principal component analysis & gene filtering with the function respectively. This method obtains highly correlated sub matrices of the gene expression dataset. Biclustering is a NP-hard problem therefore which is implemented biclustering in multi-core parallel environment to reduce the computational time of the algorithm. Besides have compared the results with other parallel & sequential algorithm to show the effectiveness of the algorithm.**

Keywords— **Microarray, gene expression, Multicore platform, Biclustering, MATLAB parallel computing, PCA, gene entropy.**

I. INTRODUCTION

Microarray helps in studying the variations of many genes simultaneously. With the development of microarray techniques, a lot of work has been done on the analysis of gene expression data. Microarray experiments, identifies co-expressed genes that share similar expression patterns [1]. Clustering is the key initial step in the analysis of gene expression data to find the co-expressed genes. In the previous studies, it has been found that the genes showing similar expression patterns are likely to be involved in same cellular processes. When the pattern that related genes showing similar transcriptional behaviour under a subset of condition is called bicluster [3, 4]. It is a type of subspace clustering. Generally, a gene only belongs to one cluster but in practicality, a gene involves many cellular processes, so a gene may belong to more than one cluster. A better solution would be to introduce fuzzy concepts at the time of gene clustering because impact of biclustering will be more obvious. Strong correlations of expression patterns between the genes indicated as co-regulation and are controlled by same regulatory mechanisms [1, 2]. One of the major characteristic of gene expression data is to find meaningful clusters or groups with respect to both genes and samples dimensions. Hartigan developed the two way clustering approach for biclustering, and Cheng and Church were the first to use this concept for microarray data analysis [22] [7]. Wei et al. proposed a parallel biclustering algorithm based on antimonite ones property of the quality of the data sets with their sizes. Tewfik et al., proposed parallel biclustering of genes with coherent evolutions [11] [9]. It finds all biclusters with a specified minimum numbers of genes and conditions in the datasets. Jiang et al., proposed mining approach for coherent gene clusters from microarray gene expression data set; they have presented two approaches namely sample gene search and gene-sample search to mine a set of coherent gene clusters [15]. The serial version of Interrelated Two-Way Clustering which is proposed and parallelized in and tested to find biclusters. Chandra et.al, has also used this concept to find biclusters in gene expression data with fuzzy approach [5] [18][6]. In this paper the concept of fuzzy c-means to cluster the genes and sample dimensions. In the next step maximize the sum of distances between the genes having maximum membership function to find the gene canters. For the preparation of biclusters the gene entropy filters and Principal component analysis for the gene shaving process. The main goal of the algorithm is to find coherent values of gene bicluster one set at a time with this new approach. Parallel approach of the algorithm on multicore processor is provided the better performance than sequential one.

II. RELATED WORKS

An easy way to comply with the journal paper formatting Clusters has been studied since years. Based on the previous idea, (Lazzeroni et al., 2000) presents the plaid models, in which the data matrix is described as a linear function of layers corresponding to its biclusters and shows how to estimate a model using an iterative maximization process. Shamir (Shamir et al., 2002) proposes a new method to obtain biclusters based on a combination of graph theoretical and statistical modeling of data. In this model, a gene responds to a condition when its expression level changes significantly at that condition of its normal level [12] [13]. In a recent work (Liu et al., 2004), a generalization of OPSM model is introduced and presented [14] [15]. The OPSM model is based on the search of biclusters in which a subset of genes induces a similar linear ordering along a subset of conditions. Some techniques search for specific structures in data matrix to find biclusters: (Gerstein et al., 2003) creates a method for clustering genes and conditions simultaneously based on the search of —checkerboard patterns in matrices of gene expression data [16]. The data is already processed by normalization in a spectral framework model (several schemes built around the idea of putting the genes on the same scale so that they have the same average level of expression across the same conditions). Evolutionary computation techniques have also been used in this research area. These techniques use aspects from natural selection within computer science, including genetic algorithms, genetic programming and evolutionary strategies. In an evolutionary technique, the following a sequential covering strategy and measuring the mean squared residue, is used based on the search of biclusters [17]. Another approach is pattern-based clustering, that captures the similarity of the patterns exhibited by a bicluster. In general a set of, given objects, a subset of these objects form a pattern-based cluster follows a similar pattern in a subset of dimensions. Wang et al., proposes a depth-first algorithm for detecting pattern based clusters [11]. In order to speed up the process and to avoid the repetition of computations, the algorithm uses a suffix tree to efficiently enumerate the possible combinations of row and column sets that represent a bicluster. Liu and Wang also proposed an exhaustive bicluster enumeration algorithm, which is based on a model that generalizes the order preserving sub matrix model [10] [18]. The objective of finding all biclusters that, after column reordering, represent coherent evolutions of the symbols in the matrix is achieved by using a pattern discovery algorithm inspired in sequential pattern mining algorithms [19]. Another algorithm used in the pattern-based clustering model is proposed in [20]. This algorithm mines only the maximal pattern-based clusters. It conducts a depth first, divide and conquer search and prunes unnecessary branches smartly. Most of these previous techniques search for one or two types of biclusters among four that have been identified in the literature [21]: biclusters with constant values, biclusters with constant values on rows or columns, biclusters with coherent values, and biclusters with coherent evolution. Constant biclusters in a gene expression matrix identify subsets of genes with equal expression values within a subset of conditions. Biclusters with constant values along rows indicate a subset of genes with expression levels that do not change across a subset of conditions, irrespective of the actual expression levels of the individual genes. Biclusters with constant columns isolate a subset of conditions for which a subset of genes have constant expression values that may differ from condition to condition. Much of the prior work, however, has focused on finding more complex relations between genes and conditions by looking for biclusters with coherent values or evolutions. Such biclusters allow for variation in the actual numerical values of the gene expression levels. Instead, they focus on the behavior of the gene expression levels across subsets of genes or conditions. Gene expression levels in biclusters of coherent values obey additive or multiplicative models on rows or columns. Our focus here, and indeed that of most researchers, is on finding subsets of genes that are up regulated or down regulated across a subset of conditions irrespective of their actual expression values or subsets of conditions that have always the same or opposite effects on a subset of genes. Most previous techniques are also greedy and will miss some biclusters that satisfy their definition of a valid bicluster. Many of these pioneering approaches used a cost function to define biclusters. In many cases, the cost function will measure the square deviation from the sum of the mean value of expression levels in the entire bicluster and the mean values of expression levels along each row and column in the bicluster [22].

III. METHODOLOGY

The proposed approach consists of three stages *namely finding* of co-regulated biclusters using Bimax algorithm, dimensionality reduction using LFDA and clustering using Fuzzy K-Means Clustering Algorithm (FKM).

FINDING OF COREGULATED BICLUSTERS USING BIMAX ALGORITHM

The diagram for finding the co regulated clusters using algorithm is shown in Fig.1. The input is gene sequence f the micro array data. Enhanced Bimax algorithm is used to display a maximal biclusters value and displays a co regulated biclusters. The Enhanced Bimax algorithm is used to measure a particular gene is present or not. It also finds the transcription sites of the co regulated biclusters. Normalization technique used to specify genes are presented in the particular group or not. The output is display the transcription factors.

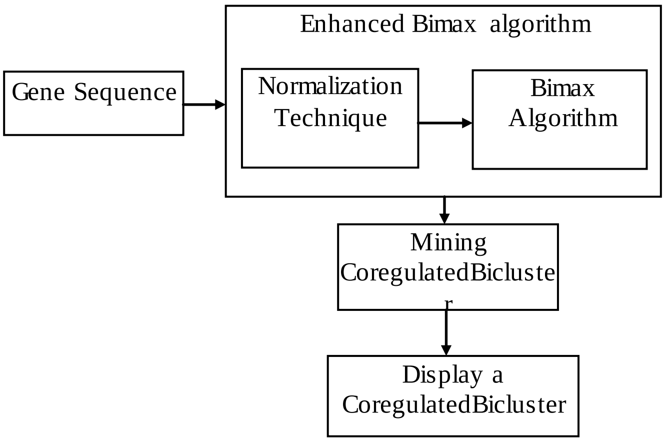

Fig. 1 Block diagram for mining co regulated bicluster.

A. Bimax Algorithm

The Bimax algorithm needs to guarantee the only optimal, inclusion-maximal biclusters which is to be generated. The problem arises because V contains parts of the biclusters found in U, and as a consequence needs to ensure that the algorithm only considers those biclusters in V that extend over CV. The parameter Z serves this goal. It contains sets of columns that restrict the number of admissible biclusters. It is used to specify the genes and conditions. It is used to specify the analysis of DNA chips and gene networks. The algorithm realizes the divide-and-conquer strategy. Fig. 1 describes an original Bimax algorithm. It consists of three procedures.

They are Enhanced Bimax, Conquer and Divide. Conquer function is call and check the condition is if the genes and conditions are equal then the partitioning is begin, otherwise it stop the process. Second step is split the data and normalization technique is used to group the splited data. It is used to find all add the maximum groups in general gene expression data. Each co regulated genes are grouping together the particular expression value and the particular situation.

B. Proposed Enhanced Bimax Algorithm

Enhanced Bimax algorithm can contain two procedures. Fig. 2 describes a flowchart for Proposed Enhanced Bimax algorithm.

(BSP) is a method for recursively subdividing a space into convex sets by hyper planes. This subdivision gives rise to a representation of the scene by means of a tree data structure known as a BSP tree. Normalization is the process of isolating statistical error in repeated measured data. Quintile normalization for instance, is normalization based on the magnitude of the measures. The goals in doing eliminate all the redundant data and ensure data dependencies. The numbers of genes that reproducibly showed and the un-normalized data and normalized data are displayed on the co-regulated biclusters. Enhanced Bimax algorithm is applied data mining technique on clustering. In the clustering similar samples and similar gene probes are organized in a fashion so that they would lie close together. It consists of three procedures. They are Enhanced Bimax, BFS and BSP.

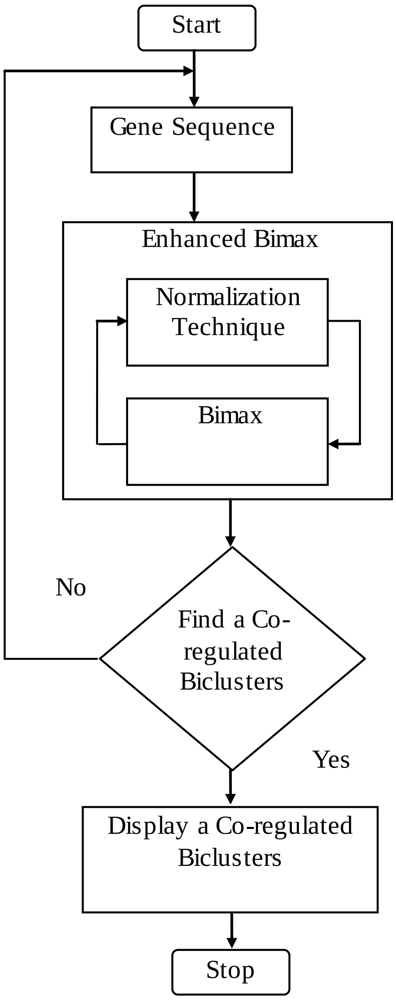

Fig. 2 Flow Chart for Proposed Enhanced Bimax Algorithm

First step is normalization technique which is used to remove the redundant data and for grouping genes in the specific conditions. Binary Space Partitioning function is call and check the condition is if the genes and conditions are equal then the partitioning is begin. Otherwise it stops the process. It specifies that a particular gene is present in the given group then it is represents a one. With these maximum groups in general gene expression data can be found. Each co regulated genes are grouping together the particular expression value and the particular situation. Otherwise the gene is not present in the given group an it is representing as zero. Fig. 2 describes a proposed Enhanced Bimax algorithm [23].

C. LOCAL FISHER'S DISCRIMINANT ANALYSIS (LFDA) FOR DIMENSIONALITY REDUCTION

Assume the training samples $\{(X_i, y_i) | X_i \in \mathbb{R}^m, y_i \in \{1,2,\dots,c\}, i = 1,2,\dots,n\}$, where X_i is the th training sample, y_i is the corresponding label of X_i , c is the number of classes and is the total number of training samples. Let n_l be the number of training samples in class ω_l and $n = \sum_{l=1}^{c} n_l$.

A. LFDA

LFDA is a recent extension to LDA that can effectively handle the multi-modal/non-Gaussian problem. It is a supervised feature projection technique that effectively combines the properties of LDA and an unsupervised manifold learning technique—Locality Preserving Projection (LPP) [13]. For more information about LPP, readers are referred to [11]. The overall idea of LFDA is to obtain a good separation of samples from different classes while preserving the local structure of point-clouds of each class [23].

The local within-class scatter matrix $S^{(lw)}$ and the local between class scatter matrix $S^{(lb)}$ used in LFDA are defined as follows

$$S^{(lb)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(lb)} (X_i - X_j)(X_i - X_j)^T \qquad (1)$$

$$S^{(lw)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(lw)} (X_i - X_j)(X_i - X_j)^T \qquad (2)$$

Where $W^{(lb)}$ and $W^{(lw)}$ are $n \times n$ matrices defined as

$$W_{i,j}^{(lb)} = \begin{cases} A_{i,j}\left(\frac{1}{n} - \frac{1}{n_l}\right), & if \ \ y_i = y_j = l, \\ \frac{1}{n} & if \ \ y_i \neq y_j \,, \end{cases} \qquad (3)$$

$$W_{i,j}^{(lw)} = \begin{cases} \frac{A_{i,j}}{n_l} & if \ \ y_i = y_j = l, \\ 0, & if \ \ y_i \neq y_j \,, \end{cases} \qquad (4)$$

The affinity matrix $A_{i,j}$ used in this work is defined as

$$A_{i,j} = \exp\left(-\frac{\|X_i - X_j\|^2}{Y_i Y_j}\right) \qquad (5)$$

Where $\mathrm{Y}_i = \|X_i - X_j{}^{(k_{nn})}\|$ represents the local scaling of data samples in the neighborhood of X_i, and X_i^k is the th-nearest neighbor of X_i.

The transformation matrix $\hat{T}_{LFDA}$ of LFDA can then be computed by maximizing the local Fisher's ratio.

$$\hat{T}_{LFDA} = \underset{T}{\arg max}\left\{trace\left[\left(T^T S^{(lw)} T\right)^{-1} T^T S^{(lb)} T\right]\right\} \qquad (6)$$

which can be solved as a generalized eigenvalue problem involving $S^{(lb)}$ and $S^{(lw)}$.

D. CLUSTERING OF GENE EXPRESSION DATA USING FUZZY K-MEANS CLUSTERING ALGORITHM (FKM)

The fuzzy k-means clustering (FKM) algorithm performs iteratively the partition step and new cluster representative generation step until convergence. The fuzzy k-means clustering algorithm partitions data points into k clusters $S_l(l = 1,2, \ldots .k)$ and clusters S_l are associated with representatives (cluster center) C_l. The relationship between a data point and cluster representative is fuzzy. That is, a membership $u_{i,j} \in [0,1]$ is used to represent the degree of belongingness of data point X_i and cluster center C_j. Denote the set of data points as $S = \{X_i\}$. The FKM algorithm is based on minimizing the following distortion:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{N} u_{i,j}^{m}\, d_{ij} \qquad (1)$$

with respect to the cluster representatives C_j and memberships $u_{i,j}$, where N is the number of data points; m is the fuzzifier parameter; k is the number of clusters; and d_{ij} is the squared Euclidean distance between data point X_i and cluster representative C_j. It is noted that $u_{i,j}$ should satisfy the following constraint:

$$\sum_{j=1}^{k} u_{i,j} = 1, for\ i = 1\ to\ N \qquad (2)$$

The major process of FKM is mapping a given set of representative vectors into an improved one through partitioning data points. It begins with a set of initial cluster centers and repeats this mapping process until a stopping criterion is satisfied. It is supposed that no two clusters have the same cluster representative. In the case that two cluster centers coincide, a cluster center should be perturbed to avoid coincidence in the iterative process. If $d_{ij} < \eta$, then $u_{i,j} = 1$ and $u_{j,l} = 1$ for l ≠ j, where η is a very small positive number. The fuzzy k-means clustering algorithm is now presented as follows.

1. Input a set of initial cluster centers $SC_n = \{C_i(0)\}$ and the value of ε. Set p = 1.

2. Given the set of cluster centers SC_n, compute $d_{i,j}$ for i = 1 to N and j = 1 to k. Update memberships $u_{i,j}$ using the following equation:

$$u_{i,j} = \left((d_{i,j})^{\frac{1}{m-1}} \sum \left(\frac{1}{\cdot}\right)^{m-1} \right) \qquad (3)$$

If $d_{i,j} < \eta$, set $u_{i,j} = 1$, where η is a very small positive number.

3. Compute the center for each cluster using Eq. (4) to obtain a new set of cluster representatives SC_{n+1}.

$$C_i(p) = \frac{\sum_{i=1} u_{i,j} X_i}{\sum^{N} \quad m} \qquad (4)$$

4. If $\|C_i(p) - C_i(p-1)\| < \varepsilon$ for j = 1 to k, then stop, where ε > 0 is a very small positive number. Otherwise set p + 1 → p and go to step 2.

The major computational complexity of FKM is from steps 2 and 3. However, the computational complexity of step 3 is much less than that of step 2. Therefore the computational complexity, in terms of the number of distance calculations, of FKM is O(Nkt), where t is the number of iterations.

IV. CONCLUSIONS

In this paper, another new approach of two way clustering of gene expression data has been proposed in multicore environment. The approach uses fuzzy C-means clustering algorithm for grouping genes and sample dimensions, two gene expression used by yeast S. cerevisiae Cho et al., and Tavazoie et al., data. Gene Entropy filtering shows the good performance in gene shaving, because it removed the maximum waste and noisy data and to rank the informative genes. Based on performance of PCA, it is a good powerful bicluster verification tool and it has the ability to remove correlation of the data. The study of results of biclustering also shows that the number of clusters of genes with two and four respectively. Size two gives larger biclusters whereas size four has given smaller one as compared to both the dataset. It can be assumed that large number of clustering with smaller biclusters shows more cohesiveness of the biclusters. The overall results demonstrated that this approach outrun parallel fuzzy interrelated two ways clustering even with reduced number of genes.

REFERENCES

1. Jinze Liu1, Jiong Yang2, and Wei Wang1 ―Biclustering in Gene Expression Data by Tendency‖ Computational Systems Bioinformatics Conference, 2004. IEEE
2. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein.Cluster analysis and display of genome-wide expression patterns. In Proc Natl Acad Sci U S A,95(25):14863-8, 1998.
3. S. Kaski, J. Nikkil, and G. Wong. Analysis And Visualization Of Gene Expression Data Using Self- Organizing Maps, Proceedings of NSIP-01, IEEEEURASIP Workshop on Nonlinear Signal and Image Processing, 2001.
4. R. Sharan and R. Shamir. Click: A clustering algorithm with applications to gene expression analysis. In ISMB, pages 307-216, 2000.
5. R. Agrawal, J.C. Gehrke, D. Gunopulos, P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, in: Proceedings of the ACM International Conference on Management of Data (SIGMOD'98), 1998, pp. 94–105.
6. C.C. Aggarwal, P.S. Yu, Finding generalized projected clusters in high dimensional spaces, in: Proceedings of the ACM International Conference on Management of Data (SIGMOD'00), 2000, pp. 70–81.
7. Hartigan,J.A. (1972) Direct clustering of a data matrix, Journal of the American Statistical Association, 67(337), 123-129.
8. Cheng,Y., Church,G.M. (2000) Biclustering of expression data, Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, ISMB'00, 93-103.
9. Yang,J. ,Wang,W., Haixun,W., Yu,P. (2002) Improving Performance of Bicluster Discovery in a Large Data Set, 6th ACM International Conference on Research in Computational Molecular Biology, RECOMB2002, Poster.
10. Yang,J. ,Wang,W., Haixun,W., Yu,P. (2003) Enhanced biclustering on expression data, 3rd IEEE Conference on Bioinformatics and Bioengineering, 321-327.
11. Wang,H. ,Wang,W., Haixun,W., Yu,P. (2002) Clustering by pattern similarity in large data sets, ACM SIGMOD International Conference on Management of Data, 394-405.
12. Lazzeroni,L., Owen, A. (2000) Plaid models for gene expression data, Technical Report Stanford University.
13. Shamir,R., Sharan,R., Tanay,A. (2002) Discovering statistically significant biclusters in gene expression data, Bioinformatics, vol. 19, Suppl. 1 2002, 136-144.
14. Liu,J., Yang, J., Wang,W. (2004) Biclustering in Gene Expression Data by Tendency, IEEE Computational Systems Bioinformatics Conference 2004, 183-193.
15. Ben-Dor,A., Chor,B., Karp,R., Yakhini,Z. (2002) Discovering local structure in gene expression data : The Order Preserving Submatrix Problem, 6th ACM International Conference on Research in Computational Molecular Biology, RECOMB2002.
16. Gerstein,M., Chang,J., Basri,R. ,Kluger,Y. (2003) Spectral Biclustering of Microarray Data : Coclustering Genes and Conditions, Journal Genome Research, vol. 13(4), 703-716.
17. Aguilar,J.S., Divina,F. (2005) Evolutionary Biclustering of Microarray Data, 3rd European Workshop on Evolutionary Bioinformatics.
18. J. Liu and W. Wang, ―Op-Cluster: Clustering by Tendency in High Dimensional Space,‖ Proc. Third IEEE Int'l Conf. Data Mining,p. 187-194, 2003.
19. J. Hipp, U. Gu¨ ntzer, and G. Nakhaeizadeh, ―Algorithms for Association Rule Mining—A General Survey and Comparison,‖ SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 58-64, 2000.
20. J. Pei, X. Zhang, M. Cho, H. Wang, and P.S. Yu, ―Maple: A Fast Algorithm for Maximal Pattern-Based Clustering,‖ Proc. Third IEEE Int'l Conf. Data Mining, p. 259-266, 2003.
21. S. C. Madeira and A. L. Oliveira, ―Biclustering algorithms for biological data analysis: A survey, IEEE Trans. Comput. Biol. Bioinform., vol. 1, no. 1, pp. 24–45, Jan.–Mar. 2004.
22. Ahmed H. Tewfik, Alain B. Tchagang, Laura Vertatschitsch ―Parallel Identification of Gene Biclusters With Coherent Evolutions‖ IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 6, JUNE 2006.
23. K. Sathishkumar, E. Balamurugan, P. Narendran "An Efficient Artificial Bee Colony and Fuzzy C Means Based Co-regulated Biclustering from Gene Expression Data" Mining Intelligence and Knowledge Exploration, Volume 8284 of the series Lecture Notes in Computer Science, p. 120-129, December 18-20, 2013.