An Efficient Artificial Bee Colony and Fuzzy C Means Based Co-regulated Biclustering from Gene Expression Data

K. Sathishkumar[1], E. Balamurugan[2], and P. Narendran[1]

[1] Gobi Arts & Science College, Gobichettipalayam
[2] Bannari Amman institute of Technology, Sathyamangalam
{Sathishmsc.vlp,rethinbs,narendranp}@gmail.com

Abstract. The gene microarray data are arranged based on the pattern of gene expression using various clustering algorithms and the dynamic natures of biological processes are generally unnoticed by the traditional clustering algorithms. To overcome the problems in gene expression analysis, novel algorithms for finding the coregulated clusters, dimensionality reduction and clustering have been proposed. The coregulated clusters are determined using biclustering algorithm, so it is called as coregulated biclusters. The coregulated biclusters are two or more genes which contain similarity features. The dimensionality reduction of microarray gene expression data is carried out using Locality Sensitive Discriminant Analysis (LSDA). To maintain bond between the neighborhoods in locality, LSDA is used and an efficient meta heuristic optimization algorithm called Artificial Bee Colony (ABC) using Fuzzy C Means clustering is used for clustering the gene expression based on the pattern. The experimental results shows that proposed algorithm achieve a higher clustering accuracy and takes lesser less clustering time when compared with existing algorithms.

Keywords: Gene expression data, Bimax Algorithm, Co-regulated Biclusters, Locality Sensitive Discriminant Analysis, Artificial Bee Colony, Fuzzy C Means.

1 Introduction

The purpose of clustering gene expression data is to reveal the natural structure inherent in the data. A good clustering algorithm should depend as little as possible on prior knowledge, for example requiring the predetermined number of clusters as an input parameter. Clustering algorithms for gene expression data should be capable of extracting useful information from noisy data. Gene expression data are often highly connected and may have intersecting and embedded patterns [1,2]. Therefore, algorithms for gene-based clustering should be able to handle this situation effectively. Finally, biologists are not only interested in the clusters of genes, but also in the relationships (i.e., closeness) among the clusters and their sub-clusters, and the relationship among the genes within a cluster (e.g., which gene can be considered as

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 120–129, 2013.

the representative of the cluster and which genes are at the boundary area of the cluster) [3, 4].

2 Related Works

K-means is a typical partition-based clustering algorithm used for clustering gene expression data. It divides the data into pre-defined number of clusters in order to optimize a predefined criterion. The major advantages of it are its simplicity and speed, which allows it to run on large datasets [5]. However, it may not yield the same result with each run of the algorithm. Often, it can be found incapable of handling outliers and is not suitable to detect clusters of arbitrary shapes. Self Organizing Map (SOM) is more robust than K-means for clustering noisy data. It requires the number of clusters and the grid layout of the neuron map as user input. Specifying the number of clusters in advance is difficult in case of gene expression data [6].

K-nearest neighbor based density estimation technique is proposed [7]. Another density based algorithm proposed by Chung et al. works in three phases: density estimation for each gene, rough clustering using core genes and cluster refinement using border genes. A density and shared nearest neighbor based clustering method is presented [8]. The similarity measure used is that of Pearson's correlation and the density of a gene is given by the sum of its similarities with its neighbors. The use of shared nearest neighbor measure is justified by the fact that the presence of shared neighbors between two dense genes means that the density around the dense genes is similar and hence should be included in the same cluster along with their neighbors.

Fuzzy C-means (FCM) is an extension of K-means clustering and bases the fuzzy assignment of an object to a cluster on the relative distance between the object and all cluster centroids. Many variants of FCM have been proposed in the past years, including a fuzzy clustering approach, FLAME [9], which detects dataset-specific structures by defining neighborhood relations and then neighborhood approximation of fuzzy memberships are used so that non-globular and nonlinear clusters are also captured.

3 Methodology

The proposed approach consists of three stages namely finding of coregulated biclusters using Bimax algorithm, dimensionality reduction using Locality Sensitive Discriminant Analysis (LSDA) and clustering using MoABC.

3.1 Finding of Coregulated Clusters Using Bimax Algorithm

The diagram for finding the coregulated clusters using algorithm is shown in Fig.1. Enhanced Bimax algorithm is used to display a maximal biclusters value and displays

a coregulated biclusters. The Enhanced Bimax algorithm is used to measure a particular gene is present or not. It also finds the transcription sites of the coregulated biclusters. Normalization technique used to specify genes are presented in the particular group or not. The output is display the transcription factors.

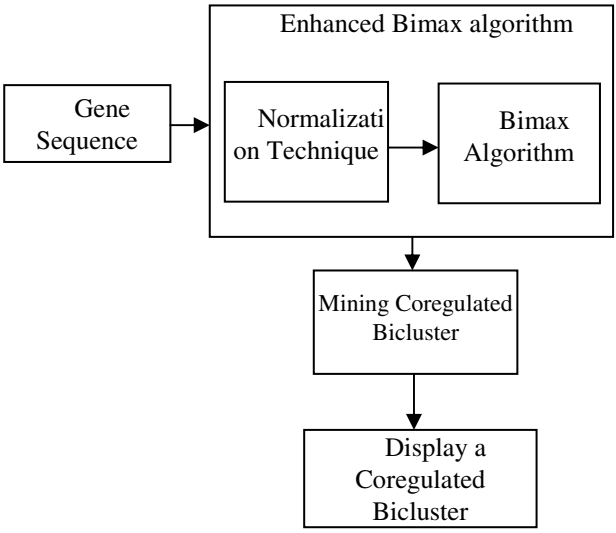

Fig. 1. Block diagram for mining coregulated bicluster

3.2 Bimax Algorithm

The Bimax algorithm needs to guarantee that only optimal, inclusion-maximal biclusters are generated. It is used to specify the genes and conditions. It is used to specify that 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 coregulated genes are grouping together the particular expression value and the particular situation [13].

3.3 Proposed Enhanced Bimax Algorithm

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

Binary Space Partitioning (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 unnormalized data and normalized data are displayed on the coregulated 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, Breadth-First Search (BFS) and BSP [13].

They are BFS and BPS combination of sequences searches the entire graph c nodes of a graph or rods, it exhaustively y Space Partitioning. First step is normalization technique used to remove the redundant data and then 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 stop 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. Otherwise the gene is not present in the given group then it is representing as zero. Fig. 2 describes a proposed Enhanced Bimax algorithm.

3.4 Locality Sensitive Discriminant Objective Function for Dimensionality Reduction.

It is observed that naturally occurring data may be generated by structured systems with possibly much fewer degrees of freedom than the ambient dimension would suggest, a number of research works have been developed with the case considering when the data lives on or close to a submanifold of the ambient space [10].

3.5 Proposed Artificial Bees Colony Based Fuzzy Clustering

The modifications carried out to improve the basic ABC algorithm and its application used to achieve fuzzy clustering is been given in this section.

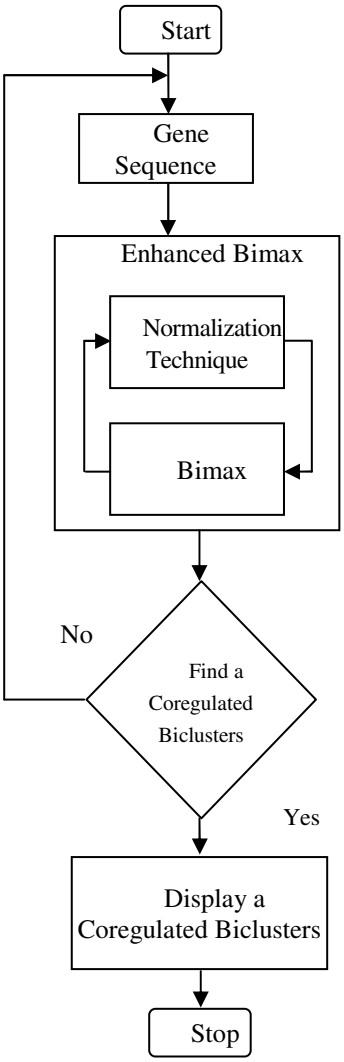

Fig. 2. Flow Chart for Proposed Enhanced Bimax Algorithm

3.6 Locality Sensitive Discriminant Analysis

A novel linear dimensionality reduction algorithm called Locality Sensitive Discriminant Analysis (LSDA). For the class of spectrally based dimensionality reduction techniques, it optimizes a fundamentally different criterion compared to classical dimensionality reduction approaches based on Fisher's criterion (LDA) or Principal Component Analysis.

3.7 Fuzzy C-Means Clustering (FCM)

FCM is a clustering algorithm which allows one data may belong to two or more clusters. It is normally used in pattern recognition [11]. It is based on minimization of the following objective function (1):

$$j_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, 1 \le m < \infty \quad (1)$$

Where,

$m =$ is any real number greater than 1

$u_{ij} =$ is the degree of membership of xi in the cluster j

x_i is the ith of d-dimensional data

c_j is the cluster centre of d-dimension data

$\left\|x_i - c_j\right\|^2$ is the distance measured of similarity between the measured data and the cluster data

3.8 Artificial Bee Colony Algorithm

It is a swarm intelligent method which inspired from the intelligent foraging behavior of honey bee swarms. It strength is its robustness and its simplicity. It is developed by surveying the behavior of the bees is finding the food source which is called nectar and sharing the information of food source the bee which is present in the nest. In the ABC the artificial agents are classify into three types; such as employed bee, the onlooker bee and the scout each of the bee plays different role in the process. The employed bee stays on a food source and in its memory provides the neighborhood of the food source. Each employed bee carries with her information about the food source and shares the information to onlooker bee. The onlooker bees wait in the hive on the dance area, after getting the information from employed bees about the possible food source then make decision to choose a food source in order to use it. The onlooker bees select the food source according to the probability of that food source. The food source with lower quantity of nectar that attracts less onlooker bees compared to ones with a higher quantity of nectar. Scout bees are searching randomly for a new solution. The employed bee whose food source has been abandoned it becomes a scout bee. The goal of the bees in the ABC model is to find the best solution. In the ABC algorithm the number of employed bees is equal to the number of onlooker bees which is also equal to the number of solutions. The ABC algorithm consist of a Maximum Cycle Number (MCN) during each cycle, there are three main parts:

- Sending the employed bees to the food sources and calculate their nectar quantities
- Selecting the food sources by the onlooker bees
- Determining the scout bee and discover a new possible food sources

Employed Bee

In the employed bee phase, each employed bee determines a new solution from the neighborhood of the current food source.

The employed bee compared the current solution with the new solution and memorizes the best one by apply the greedy selection process. When all employed bees have finished this search process, then they share the fitness value (nectar information) and the position of the food source (solution) to the onlooker bees.

Onlooker Bee

In the onlooker bee phase, after getting the information about the nectar and position of the food source each onlooker bee selects a food source with a probability of higher nectar information.

Scout Bee

If a food source position cannot be improved through fixed cycles, it is called 'limit', it means that the solution has been sufficiently exploited, and it may be removed from the population.

MoABC Based FCM

In order to perform fuzzy clustering for image segmentation using the proposed MoABC-FCM algorithm, a population of SN $(z_1, z_2, z_3 \ldots . z_{SN})$ solutions is created, where SN is the number of employed bees or onlooker bees. Each bee represents a potential solution of the fuzzy clustering problem. Each individual bee z_i in generation G is formulated using equation (2):

$$z_i(G) = (v_{i,1}, v_{i,2}, \ldots \ldots v_{i,c})^T, \ subject\ to\ 1 \leq i \leq SN \tag{2}$$

C is the number of clusters and, $v_{i,k}$ represents the kth cluster center for the ith bee.

The goal of MoABC-FCM algorithm is to determine when algorithm gets into convergence; it is converted into the optimal fuzzy partition matrix to a crisp partition matrix. The defuzzification is carried out by assigning each pixel to the cluster with the highest membership.

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, the microarray gene samples of human acute leukemia and colon cancer data are

utilized. [12] 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 LSDA is applied to identify informative genes and reduce gene dimensionality for clustering samples to detect their phenotypes.

Table 1. 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 2. Clustering for microarray gene expression data whose amount is large can be fully calculated by determining the boundary of the clusters.

Table 2. A sample of the microarray gene data to test the proposed technique

Class	ALL		AML		COLON	
Sample gene	ALL 16125 TA-Norel	ALL 23668 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. Existing clustering algorithms, such as Fuzzy C-means and Fuzzy Possibilistic C-Means Algorithm using EM Algorithm approaches and also MoABC 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 3. Performance comparison in percentage between the proposed MoABC clustering technique and other existing Techniques

Type of Gene Data	Accuracy				Correlation				Distance				Error Rate			
	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC
ALL	83.1	83.9	85.69	87.25	0.345	0.368	0.412	0.4852	0.00379	0.00346	0.00263	0.00142	0.21	0.20	0.18	0.12
AML	80.06	81.02	83.84	85.12	0.024	0.029	0.0315	0.0396	0.00364	0.00331	0.00201	0.00185	0.30	0.29	0.24	0.16
COLON	79.0	79.9	81.96	83.04	0.119	0.125	0.139	0.215	0.02029	0.02011	0.0126	0.0099	0.04	0.03	0.01	0.006

From the Table 3, it can be seen that the proposed technique MoABC 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.

5 Conclusion

Genes involved in multiple biological processes (simultaneously) may play a major role in one process while playing a minor role in another process. The importance of a gene in multiple processes has potential for further investigation. . In this paper, an effective microarray gene data clustering technique has been proposed with the aid of Bimax algorithm, LSDA and MoABC. Initially, the micro array data genes are given to the Bimax algorithm to find coregulated biclusters to reduce the space complexity of the genes, and then the dimensionality of the microarray data has been reduced with the help of LSDA 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. The experimental results show that proposed algorithm achieved better improvement in the quality of the results by using MoABC. Hence, this means of gene clustering have paved the way for effective information retrieval in the microarray gene expression data.

References

1. Belcastro, V., Gregoretti, F., Siciliano, V., Santoro, M., D'Angelo, G., Oliva, G., di Bernardo, D.: Reverse Engineering and Analysis of Genome-Wide Gene Regulatory Networks from Gene Expression Profiles Using High-Performance Computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(3), 668–678 (2012)
2. Yuan, Y., Li, C.-T.: Partial Mixture Model for Tight Clustering in Exploratory Gene Expression Analysis. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007 (2007)
3. Yin, L., Huang, C.-H.: Clustering of Gene Expression Data: Performance and Similarity Analysis. In: First International Multi-Symposiums on Computer and Computational Sciences, IMSCCS 2006 (2006)
4. Jiang, D., Pei, J., Zhang, A.: "DHC: a density-based hierarchical clustering method for time series gene expression data'. In: Proceedings of the 3rd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, USA, p. 393 (2003)
5. Dhiraj, K., Rath, S.K., Pandey, A.: Gene Expression Analysis Using Clustering. In: 3rd International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2009 (2009)
6. Yano, N., Kotani, M.: Clustering gene expression data using self-organizing maps and k-means clustering. In: SICE 2003 Annual Conference, vol. 3, pp. 3211–3215 (2003)
7. Chung, S., Jun, J., McLeod, D.: Mining geneexpression datasets using density based clustering. Technical Report, USC/IMSC, University of Southern California, No. IMSC-04-002 (2004)
8. Syamala, R., Abidin, T., Perrizo, W.: Clustering Microarray Data based on Density and Shared Nearest Neighbor Measure. In: Proceedings of the 21st ISCA International Conference on Computers and Their Applications (CATA 2006), pp. 23–25 (2006)
9. Fu, L., Medico, E.: FLAME: A novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8(3) (2007)
10. Cai, D., He, X., Zhou, K., Han, J., Bao, H.: Locality Sensitive Discriminant Analysis (2007)
11. Geng, X., Tao, F.: GNRFCM: A new fuzzy clustering algorithm and its application. In: International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII (2012)
12. Wen, J.: Ontology Based Clustering for Improving Genomic IR. Twentieth IEEE International Symposium International Journal of Data Mining and Bioinformatics 3(3), 229–259 (2009)
13. Chandran, C.P., IswaryaLakshmi, K.: Biclustering analysis of coregulatedbiclusters from gene expression data. International Journal of Computational Intelligence and Informatics 2(1) (2012)