International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Study on Performance Measure of Statistically Significant Gene Expression Data Using Biclustering Algorithms

M.Ramkumar[1], Dr.R.P.Singh[2], Dr.K.Vengatesan[3], B.Narmadha[4]

[1]*(Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, India)*
[2]*(Research Guide, Sri Satya Sai University of Technology and Medical Sciences, India)*
[3]*(Associate Professor, Sanjivani College of Engineering, India)*
[4]*(Assistant Professor, KIOT, Tamilnadu, India)*

ABSTRACT

This paper discussed about various Biclustering methods used in gene expression data to find the significant. Gene expression matrices have been extensively analyzed in two dimensions: The analysis of gene expression can be performed either according to different expression conditions, or gene dimension. These analysis correspond, respectively, to analyze the expression patterns of genes by comparing the rows in the matrix, and to analyze the expression patterns of samples by comparing the columns in the matrix.

Keywords: Biclustering, Gene Expression Data Microarray, Clustering

I.INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

II.RELATED WORK

This section discusses various works carried out by existing researchers on data mining techniques, gene expression data, microarray, statistical methods used for measure the similarity, the advantages and limitations of existing clustering techniques, different data types and data repositories which are used for mining knowledge. The cluster is group of object one with another based on the similarity between the objects. The correlation is calculated from the micro array gene expression data to form the cluster. The performance of each work is compared with existing work. Rapid retrieval of significant information from the databases has always been an important issue. Different techniques have been developed for this purpose; one of them is Data Clustering. Data clustering is methods by which clusters were made that are one way or another similar in characteristics. Clustering in computer science means unsupervised classification of data objects into different groups. It can also be referred to as partitioning of a data set into different subsets. Each data object in the subset ideally shares some common character. One of the goals of microarray data analysis is to cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic of the group. Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater the difference between groups the better or more distinct the resulting clusters.

III.BICLUSTERING TYPES

An interesting criterion to evaluate a biclustering algorithm concerns the identification of the type of biclusters the algorithm is able to find. We identified four major classes:

1. Biclusters with constant values.
2. Biclusters with constant values on rows or columns.
3. Biclusters with coherent values.
4. Biclusters with coherent evolutions.

A number of scenarios that occur in microarray experiments are not catered for by all clustering techniques and this should be taken into account when selecting a method for analysis. First, a gene may be involved in more than one biological process and may exhibit an expression profile that is a result of the regulatory effect of each process. If there are other genes that are involved in some subset of these processes, the structure should be represented by overlapping clusters. Second, a group of genes may be coexpressed under limited conditions. Several clustering methods have been developed in recent years that cater to one or more of these scenarios. These include gene-shaving , context-specific Bayesian clustering , EMMIX-GENE , interrelated two-way clustering , simultaneous clustering , coupled two way clustering, rich probabilistic models , double conjugated clustering , SAMBA, order preserving sub matrix clustering , biclustering and the plaid model . The plaid model

is one method that accommodates all the scenarios described earlier and is particularly attractive as it uses continuous gene expression levels and estimates the "usual" expression level for each gene (in the context of the data set), so that biclusters of an unusual expression pattern can be discovered. Furthermore, as a model-based clustering method, the plaid model can be naturally extended to appropriately analyze structured microarray experiments which are the focus of interest in this paper.

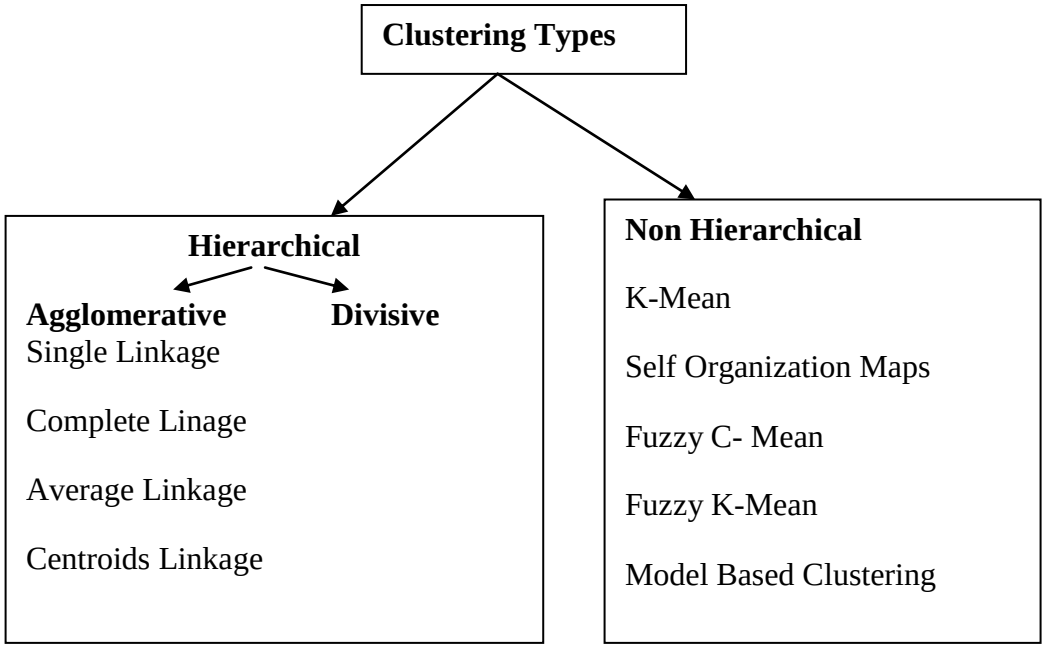

Figure 1: Clustering methods classified into hierarchical and non- hierarchical

Clustering methods can be hierarchical (grouping objects into clusters and specifying relationships, among objects in a cluster, resembling a phylogenetic tree) or non-hierarchical (grouping into clusters without specifying relationships between objects in a cluster). One way of describing clustering structures with a cluster label assigned to each object is called classification (supervised learning). Then the people perhaps, to do a search on a specific label to find out which objects belong to it. However, there are other ways of describing the discovered Structure and this depends on the clustering paradigm being followed. These paradigms reflect the different assumptions and approaches taken by researchers in the field. The figure 1 represents clustering methods classification.

What is then the difference between clustering and biclustering? Why and when should we use biclustering instead of clustering? Clustering can be applied to either the rows or the columns of the data matrix, separately. Biclustering, on the other hand, performs clustering in these two dimensions simultaneously. This means that clustering derives a global model while biclustering produces a local model. When clustering algorithms are used, each gene in a given gene cluster is defined using all the conditions. Similarly, each condition in a condition cluster is characterized by the activity of all the genes that belong to it. However, each gene in a

bicluster is selected using only a subset of the conditions and each condition in a bicluster is selected using only a subset of the genes. The goal of biclustering techniques is thus to identify subgroups of genes and subgroups of conditions, by performing simultaneous clustering of both rows and columns of the gene expression matrix, instead of clustering these two dimensions separately. We can then conclude that, unlike clustering algorithms, biclustering algorithms identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. Therefore, biclustering approaches are the key technique to use when one or more of the following situations apply:

1. Only a small set of the genes participates in a cellular process of interest.
2. An interesting cellular process is active only in a subset of the conditions.
3. A single gene may participate in multiple pathways that may or not be coactive under all conditions.

For these reasons, biclustering should identify groups of genes and conditions, obeying the following restrictions:

1. A cluster of genes should be defined with respect to only a subset of the conditions.
2. A cluster of conditions should be defined with respect to only a subset of the genes.
3. The clusters should not be exclusive and/or exhaustive: A gene/condition should be able to belong to more than one cluster or to no cluster at all and be grouped using a subset of conditions/genes.

First, we consider microarray experiments for which an a priori group structure is available for the genes or samples. In this case, we partially supervise the plaid model algorithm to favor biclusters that correspond to the external grouping, so that biclusters can be interpreted as features relating to one or more a priori groups. We compare the results of a partially supervised analysis to the results of an unsupervised analysis for an experiment investigating forms of tuberculosis. Second, we consider microarray experiments in which the expression levels of a set of genes are measured over time for several samples. For this type of data, we extend the plaid model so that instead of clustering single expression levels, whole time series of expression levels are clustered. This allows complete three-way microarray data sets to be analyzed, obviating the need for collapsing such data sets to a two-way data structure, which in some cases can lead to a substantial loss of information

IV.PLAID MODEL

The plaid model can be fully supervised by clustering complete a priori groups instead of individual genes or samples. However, we would like to allow for misclassification, experimental error, and the presence of biclusters in the data that are unrelated to the external grouping. Therefore, we do not consider a fully supervised approach to be appropriate. Partial supervision is preferable and can be implemented by using a supervised model to start the search for a layer, reverting to the unsupervised model after a set number of iterations.

For the binary least squares algorithm, we prefer to address the problem of inconsistent biclusters by the use of search models, simplified models that represent the features of a bicluster that are considered to be most

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

important. We propose that the search model is used within the layer iterations, and then the full plaid model is fitted before pruning the bicluster and back fitting. It is not appropriate to use a simplified model throughout the algorithm as we expect to see gene and sample effects in practice and do not wish to treat these effects in the same way as pure error.

V.EEW-SC ALGORITHM

An EEW-SC (Enhanced Entropy Weighting Subspace Clustering) technique used to measure similarity between the pair of gene in high dimensional gene expression, that can be written as equ (1).

$$EEW\text{-}SC = \min\left\{\min_i\left(\frac{c\sum_{k=1}^{D} w_{ik}(x_{ik}-v_{ik})^2}{\sum_{k=1}^{D} w_{ik}(v_{ik}-v_{ok})^2}\right), n\right\} \quad (1)$$

Where, V= {v1, v2 …vc} is the cluster center matrix and W= {w1, w2…wc} is the weight matrix.

VI.SAMA

This model is used to represent the biclusters which are uniformly related with another element which is expressed by a log function. In which weight of each edge (u, v) to $\log \frac{1-p_c}{1-p_{u,v}} > 0$ and the weight of the each non-edge (u, v) to $\log \frac{1-p_c}{1-p_{u,v}} < 0$, then finally it is assumed that the score of H is weight. In another alternative model, each edge of a Bicluster occurs with constant probability $p_c > max_{(u,v)} U*V\ p_{u,v}$ the SAMA algorithm describes p_c.

$$Log\ L(H) = \sum_{(u,v)\xi E} \log \frac{p_c}{p_{u,v}} + \sum_{(u,v)\xi E} \log \frac{1-p_c}{1-p_{u,v}} \quad (3)$$

VII.CLICK

The probabilistic mechanism used to identify the highly connected components in a pair of elements, in which techniques used Cluster Identification via Connectivity Kernels. The pair of similarity between the elements are normally distributed and assume, the weight wij of an edge (i, j) is defined as the probability that vertices i and j

are in the same cluster. The proximity graph also constructed using CLICK. The CLICK follows iterative method to find the minimum cut in the in the proximity graph and recursively divide the data set into a set of associated mechanism from the minimum cut.

VIII.BICLUSTERING ISA

The Iterative Signature Algorithm is a novel method used to find the correlated biclusters .The rows and columns are may be belong to multiple modules, which are overlapped with one another. Starting with preliminary set of genes, all samples (conditions) are scored with reverence to this gene set and those samples are selected for which the score exceeds a confident threshold. In the similar way, all genes are scored concerning the chosen samples and a new set of genes is preferred based on another user distinct threshold. The entire process is continual until the set of genes and the set of samples congregate and do not change anymore.

IX.BIMAX

Main objective of BiMax algorithm is to find maximum biclusters from the data set. In which algorithms either row are added with another row or row added with column to partition the group based on the divide and conquer technique. Consider C_U, C_V is two set of column partitions and G_U, G_W are set of two row partitions. A submatix U is derived from ($G_U U G_W$,C_U) and V is derived from submatix $G_W U G_V$,C_V. The main operation of BiMax algorithm is generation of inclusion-maximal bicluster represented in figure 2.

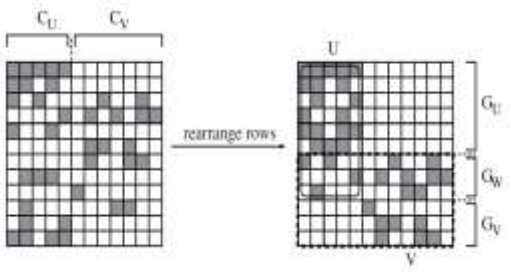

Figure 2: Formation of Bicluster using BiMax

This can be done by iterating these steps such as rearrange the rows and columns to concentrate ones in the upper right of the matrix, divide the matrix into two sub matrices and whenever in one of the sub matrices only ones are found, this sub matrix is returned.

X.CONCLUSION

The gene expression data analysis is important filed in biological area, which follows different techniques to group the related genes. Currently, clustering is such a utensil broadly used in gene expression data analysis to achieve biological information. A primary aim of such an analysis will be the detection of groups of genes that demonstrate similar expression patterns that can smooth the progress of the scientist to conduct suitable diagnosis and conduct of patients. The clustering techniques are very useful to find the co-regulated gene, in

which each method has own limitations and predefined number of clusters. To defeat this, a new Biclustering method will develop and experience with the pattern recognition.

REFERENCES

[1.] Alter O., Brown P.O. and Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA, Vol. 97(18):10101-10106, August 2009.

[2.] Ding, Chris. Analysis of gene expression profiles: class discovery and leaf ordering. In Proc. of International Conference on Computational Molecular Biology (RECOMB), pages 127-136, Washington, DC, April 2012.

[3.] Danasingh Asir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany Jebamalar Leavline, "Improving the Accuracy of the Supervised Learners using Unsupervised based Variable Selection", Asian Journal of Information Technology, 13.9 (2014): 530-537.

[4.] Arifovic, Jasmina, "Genetic algorithm learning and the cobweb model", Journal of Economic dynamics and Control, 18.1 (2010): 3- 28.

[5.] Hommes, Cars H, "On the consistency of backward-looking expectations: The case of the cobweb", Journal of Economic Behavior & Organization, 33.3 (2011): 333-362.

[6.] Alejos, Óscar, and Edward Della Torre, "The generalized cobweb method", Magnetics, IEEE Transactions on 41, 5 (2005): 1552-1555.

[7.] Zhao, Zhechong, and Lei Wu, "Stability analysis for power systems with pricebased demand response via Cobweb Plot", Proc. IEEE PES General Meeting, 2013.

[8.] Yuni Xia,Bowei Xi, "Conceptual Clustering Categorical Data with Uncertainty", 19th IEEE International Conference on Tools with Artificial Intelligence.

[9.] Moon, Tood K, "The expectation-maximization algorithm", Signal processing magazine, IEEE 13, 6 (2011): 47-60.

[10.] Brankov, Jovan G., et al. "Similarity based clustering using the expectation maximization algorithm", Image Processing, 2002, Proceedings, 2002 International Conference, Vol. 1, IEEE, 2002.

[11.] Lagendijk, Reginald L., Jan Biemond, and Dick E. Boekee, "Identification and restoration of noisy blurred images using the expectation-maximization algorithm", IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], 38 (7) (2011).

[12.] Vengatesan K., and S. Selvarajan: Improved T-Cluster based scheme for combination gene scale expression data. International Conference on Radar, Communication and Computing (ICRCC), pp. 131-136. IEEE (2012).

[13.] Kalaivanan M., and K. Vengatesan.: Recommendation system based on statistical analysis of ranking from user. International Conferenceon Information Communication and Embedded Systems (ICICES), pp.479-484, IEEE, (2013).

[14.] K.Vengatesan, S. Selvarajan: The performance Analysis of Microarray Data using Occurrence Clustering. International Journal of Mathematical Science and Engineering, Vol.3 (2) ,pp 69-75 (2014).