ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A STUDY ON MICROARRAY GENE EXPRESSION DATA AND CLUSTERING ANALYSIS

K.Sathishkumar*, Dr.V.Thiagarasu, E.Balamurugan
ASSISTANT PROFESSOR, DEPT. OF INFORMATION TECHNOLOGY. GOBI ARTS & SCIENCE COLLEGE (AUTONOMOUS), GOBICHETTIPALAYAM,INDIA
ASSOCIATE PROFESSOR, DEPT. OF COMPUTER SCIENCE, GOBI ARTS & SCIENCE COLLEGE(AUTONOMOUS), GOBICHETTIPALAYAM,INDIA
ASSOCIATE PROFESSOR, DEPARTMENT OF (IT & IS) BLUECREST COLLEGE, ACCRA-NORTH, GHANA

ABSTRACT

After genome sequencing, DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Biologists are hugely weighed down by the massive amount of unparalleled qualities of genome-wide data produced by the DNA Microarray experiment.Clustering is the process of grouping data objects into set of disjoint classes called clusters so that objects within a class are highly similar with one another and dissimilar with the objects in other classes.Clustering is one of most useful tools for the microarray gene expression data analysis. The remote method provides a divide-and–conquer strategy to extract information from expression shape. In this paper, the highlighted procedures followed by different categories provide a structure for further development with better understanding.

Keywords: Microarray, Clustering, Gene expression data, Biclustering

INTRODUCTION

Microarray is a comparatively new technology and a chip-based high throughput technology to investigate the expression levels of thousands of genes simultaneously, compared with the traditional approach to genomic research. The eminence of DNA microarray technology is the aptitude to be used to simultaneously monitor and study the expression levels of thousands of genes, relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes [1]. Functional genomics can be better implicit when the veiled patterns in gene expression data is elucidated, however, it is very challenging to comprehend and construe this due to the complexity of biological networks and large number of genes.

The most important area of microarray bioinformatics is possibly the data clustering analysis. Clustering is an exceptional preference for initial data analysis and data mining processes. To perceive and identify appealing patterns of expression across multiple genes and experiments, reveal natural structures and compress high-dimensional array data clustering must be ascertained to allow easier management of data set. This data reduction method is a simple tool yet powerful method of organising genes based on their interdependence behaving similarly over the different conditions in different mutants, patients or at different time points in a time series during an experiment with similar expression patterns and properties into a set of disjoint groups based on specific features so that the underlying structures can be acknowledged and explored.

PROCEDURES OF CLUSTERING ANALYSIS

The procedures of cluster analysis are the feature selection, cluster algorithm selection, cluster validation and result interpretation [4] [5]. The intimately connected steps of cluster analysis with feedback pathways are shown in the following Figure1.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

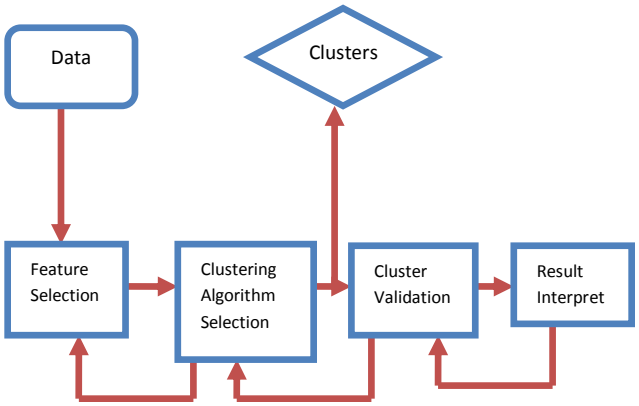

Figure 1: Cluster Analysis

A. Feature selection

Microarray experiments provide a expression information of large number of genes (from 103 to 104 or105). It is essential to consider which feature (gene) subset will be employed in clustering analysis, by eliminating the least interesting and highlight the most interesting genes. Distinctive features from a set of candidates are neatly selected, while feature extraction exploits some alteration to produce useful and novel features from the original ones which are very essential to the efficiency of clustering purpose [5].

B. Cluster algorithm design or selection

Different clustering algorithms and methods have been developed to improve the preceding ones, unravelling the problems and fit for specific fields [6]. There is no absolute clustering method that can be universally used to solve all problems. So in order to select or generate a suitable clustering strategy, it is vital to investigate the features of the problem.

As Xu and Wunsch revealed the step is usually combined with the selection of a corresponding proximity measure and the construction of a criterion function [5]. Patterns are grouped according to whether they resemble each other. Once a proximity measure is chosen, the construction of a clustering criterion function makes the partition of clusters an optimizing problem.

C. Clustering validation

Finding the number of clusters in a dataset and many of this method have been proposed some of which are silhouette index, Dunn's index, and Davis-Bouldin index for gene expression data which evaluates the partitions generated using clustering algorithm and find the pre-eminent partition based on intra-cluster and intercluster distance. It is vital to evaluate diverse clustering outcome, the quality and reliability of clusters before deciding on the finest data distribution [7].

D. Result explanation

Assessing the results and interpreting the clusters found are as significant as generating the clusters [8]. The objective of clustering is to solve the encountered problem efficiently and offer the users with significant understanding of their original data.

ANALYSIS OF CLUSTERING TECHNIQUES

One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. On one hand, co-expressed genes can be grouped in clusters based on their expression patterns [2, 3]. Some clustering algorithms, such as Distance metric, hierarchical Clustering and K-means can be used both to group genes and to partition samplesboth the gene-based and sample-based clustering approaches search exclusive and exhaustive partitions of objects that share the same feature space.

Distance metric

In order to group together similar objects, the meaning and measure of similarity has to be defined which is referred to as the distance metric which clustering is highly dependent on. A distant metric is a function that takes two points in a dimensional space where this should be symmetrical, positive and triangle unequal [9].

To calculate the distance between clusters different distance linkages are involved which affects the complexity and performance of the clustering. Single linkage calculates the distance between the closest neighbors. Complete linkage calculates the distance between the furthest neighbors. Centriod linkage defines the distance between two clusters.

Average linkage measures the average distance between members of different clusters. Average and complete linkage are the preferred methods for microarray data analysis [10].

Many diverse clustering techniques have extensively been under development [6]. The most widely used techniques in analysis of gene expression data which are applied in the early stages and proven to be useful are Hierarchical clustering [10], K-means clustering [8] and Self-organized maps (SOM) [11].

1. *Hierarchical clustering*

Hierarchical clustering is the first and most common clustering method applied to gene expression data which is developed on the basis of a single layered neural network. A hierarchical series of nested clusters are generated by grouping genes with similar pattern of expression across a range of samples located near each other. Hierarchical clustering calculates all pairs-wise distance relationships between genes and experiments to merge pairs of values that are most similar for the formation of a node [1]. The inter-cluster distance groups together these clusters to make a higher level cluster which can be graphically illustrated by a tree, called dendrogram representing the clusters and relationship between them. This is repeated, comparing genes or new clusters until all clusters are joined. These methods are either agglomerative algorithms (bottom-up approach) which joins clusters in a hierarchical manner or the more rapid dividing hierarchically.

The drawbacks of this method are its high computational intricacy, lack of robustness, vagueness of termination criteria and failure with large number of genes as data sets grow in complexity.

2. *K-means Clustering*

K-means clustering is a simple and fast method used commonly due to its straightforward implementation and small number of iterations. This algorithm divides the data set into k disjoint subsets. An estimation of the number of clusters (*k*) is made by the user and calculated as an input where the computer randomly assigns each gene to one of the k clusters [11].

The distance between each gene and the centre of each cluster is promptly calculated resulting in an optimal grouping of data to clusters where the genes inside every cluster are as close to the centre of the cluster as possible while at the same time there is maximal distance between genes of different clusters. This method is useful if different values of *k* are attempted and it only gives the number of clusters not the relationship between them like hierarchical clustering.

The drawbacks of this method are the lack of prior knowledge of the number of gene clusters in a gene expression data which results in the changing of results in the altering of results in successive runs since the initial clusters are selected randomly and the quality of the attained clustering has to be assessed.

3. *Self-Organized Maps Clustering (SOM)*

SOM is a reasonably fast and easy to implement method, scalable to large data sets. It is intimately related to multidimensional scaling and its objective is to represent all data points in the source space by points in a target space where the distance and proximity relationships are preserved. At the input, the data objects are presented and output neurons are organized with a sample neighborhood grid structure [12].

The remarkable features of SOM is that it generates an intuitively appealing map of a high dimensional data set and places similar clusters near each other so that the neighbouring clusters in this grid are more related than clusters that are not neighbours. SOM is trained through competitive learning for the distribution of the input data set which provides a relatively robust approach than k-means in the clustering of highly noisy data. However SOM requires

users to input the number of clusters and the grid structure of the neuron map. After the completion of the training, clusters are identified by mapping all data points to the output neurons.

The drawbacks of this method is that it is not effective since the main interesting patterns may be merged into only one or two clusters and cannot be identified.

Table1: Clustering Techniques and Its Limitations

S.No	Cluster Techniques	Author's Name	Functions	Limitations
1	Hierarchical Clustering	M.B. Eisen and P.O. Brown(1999)	calculates all pairs-wise distance relationships between genes	High computational intricacy, lack of robustness, vagueness of termination criteria
2	K-means Clustering	S. Tavazoie.et al.,(1999)	straightforward implementation and small number of iterations	Lack of prior knowledge of the number of gene clusters, Depend on Random values.
3	Self-Organized Maps Clustering (SOM)	T. Kohonen(1995)	scalable to large data sets, high dimensional data set	patterns may be merged into only one or two clusters and cannot be identified

The above Table1 shows the traditional clustering techniques and limitations based on the criteria.

MICROARRAY DATA CLUSTERING ANALYSIS

Clustering gene expression data can be categorized into the three groups, 1) gene-based, 2) sample-based and 3) subspace clustering as both genes and samples is required to be clustered significantly.

1. *Gene-based clustering*

In such gene-based clustering, the genes are treated as the objects, while the samples are the features. Clustering algorithms for gene expression data should be competent of extracting useful information from a high level of background noise [13].

2. *Sample-based clustering*

To find the substructure of the sample, regards the samples as the objects and the genes as the features. Samples are generally related to various disease or drug effects within a gene expression matrix. Only a small subset of genes whose expression levels strongly correlate with the class distinction, rise and fall coherently and exhibiting fluctuation of a similar shape under a subset of conditions, called the informative genes that participate in any cellular process relevant. The remaining genes are regarded as noise in the data as they are irrelevant to the sample of interest. By focusing on a subset genes and conditions of interest, the noise levels induced by other genes and conditions can be lowered which is characterized by co-clustering [14].

3. *Subspace clustering*

To find subset of objects such that the objects emerge as a cluster in a subspace created by a subset of the feature [13]. Genes and samples are treated symmetrically such that either genes or samples can be regarded as objects or features. A single gene may participate in multiple pathways that may or may not be coactive under all conditions Subspace clustering [15] techniques confine coherence exhibit by the blocks within gene expression matrices. A block is a sub-matrix defined by a subset of genes on a subset of samples.

4. *Biclustering*

Biclustering performs simultaneous clustering on the row and column dimension of the data matrix where the gene exhibits highly correlated activities for every condition instead of clustering these two dimensions separately which distinct classes of clustering algorithms that perform simultaneous row-column clustering to identify sub matrices, subgroups of genes and subgroups of conditions. Clustering derives a global model while Biclustering produces a local model [16][17]. Unlike clustering algorithms, Biclustering algorithms identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions, each gene and condition in a

bicluster are only a subset of the gene and condition. In biclustering, if some points are similar in several dimensions they will be clustered together in that subspace proved of great value of finding the interesting patterns in the microarray expression data.

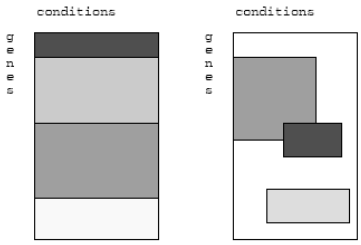

Figure 2. Clustering and Biclustering

Figure 2. Left: Traditional clustering searches a partition of all genes into *k* disjoint groups. Right: Biclustering searches for one or a set of blocks containing a consistent local pattern. Three biclusters are shown. (Note that it is not generally possible to display several biclusters at the same time as contiguous blocks.)

CONCLUSION

Clustering methods are rather effortless to implement and have a reasonable computational complexity and the performance may vary significantly with diverse data sets but there is no absolute finest algorithm in it. Some of the major weaknesses are the calculations in time, variations of densities in the data space resulting in overlapping clusters, cluster validation, presence of irrelevant attributes, Accuracy, high level of background noise, no prior knowledge and the dimensionality curse. To overcome the problems, these new developed approaches can be used to depict the primary structure of the genetic network. Finally, Biclustering algorithm has demonstrated the significant improvement, its weakness and inadequacy of clustering algorithms, but there's always scope for improvement.

REFERENCES

1. M.B. Eisen and P.O. Brown, "DNA arrays for analysis of gene expression", Methods Enzymol, vol. 303, pp. 170-205, P.O. 1999.
2. Ben-Dor A., Shamir R. and Yakhini Z. Clustering gene expression patterns. Journal of Computational Biology, 6(3/4):281–297, 1999.
3. Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David . Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95(25):14863–14868, December 1998.
4. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield D.D., and Lander E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, Vol. 286(15):531–537, October 1999.
5. R. Xu and D. Wunsch, "Survey of clustering Algorithms", IEEE Trans on Neural Networks. Vol. 16, no. 3, pp.645-678, 2005.
6. B. Everitt, "Cluster analysis" 1st ed. Heinemann, London, 1980.
7. N. Bolshakova, F. Azuaje, "Cluster validation techniques for genome expression data", Signal processing, 83, pp.825-833, 2003.
8. A.K. Jane and R.C. Dubes, "Algorithm of clustering data",Prentice Hall, Englewood Cliff, NJ, 1998.
9. E.Shay,"Microarray cluster analysis and applications", Available at: http://www.science.co.il/enuka/Essays/Microarray-Review.pdf, Jan, 2003.
10. M.B. Eisen, T.P. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", Proc.Natl. Acad. Sci. USA, 95(25): 14863-14868, December 1998.
11. S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, "Systematic determination of genetic network architecture", Nature Genet, pp. 281-285, 1999.
12. T. Kohonen, "Self-organising maps." Springer, Berlin, 1995.
13. D. Jiang, C. Tang, A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey", IEEE, vol. 16, no. 11, Nov. 2004.

14. ErfanehNaghieh and YonghongPeng. *"Microarray Gene Expression Data Mining: Clustering Analysis Review",*UK,pp.1-4.
15. R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", Proc. ACM SIGMOD international conference on Management of Data, pp. 94-105, 1998.
16. S.C. Madeira and A.L. Oliveria, "Biclustering Algorithms for Biological Data Analysis: A survey", IEEE, vol. 1, no. 1, Jan- March 2004.
17. K. Sathishkumar, Dr.V.ThiagarasuM.Ramalingam, **"**Biclustering of Gene Expression Using Glowworm Swarm Optimization and Neuro-Fuzzy Discriminant Analysis",International Journal of Advanced Research in Computer Science and Software Engineering,Vol.4,pp.188-196,ISSN:2277-128X,2014.