www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 8950-8956

A Survey: Gene Selection Methods VIA Spectral Biclustering

Dr.V.Anuradha* P.Ramya **
[*]Assistant Professor, SreeSaraswathiThyagaraja College
Pollachi – 642 107, Coimbatore, Tamil Nadu, India
Email: mailanuvinu@yahoo.co.in
******Research Scholar of Computer Science, SreeSaraswathiThyagaraja College
Pollachi – 642 107, Coimbatore, Tamil Nadu, India
Email: ramyasri2291@gmail.com

Abstract: Gene selection is an important issue in microarray data processing. Microarray gene expression data usually consists of a large amount of genes. Spectral Bi-clustering is used for the selection of publically available datasets. The existing work semi unsupervised gene selection method finds much smaller and informative gene subsets without class information a priori. It uses gene ranking and gene combination selection methods of semi unsupervised gene selection where the gene combinations are selected based on the similarity between genes. The time efficiency in this method is very high. Compared to the previous work, our method can make accurate predictions with smaller gene subsets and able to identify the cancer in a single or two gene combinations. In this paper, we study a new and efficient semi-unsupervised gene selection method which results in much smaller gene subsets without prior subtype knowledge. Also it reduces the number of genes combinations.

Index Terms: Spectral Bi-clustering, Gene ranking, Gene combination, Semi-unsupervised gene selection method.

I.INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups. When representing data with fewer clusters necessarily loses certain new details, but achieves simplification.

It represents many data objects by few clusters, and hence, it models data by its clusters. Figure 1 shows a clustering process contain their objects. Traditionally clustering techniques are broadly divided in hierarchical and partitioning. [1]

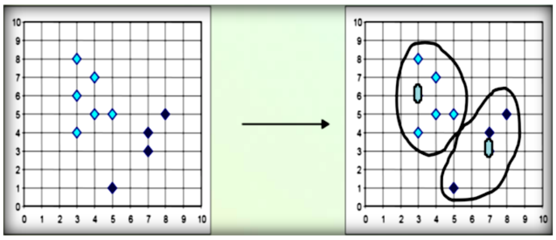


Figure 1: Clustering process

Clustering studies are based on the idea that genes that are contained in a particular pathway should be co-regulated and therefore should exhibit similar patterns of expression. Figure 2 shows a 2D representation of a clustering result. In this example, two types of genes, each one associated with a different biological function (A and B), are clustered based on their expression profiles.

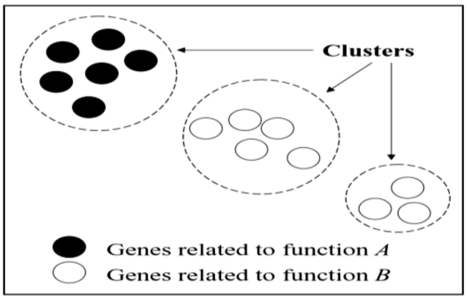

Clusters
● Genes related to function *A*
○ Genes related to function *B*

Figure 2: 2D representation of a clustering

Each cluster is encircled, and the genes that are linked to each cluster are displayed randomly within the correspondent circle.

Clustering algorithms are useful for: (a) measuring the similarity between genes according to their expression patterns under different conditions; or (b) measuring the similarity between samples described by the expression levels of a set of genes. [2]

Several journal papers have overviewed microarray data acquisition and analysis issues. Duggan et al introduced basic experimental and data analysis techniques for cDNA microarrays. Lockhart and Winzeler presented microarray methods and applications. They discussed goals for the development of more effective expression data analysis techniques. Sherlock overviewed data analysis problems and procedures based on self- organising maps (SOMs), k-means and hierarchical clustering [3].

Quackenbush reviewed general aspects in microarray data analysis, ranging from data normalisation and distance metrics, through clustering algorithms and principal component analysis, to supervised classification methods. Wu has studied important topics for the analysis of gene expression profiles, which involve: data acquisition, pre-processing, traditional clustering methods such as hierarchical cluster and SOMs, and statistical procedures for hypothesis testing. Azuaje discussed data mining and management problems, including discovery goals, methods and applications in a number of biomedical domains. Unlike the papers referred above, this review focuses on clustering approaches to analysing microarray data [4,5].

II.MICROARRAY DATA CLUSTERING MODELS

A clustering algorithm generally requires the data to be described by a matrix of values. In a microarray data experiment, an element of such a matrix may represent, for instance, the gene expression value associated with a particular biochemical perturbation. Other methods process a matrix of pair wise values, where each of them is associated with the similarity (or dissimilarity) value between two objects. A matrix element may define, for instance, the similarity between two genes expressed under a specific biological condition.

Clustering algorithms typically aim to optimise a partition quality measure. These measures may related to: (a) the heterogeneity of the clusters, also known as their intra-cluster distances; and (b) their separation from the rest of the data, also referred to as the inter-cluster distances. Therefore, clustering algorithm designers and users may need to define methods not only to assess the distance between genes (or samples), but also intra- and inter-cluster distances. Several sample-to-sample, intra- and inter-cluster metrics have been applied and/or combined in microarray data clustering models[3].

The selection of a clustering algorithm depends on the statistical nature of the data, problem and user requirements, and the attributes and constraints exhibited by the algorithms available. Clustering algorithms may be categorised into different types, according to the way they process and partition the data. Hasti and colleagues for example, define three major types of clustering techniques: combinatorial algorithms, mixture modelling and mode seeking.

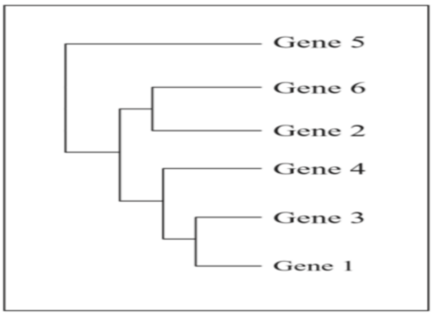


Figure 3: A dendrogram representing the hierarchical clustering of a collection of genes

III.MICROARRAY DATA CLUSTERING ALGORITHMS

Hierarchical clustering may be implemented by applying agglomerative and divisive paradigms. These techniques aim to produce a tree-like structure in which the nodes represent subsets of an expression data set. The agglomerative paradigm starts at the bottom of the data hierarchy (individual genes or samples). At each hierarchical level, it recursively merges a selected pair of clusters into a single cluster. An agglomerative technique depends on the approach used to assess the separation between clusters, such as single and average linkage methods[4]. Divisive strategies perform a top-down approach because they begin to process the entire data set as a single cluster, and recursively divide the existing clusters into two clusters at each iteration. Figure 3 illustrates a tree (also known as dendrogram) obtained after performing a hierarchical clustering of a collection of genes. Each column is associated with a gene, and the branch lengths represent the distance or dissimilarity between genes.

Azuaje and Bolshakova[3] categorise microarray data clustering algorithms into three major types: (a) hierarchical clustering, (b) models based on iterative relocation and (c) adaptive systems and other advances. Hierarchical models may be defined as above. Models based on iterative relocation consist of a number of 'learning' steps to search for an optimal partition of samples. Such processes commonly require: (1) the specification of an initial partition of objects into a number of classes; (2) the definition of a number of clustering parameters to implement the search process and assess the adequacy of its outcomes; (3) a set of procedures to transform the structure or composition of a partition; and (4) a repetitive sequence of such transformation procedures.

A.ADVANCES IN CLUSTERING-BASED ANALYSIS OF MICROARRAY DATA

Bioscientists may choose clustering solutions from a diverse and extensive collection of algorithms originating from statistical learning, pattern recognition and machine learning. Wu et al. [8] have proposed a novel approach to predicting gene function based on expression data. They showed the importance of applying multiple clustering methods to discover relevant biological patterns. The clustering methods applied were: hierarchical clustering, k-means, SOMs, 'block up/ down' and 'top-N'. [8] Each method may produce partially overlapping expression clusters. A class prediction provided by a clustering algorithm is associated with a probability value, P, which may assign a gene to multiple functional categories. Figure summarises the sequence of analysis steps of this approach.

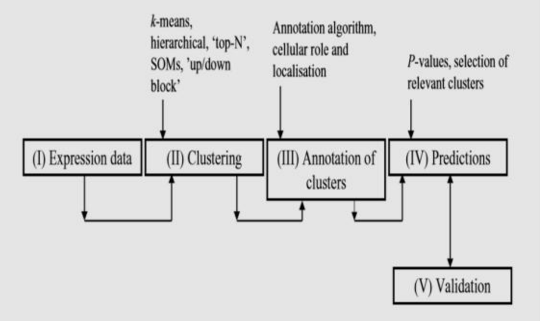

k-means, hierarchical, 'top-N', SOMs, 'up/down block'
Annotation algorithm, cellular role and localisation
P-values, selection of relevant clusters
(I) Expression data
(II) Clustering
(III) Annotation of clusters
(IV) Predictions
(V) Validation

Figure 4: A gene function prediction and validation system based on the generation of multiple expression clusters.

Gene expression data is orderly in a data matrix, where each gene corresponds to one row and each condition to one column. Every element of this matrix represents the expression level of a gene under a particular condition, and is represented by a real number. [3]

Gene expression matrices have been extensively analyzed in two dimensions: the gene dimension and the condition dimension. This corresponds to the:

- Analysis of expression patterns of genes by comparing rows in the matrix.
- Analysis of expression patterns of samples by comparing columns in the matrix.

Common objectives pursued when analyzing gene expression data include:

1) Grouping of genes according to their expression under multiple conditions.

2) Classification of a new gene, given its expression and the expression of other genes, with known classification.

3) Grouping of conditions based on the expression of a number of genes.

4) Classification of a new sample, given the expression of the genes under that experimental condition.

Clustering methods can be applied to either the rows or the columns of the data matrix, separately. Biclustering methods, on the other hand, perform clustering in the two dimensions simultaneously. This means that clustering methods derive a global model while biclustering algorithms produce a local model. When clustering algorithms are used, each gene in a given gene cluster is define using all the conditions. Similarly, each condition in a condition cluster is characterized by the activity of all the genes. 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 co-active under all conditions.

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

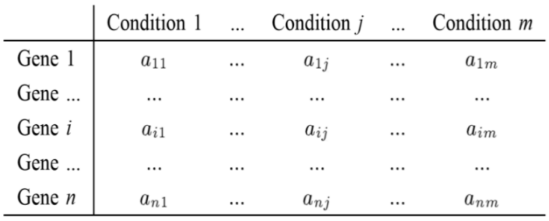
- A cluster of genes should be defined with respect to only a subset of the conditions.
- A cluster of conditions should be defined with respect to only a subset of the genes.

The clusters should not be exclusive and/or exhaustive: a gene or 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 or genes, respectively. Additionally, robustness in biclustering algorithms is especially relevant because of two additional characteristics of the systems under study. The first characteristic is the sheer complexity of gene regulation processes that require powerful analysis tools. The second characteristic is the level of noise in actual gene expression experiments that makes the use of intelligent statistical tools indispensable.

PROBLEM FORMULATION

We will be working with an n by m matrix, where element a_{ij} will be, in general, a given real value. In the case of gene expression matrices, a_{ij} represents the expression level of gene i under condition j. Table I illustrates the arrangement of a gene expression matrix.

A large fraction of applications of biclustering algorithms deal with gene expression matrices. However, there are many other applications for biclustering. For this reason, we will consider the general case of a data matrix, A with set of rows X and set of columns Y, where the elements a_{ij} corresponds to a value representing the relation between row i and column j [3].



	Condition 1	...	Condition j	...	Condition m
Gene 1	a_{11}	...	a_{1j}	...	a_{1m}
Gene ...	...	...	...	...	...
Gene i	a_{i1}	...	a_{ij}	...	a_{im}
Gene ...	...	...	...	...	...
Gene n	a_{n1}	...	a_{nj}	...	a_{nm}

Table 1: Gene Expression Data Matrix

IV.BICLUSTERING APPLICATIONS

Large datasets of clinical samples are an ideal target for biclustering [11]. As such, many applications of biclustering are performed using gene expression data obtained using microarray technologies that allow the measurement of the expression level of thousands of genes in target experimental conditions. In this application domain, we can use biclusters to associate genes with specific clinical classes or for classifying samples, among other possible interesting applications.

However, and even though most recent applications of biclustering are in biological data analysis, there exist many other possible applications in very different application domains. Examples of these application areas are: information retrieval and text mining; collaborative filtering, recommendation systems, and target marketing; database research and data mining; and even analysis of electoral data [3].

A.BIOLOGICAL APPLICATIONS

Cheng and Church [14] applied biclustering to two gene expression data matrices, specifically to the Yeast Saccharomyces Cerevisiae cell cycle expression data with 2884 genes and 17 conditions and the human B-cells expression data with 4026 genes and 96 conditions. Yang et al. [15, 16] also used these two datasets. Wang et al. [26] and Liu and Wang also used the Yeast data [27].

Lazzeroni et al. [9] also used biclustering to identify biclusters in Yeast gene expression data: the rows of the data matrix represented 2467 genes and the columns were time points within each of 10 experimental conditions. Furthermore, experiments one to three examined the mitotic cell cycle; experiments four to six tracked different strains of Yeast during sporulation; experiments seven to nine tracked expression following exposure to different types of shocks and experiment ten studied the diauxic shift.

Segal et al. [18,19] used two gene expression data matrices. They first analyzed the Yeast stress data, which characterizes the expression patterns of yeast genes under different experimental conditions by selecting 954 genes with significant changes in gene expression and the full set of 92 conditions.

Their model identifies groupings based on similarities of gene expression, the presence of known transcription factor binding sites within the gene promoters and functional annotation of genes. They identify expected gene clusters, that display similar gene expression patterns and are known to function in the same metabolic processes. They also discover new groupings of genes based on both expression levels and transcription factor binding sites. Secondly, they used the Yeast Compendium data, which observed the genomic expression programs triggered by specific gene mutations. The goal of these experiments is to assign hypothetical functions to uncharacterized genes by their deletion to known expression programs.

V.COMPARATIVE AND THEIR RESULTS

Getz et al. [20] applied biclustering to two gene expression data matrices containing cancer data. The first data matrix was constituted by 72 samples collected from acute Leukemia patients at the time of diagnosis using RNA prepared from the bone marrow mononuclear cells of 6817 human genes: 47 cases were diagnosed as ALL (Acute Lymphoblastic Leukemia) and the other 25 as AML (Acute Myeloid Leukemia). They identified a possible diagnosisto leukemia by identifyingdifferent responses to treatment, and the groups of genes to be used as the appropriate probe.

Busygin et al. [21] and Kluger et al. [22] also used these Leukemia data. The second gene expression matrix used by Getz et al. contained 40 colon tumor samples and 22 normal colon samples and 6500 human genes from which they choosed the 2000 of greatest minimal expression over the samples.

Muraly and Kasif [23] also used these two datasets. Sheng et al. [24] also used leukemia expression data. The data matrix was this time constituted by 72 samples collected from acute Leukemia patients which were now classified into three types of Leukemia: 28 cases were diagnosed as ALL (Acute LymphoblasticLeukemia), 24 as AML (Acute MyeloidLeukemia) and the remaining 20 as MLL (Mixed-Linkage Leukemia). The expression level of 12600 human genes was available.

Tang et al. [25] applied ITWC to a gene expression matrix with 4132 genes and 48 samples of Multiple Sclerosis patients and Ben-Dor et al. [10] used a breast tumor dataset with gene expression data from 3226 genes under 22 experimental conditions. Tanay et al. [11] used SAMBA to perform functional annotation in Yeast data using an expression matrix with 6200 genes and 515 experimental conditions. They also applied biclustering to human cancer data. The Lymphoma dataset they used is characterized by well defined expression patterns differentiating three types of lymphoma: Chronic Lymphocytic Leucemia (CLL), Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL).

Kluger et al. [23] also used the Lymphoma expression data used by Tanay et al. but also applied biclustering to two extra gene expression matrices: a breast tumor dataset and a central nervous system embryonal tumor dataset.

All the previous applications of biclustering analyzed biological data from gene expression matrices obtained from microarray experiments. However, biclustering can also reveal to be interesting to analyze other kind of biological data. For example, Liu and Wang used a dataset with drug activity data: a matrix with 10000 rows and 30 columns where each row corresponds to a chemical compound and each column represents a descriptor/feature of the compound. The values in the data matrix ranged from 0 to 1000.

VI.CONCLUSION

We have presented a comprehensive survey of the models, methods and applications developed in the field of biclustering algorithms. The previous work focussed on identifying the cancer by using Cosine similarity algorithm for lymphoma and liver cancer microarray datasets. The existing work semi unsupervised gene selection method finds much smaller and informative gene subsets without class information apriori. The gene expression data sets contain thousands of genes while the number of tissue samples ranges from tens to hundreds. The raw data in many cancer gene-expression data sets can be arranged in a matrix. It uses gene ranking and gene combination selection methods of semi unsupervised gene selection where the gene combinations are selected based on the similarity between genes and the best eigenvector (angle of gene) to find out the cancer. The time efficiency in this method is very high.

In "Gene Selection Method via Spectral Biclustering", the similarity algorithm for lymphoma and liver cancer microarray datasets are used, where similar genes are tested using all the combinations of dice to find out the cancer. Compared to the previous work, our method can make accurate predictions with smaller gene subsets and able to identify the cancer in a single or two gene combinations. The gene subsets selected by the previous methods are usually too large. In this paper, we study a new and efficient semi-unsupervised gene selection method which results in much smaller gene subsets without prior subtype knowledge. Also it reduces the number of genes combinations needed to predict particular type of cancer. From biological and clinic point of view, finding the small number of important genes can help medical researchers to identify the cancer genes.

REFERENCES

1) Pavel Berkhin "A Survey of Clustering Data Mining Techniques"
2) Francisco Azuaje "Clustering-based approaches to discovering and visualising microarray data patterns" HENRY STEWART PUBLICATIONS 1467-5463.. VOL 4. NO 1. 31–42. MARCH 2003
3) Sara C. Madeira and Arlindo L. Oliveira "Biclustering Algorithms for Biological Data Analysis: A Survey "..INESC-ID TEC. REP. 1/2004, JAN 2004
4) Azuaje, F. and Bolshakova, N. (2002), "Clustering genomic expression data: Design and evaluation principles", in Berrar, D., Dubitzky, W. and Granzow, M., Eds, 'A Practical Approach to Microarray Data Analysis', Kluwer Academic Publishers, London, pp. 230–245.
5) Quackenbush, J. (2001), "Computational analysis of microarray data", Nature Rev. Genet., Vol. 2, pp. 418–427.
6) Granzow, M., Berrar, D., Dubitzky, W. et al. (2001), "Tumor identification by gene expression profiles: A comparison of five different clustering methods", ACM-SIGBIO Lett., Vol. 21, pp. 16–22.
7) Hasti, T., Tibshirani, R. and Friedman, J. (2001), "The Elements of Statistical Learning", Springer, New York.
8) Wu, L., Hughes, T., Davierwala, A. et al. (2002), "Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters", Nature Genet., Vol. 31, pp. 255–265.
9) Laura Lazzeroni and Art Owen. "Plaid models for gene expression data". Technical report, Stanford University, 2000.
10) Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. "Discovering local structure in gene expression data: The order–preserving submatrix problem". In Proceedings of the 6th International Conference on Computacional Biology (RECOMB'02), pages 49–57, 2002.
11) Dan Gusfield. Algorithms on strings, trees, and sequences. Computer Science and Computational Biology Series. Cambridge University Press, 1997.
12) Ren Peeters. "The maximum edge biclique problem is NP-complete". Discrete Applied Mathematics, 131(3):651–654, 2003.
13) Amos Tanay, Roded Sharan, and Ron Shamir. "Discovering statistically significant biclusters in gene expression data". In Bioinformatics, volume 18 (Suppl. 1), pages S136–S144, 2002.
14) Yizong Cheng and George M. Church. "Biclustering of expression data". In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00), pages 93–103, 2000.
15) Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. Æ-clusters: "Capturing subspace correlation in a large data set". In Proceedings of the 18th IEEE International Conference on Data Engineering, pages 517–528, 2002.
16) Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. "Enhanced biclustering on expression data". In Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering, pages 321–327, 2003
17) Haixun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. "Clustering by pattern similarity in large data sets". In Proceedings of the 2002 ACM SIGMOD

International Conference on Management of Data, pages 394–405, 2002.

18) Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman, and Daphne Koller. “Rich probabilistic models for gene expression”. In Bioinformatics, volume 17 (Suppl. 1), pages S243–S252, 2001.

19) Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman, and Daphne Koller. “Decomposing gene expression into cellular processes”. In Proceedings of the Pacific Symposium on Biocomputing, volume 8, pages 89–100, 2003.

20) G. Getz, E. Levine, and E. Domany. “Coupled two-way clustering analysis of gene microarray data”. In Proceedings of the Natural Academy of Sciences USA, pages 12079–12084, 2000.

21) Stanislav Busygin, GerritJacobsen, and Ewald Kramer. “Double conjugated clustering applied o leukemia microarray data”. In Proceedings of the 2nd SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data, 2002.

22) Yuval Klugar, Ronen Basri, Joseph T. Chang, and Mark Gerstein. “Spectral biclustering of microarray data: coclustering genes and conditions”. In Genome Research, volume 13, pages 703–716, 2003

23) T. M. Murali and Simon Kasif. “Extracting conserved gene expression motifs from gene expression data”. In Proceedings of the Pacific Symposium on Biocomputing, volume 8, pages 77–88, 2003

24) Qizheng Sheng, Yves Moreau, and Bart De Moor. “Biclustering micrarray data by gibbs sampling”. In Bioinformatics, volume 19 (Suppl. 2), pages ii196–ii205, 2003

25) Chun Tang, Li Zhang, Idon Zhang, and Murali Ramanathan. “Interrelated two-way clustering: an unsupervised approach for gene expression data analysis”. In Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, pages 41–48, 2001.

26) Inderjit S. Dhillon.” Co-clustering documents and words using bipartite spectral graph partitioning”. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’01), pages 269–274, 2001.

27) Jinze Liu and Wei Wang. Op-cluster: ”Clustering by tendency in high dimensional space”. In Proceedings of the 3rd IEEE International Conference on Data Mining, pages 187–194, 2003.