A Survey on Biclustering

Shruthi. M. P
Department of Computer Science and Engineering
Adhiyamaan College of Engineering

Saravana Kumar. E
Department of Computer Science and Engineering
Adhiyamaan College of Engineering

Abstract

A variety of clustering approaches are used for the analysis of gene expression obtained from microarray experiments. However, the results are limited when using standard clustering methods. These results are obligatory by the existence of various experimental conditions where the activity of the genes is unrelated. The same limitation exists when traditional clustering algorithms is performed. For this reason, a number of algorithms which simultaneously clusters the rows and columns in a gene expression matrix. This simultaneous clustering, usually called as biclustering which finds the subgroups of genes and subgroups of columns, where genes exhibit correlated activities for each and every condition. This type of biclustering algorithms were used in many fields such as information retrieval and data mining. This paper analyses a large number of biclustering algorithms, used for mining gene expression. It also classifies the genes in accordance with the type of biclusters they can find and used to perform the search and also the target applications.

Keywords: Clustering, Biclustering, Microarray Data or Gene Expression Data

I. INTRODUCTION

DNA chips and different methods are used to measure the expression level of genes in any of the human beings, with various diverse samples (conditions). The samples may compare to various time focuses or distinctive experimental conditions. In different cases, the samples may have originated from various organs, from destructive or sound tissues, or even from various people. Basically envisioning this sort of information, which is broadly called gene expression information or essentially expression information, is testing and extricating organically pertinent learning is harder still. For the most part, gene expression information is organized in an matrix form, where every genes and conditions compares to rows and columns respectively. Every component of this matrix represents the expression level of a gene under a specific condition, and is numeric representation of expression which is normally the logarithm of the relative plenitude of the mRNA of the quality under the specific condition.
Biologically relevant clusters are identified by various clustering algorithms. Clustering methods can be utilized to group either genes or conditions. Significantly difficult to apply clustering algorithms to gene expression data. Many activation components are common to a group of genes at a certain condition but behaves independent under other conditions. Finding such nearby expression samples might be the way to revealing numerous genetic pathways that are not clear generally. It is accordingly very alluring to move past the clustering techniques worldwide, and to create algorithmic methodologies equipped for finding nearby samples in microarray information.

Clustering of rows and columns in a data matrix can be done separately and it derives a global model. But biclustering methods performs clustering in two dimension that is it derives local model[4]. In other words biclustering algorithms identifies the genes under a specific condition. But clustering algorithms identifies genes under all the specified conditions. Hence biclustering can be used for any of these following objectives:

1) A particular set of genes participate in a cellular process.
2) 2. When the cellular process is active at a particular subset of condition.
3) 3. A single gene may participate in multiways that may not co-active under all specified conditions.

This paper is about a survey on biclustering of gene expression data.

II. RELATED WORK

A. *Clustering Techniques*

Eisen,Spellman,Brown,&Botstein,(1998), clustering algorithms[7] were used to identify subsets of co-regulated genes similarly expressed across all the samples.

Golub et al.,(1999) and Alizadeh et al.,(2000) applied clustering algorithms to cluster samples into homogenous groups based on their gene profiling[1]. However single gene may participate in one or more conditions and it may not co-active under all the conditions.

Several studies have been done to cluster the gene, however sample exhibits similar behaviour under all conditions.
Hence a new family of algorithms called as biclustering have been proposed from the seminal work of Cheng and Church Cheng and Church,(2000).

B. Biclustering Algorithms

Cheng and Church,(2000) introduced the first algorithm to bicluster gene expression data[5]. They propose to use a mean squared residue (MSR) of a bicluster as a objective function to greedily extracts biclusters that satisfy a homogeneity constraint. It generates the row and column cluster randomly and then improves the biclusters to minimize the MSR value. Only one bicluster is identified at a time and then replaced by random values before identifying the next cluster.

Tanay,Sharon&Shamir,(2002) proposed an algorithm known as the statistical algorithmic method [20]for bicluster analysis which models the matrix as a bi-partite graph and attempts to find bi-cliques in the graph. Edges are assigned weights corresponding to being up-regulated or down-regulated and heavy subgraphs represent a bicluster. It is also capable of discovering non-complete bi-cliques in a bi-partite graph, thereby showing robustness to the presence of noise in the data

Murali&Kasif,(2002) proposed an algorithm known as Xmotifs algorithm [17].It uses a greedy method that attempts to find biclusters in discretized data. It works by computing the column cluster corresponding to each row using its statistical analysis as compared to a uniform distribution.

Ben-Dor,Chor,Karp& Yakhini,(2003) proposed an algorithm called order preserving submatrix (OPSM), that defines a bicluster as an order preserving submatrix[2]. It builds the biclusters iteratively by first generating and then growing partial biclusters. Each bicluster is based on a score given by the probability that it will grow to some fixed target size. Best partial biclusters are kept at each iteration.

Bergmann,Ihmels(2003) proposed an iterative signature algorithm[3],which is a nondeterministic model to find biclusters having two essential properties.

1) Row in a bicluster should have an average value above a given threshold.
2) The columns in a bicluster also have an average value above a certain threshold.

It starts with random initial cluster and iteratively updates the rows and columns convergence. It can find both up-regulated and down-regulated biclusters.

Prelic et al.,(2006) proposed bimax algorithm[18] that discretizes the data into 0's and 1's and search for the biclusters.It uses a divide and conquer technique to divide the matrix into a checkerboard structure.

Cho&Dhillon(2008) It uses minimum sum squared residue co-clustering as an objective function[6]. Local search strategy is incorporated that improves the final biclusters considering a single row/column at a time. Thus one batch algorithm updates the biclusters followed by the local search to adjust the clusters at a finer level.

Bozdag,parvin(2009) The correlated pattern biclustering algorithm uses the pearson correlation coefficient[4] to find biclusters having high row-wise correlation. It is done by selecting a reference row and then adding other rows having a high correlation. Several runs are made and then extracted the biclusters.

Hochreiter et al.,(2010) Factor analysis for bicluster acquisition(FABIA) uses factor analysis where a matrix is considered to be a sum of biclusters and some noise[12]. Each bicluster is considered of a sparse row and column vector. Factor analysis is used to minimize the error between real and modelled data.

Pontes et al.,(2013) introduced the Evolutionary Biclustering[19] which uses a genetic algorithm to find biclusters in an evolutionary manner. They propose a weighted combination of four objective functions to derive a unified objective function. Hence accurate biclusters are extracted

III. TOOLS

For clustering the genes, there are various types of tools for analysis.

Lattice Miner (LM) is a formal concept analysis webtool for the construction, visualization and manipulation of concept lattices. It allows the generation of formal concepts and association rules as well as the transformation of formal contexts via apposition, subposition, reduction and object/attribute generalization [4]. The manipulation of concept lattices through approximation, projection, selection and also allows drawing nested line diagrams.

Formal concept analysis based Association rule Miner (FAM) was designed on the basis of user' facility such as context editing, concept and lattice exploring, query submitting and showing the association rules in response to the query[3].

SPECLUST is a webtool for hierarchical clustering of peptide mass spectra.Mass spectra are clustered according to the peptide masses, such that mass. Hierarchical clustering of Mass Spectra (MS) with SPECLUST can in particular be useful for MS-screening of large proteomic data sets derived from 2D-gels.

Mixture Modelling (Mixmod) webtool fits mixture models to a given data set with a density estimation, a clustering or a discriminant analysis purpose. A large variety of algorithms are grouped together to estimate the mixture of parameters (EM, CEM, SEM) in order to get complete data. Mixmod is currently focused on multivariate Gaussian mixtures and fourteen different Gaussian models.Mixmod is interfaced with Scilab and Matlab.

IV. DATASETS

PACKAGE	*LIST OF DATASETS*
GeneARMA	*Time-course microarray with periodic gene expression*
Maanova	*N-dye Micro 18-array affymetrix experiment*

Adegenet	*Genetic and Genomic*
SNPMClust	*Dose-response microarray*
DCGL	*Differential co-expression and regulation analysis*
Biclust	*BicatYeast*
EMA	*Easy Microarray data Analysis*
FBN	*SNP microarray*

A large number of datasets are available readily. Real datasets are also used to test the aspects and to verify the performance of the algorithm.

V. APPLICATIONS

Biclustering can be applied whenever there is a need of data to be analysed and that has to be in the form of matrices. There exist many applications in different application domains. Examples of some application areas are: information retrieval and text mining, database research and data mining. Biclustering is applied to cluster yeast data[5][8][9], leukaemia cancer[10], movielens dataset[14], simultaneous clustering of documents and words[12],electoral data[11]. Hence biclustering can be applied in wide variety of application.

VI. CONCLUSION

Biological validation of biclusters is an open issue and has been subjected as a huge research. Hence there is a need for continuous work in construction of biologically significant groups of biclusters in large microarray data. It is believed that this review will be helpful for academicians and researches to select appropriate approach and to apply if for analysing the data.

REFERENCES

[1] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403, 503–511 .
[2] Ben-Dor, A., Chor, B., Karp, R., & Yakhini, Z. (2003). Discovering local structure in gene expression data: the order-preserving submatrix problem. Journal of Computational Biology, 10, 373–384.
[3] Bergmann, S., Ihmels, J., & Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E, 67, 31902.
[4] Bozdag, D., Parvin, J. D., & Catalyurek, U.V.(2009). A biclustering method to discover co-regulated genes using diverse gene expression datasets. Bioinformatics and computational Biology (pp.151-163). Springer.
[5] Cheng, Y., & Chruch, G.M. (2000). Biclustering of expression data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (pp.93-103).
[6] Cho,H.,& Dhilllon, I.S.(2008). coclustering of human cancer microarray using minimum sum squared residue coclustering. In IEEE/ACM Transactions on Computational Biology and Bioinformatics(TCBB):5(pp. 385-400).
[7] Eisen, m. B., Spellman, P.T., Brown, P.O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Science, 95, 14863.
[8] 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.
[9] 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.
[10] 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.
[11] Hartigan, J.A. (1972). Direct clustering of a data matrix. Journal of the American Statistical Association, 67(337), 123-129.
[12] Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., et al. (2010).FABIA:factor analysis for bicluster acquisition. Bioinformatics, 26,1520-1527.
[13] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretical co-clustering. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03), pages 89–98, 2003.
[14] 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.
[15] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. Technical report, Stanford University, 2000.
[16] Madeira, S. C., & Oliveira, A.L.(2004). Biclustering algorithms for biological data analysis: a survey. IEEE Transactions on computational Biology and Bioinformatics, 1, 24-25.
[17] Murali, T. M., & Kasif, S. (2002). Extracting conserved gene expression motifs from gene expression data. In Biocomputing 2003: Proceedings of the Pacific Symposium, Hawaii, USA(p.77). 3-7 January 2003.
[18] Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., et al.(2006). A systematic comparison and evaluation of bicluserting methods foe gene expression data. Bioinformatics, 22, 1122-1129.
[19] B. Pontes, F. Divina, R. Gir´aldez, and J. S. Aguilar-Ruiz. Virtual error: A new measure for evolutionary biclustering. In E. Marchiori, J. H. Moore, and J. C. Rajapakse, editors, Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, volume 4447 of Lecture Notes in Computer Science, pages 217–226. Springer, 2007.
[20] Tanay, A., Sharon, R., & Shamir, R.(2002). Biclustering gene expression data. In Proceedings of International Conference on Intelligent Systems for Molecular Biology.