International Journal of Computer Science and Business Informatics

IJCSBI.ORG

A Survey on Bi-Clustering and its Applications

K. Sathish Kumar
Assistant Professor in Information Technology
Gobi Arts & Science College (Autonomous)
Gobichettipalayam

M. Ramalingam
Assistant Professor in Information Technology
Gobi Arts & Science College (Autonomous)
Gobichettipalayam

Dr. V. Thiagarasu
Associative Professor of Computer science
Gobi Arts & Science College (Autonomous)
Gobichettipalayam

ABSTRACT
Biclustering is the process of immediate taking apart of the set of samples and the set of their attributes into classes. Samples and attributes are classified together which is believed to have a high importance to each other. Though, the outcome of the application of classic clustering methods to genes is limited. These limited results are forced by the survival of a number of investigational conditions where the activity of genes is not correlated. For this purpose, a number of algorithms that perform real-time clustering of the expression matrix have been proposed. In this survey, analysis about the most widely used biclustering techniques and their associated applications regarding various fields. This survey presents an study of several biclustering algorithms proposed by various authors to deals with the gene expression data efficiently. The existing algorithms are analyzed thoroughly to identify their advantages and limitations. The performance evaluation of the existing algorithms is carried out to determine the best approach. Then, in order to improve the performance of the best approach, a novel approach is been proposed in this paper.

Keywords
Biclustering, simultaneous clustering, co-clustering, Data Mining, Gene Expression Data, Gene Selection.

INTRODUCTION
Analyzing variations in expression levels of genes under different conditions (samples) is significant to recognize the basic complex biological processes that the genes take part in. In gene expression data analysis,

expression levels of genes in each sample are characterized by a real-valued data matrix with rows and columns representing the genes and the samples, correspondingly. The objective is to identify genes that have correlated expression values in a variety of samples [1].

Gene expression matrices have been widely investigated in two dimensions, that is, the gene dimension and the condition dimension. This corresponds to the [2]:

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

However, applying clustering algorithms to gene expression data runs into an important complexity. Numerous activation patterns are familiar to a group of genes only under definite experimental conditions [3]. It is then highly enviable to move further than the clustering model [4].

This paper proceeds as follows. In the next section, the background study is described. Section 3 describes related works in this field, etc.

1. SURVEY

Biclustering, which has been applied intensively in molecular biology explore in recent times, gives a structure for identifying hidden substructures in large high dimensional matrices. Tanay et al. [5] said that a bicluster as a subset of genes that together react upon a subset of conditions. Biclustering algorithms might have two different objectives; one is to find one bicluster or to identify a given number of biclusters.

Cheng and Church's Algorithm (CC) [6] describe an underneath a user-defined thresholdδ. In order to identify the largest δ -bicluster in the data, they recommend a twophase approach: first, rows and columns are removed from the original expression matrix until the above limitation is satisfied. Afterward, previously deleted rows and columns are added to the resulting submatrix as long as the bicluster score does not exceed. This process is iterated numerous times where biclusters are covered with random values previously.

In [7] an improved form of CC algorithm were proposed in which avoids the problem of random interference caused by covered biclusters. Samba [8] introduced a graph-theoretic methodology to biclustering in grouping with a statistical data model. In this structure, the expression matrix is modeled as a bipartite graph, a bicluster is defined as a subgraph, and a likelihood score is used in order to assess the importance of observed subgraphs. A related heuristic algorithm called Samba aims at identifying highly important and

distinctive biclusters. In a recent investigation, this approach has been extended to integrate multiple types of experimental data.

In [9], Order Preserving Submatrix Algorithm (OPSM) is a bicluster defined as a sub matrix that conserves the order of chosen columns for all of the selected rows. Also, it can be said that, the expression values of the genes inside a bicluster persuade a matching linear ordering across the selected samples. Based on a stochastical model, the authors implemented a deterministic algorithm to discover large and related important biclusters. This idea has been taken up in a recent investigation by [10].

Tang et al. [11] proposed the Interrelated Two-Way Clustering (ITWC) algorithms that come together the results of the data matrix to generate biclusters. After normalizing the rows of the data matrix, they calculate the vector-angle cosine value between each row and a pre-defined steady pattern to test the row values vary much among the columns and remove the ones with little variation. After that they utilize a correlation coefficient as similarity measure to measure the strength of the linear relationship between two rows or two columns, to carry out two-way clustering. As this similarity measure based on the pattern and not on the absolute magnitude of the spatial vector, it also allows the identification of biclusters by means of coherent values represented by the multiplicative model.

The worst-case running-time complexity of BiMax for matrices comprising disjoint biclusters is O (nmb) and meant for arbitrary matrices is of order O (nmb min {n, m}) Noureen and Qadir [12]. In [13] It main goal is to find market segments between tourists who are similar to each other, therefore allowing a targeted marketing mix to be flourished. In general data used to segment tourists are illustrated. Small samples and many questions give rise to serious methodological problems that have usually been addressed by means of factor-cluster analysis to reduce the dimensionality of data.

In [14] The technique is depends on a force-directed graph where biclusters are represented as feasible overlapped groups of genes and conditions. In [15] introduced an expression pattern based biclustering approach, CoBi, for combining both positively and negatively keeping up genes from microarray expression data. Regulation pattern and resemblance in degree of fluctuation are accounted for as computing likeness among two genes. Unlike conventional biclustering approaches, which utilize greedy iterative approaches, it uses a BiClust tree that requires single pass over the entire dataset to find a set of biologically related biclusters. Biclusters determined from various gene expression datasets by the technique show highly improved functional categories. MSBE Biclustering algorithm [16] and the threshold of the average relationship score is a user input factor to allow the user to control the excellence of the biclustering results.

In [17], a fuzzy biclustering technique is introduced, it is based on formulating the one-way clustering along the row and column dimension as a normalized graph cut problem. The graph cut problem is after that solved by a spectral decomposition, followed by K-mean clustering of the eigenvectors. The biclustering of the row and column dimensions is accomplished by a three-stage procedure. Initially, the original data matrix experiences one-way clustering in the row dimension to gain k clusters. After that, a novel pattern matrix where each row is specified by the average number of rows that belong to the same cluster in the original data matrix is calculated. Again, the new data matrix then experience the same one-way clustering in the column dimension to obtain k' clusters. Lastly, a table of fuzzy relation coefficients that share each of the k row clusters to each of the k' column clusters are worked out. By calculating the new data matrix by means of the result of the initial stage clustering, the fuzzy biclustering algorithm attains a biclustering of the original data matrix.

2. PROBLEMS AND DIRECTIONS

Clustering the microarray data is based on user defined threshold value, this influence the quality of biclusters produced. This value will affect the quality of biclusters by missing some of the genes or by including a number of unnecessary genes. Once a bicluster is produced, their entries are replaced by random numbers, avoiding the identification of overlapping biclusters. Problem of finding the minimum set of biclusters to cover all the elements in a data matrix is extremely tough.

4. MOTIVATION

Biclustering is a significant approach in microarray data analysis. The primary basis for using bi-clustering in the analysis of gene expression data are (1) similar genes might show similar behaviors, not all conditions, (2) genes may contribute in more than one purpose, resulting in one regulation model. By biclustering algorithms, one may perhaps attain sets of genes that are co-regulated under subsets of state of affairs. Some of the biclustering algorithm is based on the user defined threshold value, this influences the quality of biclusters produced. To prevail over these problems, build up an algorithm to discover an optimal bicluster with threshold value rather than user defined threshold.

3. CONCLUSIONS

A complete survey of the models, methods, and applications developed in the field of biclustering algorithms are investigated and analyzed. The list of applications presented is by no means comprehensive, and all-inclusive list of potential applications would be prohibitively extended. The list of accessible algorithms is also very composite, and many combinations of thoughts can be personalized to obtain new algorithms potentially more

effectual in exacting applications. The modification and validation of biclustering methods by comparison with known biological data is surely one of the most important open issues. Another motivating region is the application of robust biclustering techniques to new and existing application domains.

REFERENCES

[1] Doruk Bozdag, Ashwin S. Kumar and Umit V. Catalyurek, Comparative Analysis of Biclustering Algorithms, 2010.

[2] Sara C. Madeira and Arlindo L. Oliveira, Biclustering Algorithms for Biological Data Analysis: A Survey, INESC-ID TEC. REP. 1/2004, JAN 2004.

[3] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulus, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM/SIGMOD International Conference on Management of Data, pp. 94–105, 1998.

[4] 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), pp. 49–57, 2002.

[5] A. Tanay, R. Sharan and R. Shamir,Discovering statistically significant biclusters in gene expression data. Bioinformatics, Vol. 18, pp. 136-144, 2002.

[6] Y. Cheng and G. M. Church. Biclustering of expression data. In Proc. of the International Conference on Intelligent Systems for Molecular Biology, pp. 93–103, 2000.

[7] Yang, J., Wang, H., Wang, W., Yu, P.S., (2003) Enhanced Biclustering on Expression Data. BIBE 2003, pp. 321-327.

[8] Tanay, A., Sharan, R., Kupiec, M., Shamir, R., (2004) Revealing Modularity and Organization in the Yeast Molecular Network by Integrated Analysis of Highly Heterogeneous Genomewide Data, PNAS, pp. 101-9, 2981-2986.

[9] Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z., (2002) Discovering Local Structure in Gene Expression Data: The Order-Preserving Sub-Matrix Problem, Proceedings of the 6th Annual International Conference on Computational Biology, pp. 49-57.

[10] Liu, J., Wang, W., (2003) OP-Clusters: Clustering by tendency in high dimensional space, Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), pp. 187-194.

[11] 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, pp. 41–48, 2001.

[12] Noureen, N., Qadir, M.A., BiSim: A Simple and Efficient Biclustering Algorithm, Soft Computing and Pattern Recognition, SOCPAR '09. International Conference of 2009, pp. 1 – 6.

[13] Sara Dolnicar, Sebastian Kaiser, Katie Lazarevski, Friedrich Leisch, Biclustering Overcoming Data Dimensionality Problems in Market Segmentation, Journal of Travel Research.Vol. 51 No. 1 41-49, 2012.

[14] Rodrigo Santamarı´a*, Roberto Thero´ n and Luis Quintales, BicOverlapper: A tool for bicluster visualization, Vol. 24 no. 9 2008, pp. 1212–1213.

[15] Swarup Roya, , Dhruba K Bhattacharyyab, Jugal K Kalitac, CoBi: Pattern Based Co-Regulated Biclustering of Gene Expression Data, Preprint submitted to Elsevier March 9, 2013.

[16] Liu X, Wang L: Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics 2007, Vol. 23, No. 1, pp. 50-56.

[17] Koutsonikola VA, Vakali A. A Fuzzy Bi-clustering Approach to Correlate Web Users and Pages. Int. J. Knowledge and Web Intelligence 2009, Vol. 1 No.1-2, pp. 3-23.

This paper may be cited as:

Sathish Kumar, K., Ramalingam, M. and Thiagarasu, V. 2014. A Survey on Bi-Clustering and its Applications. *International Journal of Computer Science and Business Informatics, Vol. 12, No. 1, pp. 65-70.*