A survey on biological data analysis by biclustering

Majid Rastegar-Mojarad  
Faculty of Electrical Engineering,  
Persian Gulf University,  
Bushehr, Iran  
rastegar_m@comp.iust.ac.ir

Saeed Talatian-Azad  
Faculty of Computer Engineering,  
Islamic Azad University,  
Bushehr, Iran  
s.talatian@mail.sbu.ac.ir

Behrouz Minaei-Bidgoli  
Assistant Professor of School of  
Computer Engineering,  
Iran Univ. of Scie. & Technology,  
Tehran, Iran  
minaebi@cse.msu.edu

Abstract—Several non-supervised machine learning methods have been used in the analysis of gene expression data obtained from microarray experiments. Recently, biclustering, a non-supervised approach that performs simultaneous clustering on the row and column dimensions of the data matrix, has been shown to be remarkably effective in a variety of applications. The discovery of biclusters, which denote groups of items that show coherent values across a subset of all the transactions in a data set, is an important type of analysis performed on real-valued data sets in various domains, such as biology. In this survey, we analyze several of existing approaches to biclustering that use in biological data analysis.

Keywords; data mining, biclustering, biological data analysis

I. INTRODUCTION

DNA chips and other techniques measure the expression level of a large number of genes, perhaps all genes of an organism, within a number of different experimental samples (conditions). The samples may correspond to different time points or different environmental conditions. In other cases, the samples may have come from different organs, from cancerous or healthy tissues, or even from different individuals. Simply visualizing this kind of data, which is widely called gene expression data or simply expression data, is challenging and extracting biologically relevant knowledge is harder still [1].

Several non-supervised machine learning methods have been used in the analysis of gene expression data obtained from microarray experiments. Recently, biclustering, a non-supervised approach that performs simultaneous clustering on the row and column dimensions of the data matrix has been shown to be remarkably effective in a variety of applications. The goal of biclustering is to find subgroups of genes and subgroups of conditions, where the genes exhibit highly correlated behaviors. In the most common settings, biclustering is an NP-complete problem, and heuristic approaches are used to obtain sub-optimal solutions using reasonable computational resources.

Biclustering can be applied whenever the data to analyze has the form of a real-valued matrix $A$, where the set of values $A_{ij}$ represent the relation between its rows $i$ and its columns $j$. An example of this kind of data are the gene expression matrices. Moreover, it can be applied when the data can be modeled as a weighted bipartite graph. Furthermore, biclustering can be used when the goal is to identify sub-matrices described by a subset of rows and a subset of columns with certain coherence properties. Large datasets of clinical samples are an ideal target for biclustering. 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.

DNA microarray data are usually arranged in a matrix, where each row corresponds to a gene and each column an experimental condition. Each entry in the matrix records the expression level of a gene as a real number, which is usually derived by taking the logarithmic of the relative abundance of the mRNA of that genes in a specific condition [2]. An important objective of analyzing this kind of data is the classification of genes and conditions and the identification of regulatory process. With the aim of analyzing such groups and samples, clustering has an important role in the exploratory analysis of microarray data [3].

Hartigan’s pioneering work on direct clustering was the first to reveal the potential of co-clustering, also called biclustering, [4]. In a two dimensional matrix, co-clustering aims at identifying homogeneous local patterns, each of which consists of a subset of rows and a subset of columns. In particular, co-clustering has attracted genomic researchers, because the co-clustering model is compatible with our understanding of cellular processes, where a subset of genes are coregulated under certain experimental conditions, but behave almost independently under other conditions [5]. The paper by Madiera and Oliveira provides an extensive survey on the application of co-clustering to biological data analysis [6]. Another interesting survey on biclustering algorithms is also in [7],Cheng and Church [8] are considered to be the first to apply co-clustering to gene expression data. They proposed a greedy search heuristic that generates biclusters, one at a time, based on a
homogeneity constraint, called mean squared residue. Since then, several similar residue based co-clustering approaches have been proposed to enhance the work of Cheng and Church.

In this paper, we survey recently biclustering algorithms to analyze biological data.

II. EARLIER WORKS

One of the earliest co-clustering formulations, block clustering was introduced by Hartigan who called it “direct clustering” [4], and it refers to the “simultaneous clustering” of both rows and columns of a data matrix [9]. Hartigan introduced various co-clustering quality measures and models, including the partitional model and gave a greedy algorithm for a hierarchical co-clustering model, where the partitions of row and columns can be described in a hierarchical manner by trees.

Cheng and Church [8] are considered to be the first to apply co-clustering to gene expression data. They proposed a greedy search heuristic to generate biclusters that satisfy a certain homogeneity constraint, called mean squared residue. The algorithm produces one co-cluster at a time using a low mean squared residue (except the trivial case where all genes have constant expression values) as the criterion for identifying a co-cluster.

A sequence of node (i.e., row or column) deletions and additions is applied to the original matrix, while the mean squared residue of the co-cluster is kept under a given threshold. After each co-cluster is produced, the elements of the co-cluster are replaced with random numbers, and then, the same procedure is applied to the modified matrix to generate another, possibly overlapping, co-cluster until the required number of co-clusters is found. Since Cheng and Church [8], several similar approaches have been taken within gene expression analysis [10], [11], all of which basically utilize the definition of the mean squared residue and try to enhance the original work.

Recently, Cho et al. [12] developed two minimum sum squared co-clustering algorithms: one with its objective function based on the partitioning model proposed by Hartigan [4] and the other one based on the mean squared residue formulated by Cheng and Church [8].

III. HEURISTIC ALGORITHMS

A good survey on biclustering is available in literature [6], with a categorization of the different heuristic approaches made as follows:

- Greedy iterative search [8, 10]: make a locally optimal choice, in the hope that this will lead to a globally good solution.
- Exhaustive biclustering enumeration: the best biclusters are identified, using an exhaustive enumeration of all possible biclusters existent in the data, in exponential time [14].
- Distribution parameter identification [1]: identify best fitting parameters by minimizing a criterion through an iterative approach.

A coupled two-way iterative method [13] has been devised to iteratively generate a set of biclusters, at a time, in cancer datasets. In the process it repeatedly performs one way hierarchical clustering on the rows and columns of the data matrix, while using stable clusters of rows as attributes for column clustering and vice versa. The Euclidean distance is used as the similarity measure, after normalization of the data.

Yang et al. [10] point out that random number replacement can interfere with the future discovery of co-clusters, especially ones that have overlap with the discovered ones. They present an algorithm called Flexible Overlapped biclustering (FLOC) that simultaneously produces k co-clusters whose mean residues are all less than a predefined constant. FLOC incrementally moves a row or column out of or into a co-cluster depending on whether the row or column is already included in that co-cluster or not, and this move is called an action. The idea of action is very similar to our incremental LS strategy in co-clustering. Bipartite graphs are employed in Ref. [14], with a bicluster being defined as a subset of genes that jointly respond across a subset of conditions. The objective is to identify the maximum-weighted subgraph. Here a gene is considered to be responding under a condition if its expression level changes significantly, under that condition over the connecting edge, with respect to its normal level. This involves an exhaustive enumeration, with a restriction on the number of genes that can appear in the bicluster. A simultaneous discovery of all biclusters is made at the same time. It may be noted that in all these methods it is possible to generate overlapped gene clusters.

The Plaid model [1] tries to capture the approximate uniformity in a submatrix of the gene expression data, while discovering one bicluster at a time in an iterative process. The input matrix is described as a linear function of variables corresponding to its biclusters, and an iterative maximization process is pursued for estimating the function. It searches for patterns where the genes differ in their expression levels by a constant factor.

IV. SPECTRAL BICLUSTERING

Many algorithms have been proposed in literature for biclustering gene expression data. Spectral biclustering of microarray data is proposed in [15]. In [15] have developed a method that simultaneously clusters genes and conditions, finding distinctive “checkerboard” patterns in matrices of
gene expression data, if they exist. In a cancer context, these checkerboards correspond to genes that are markedly up- or down-regulated in patients with particular types of tumors. This method, spectral biclustering, is based on the observation that checkerboard structures in matrices of expression data can be found in eigenvectors corresponding to characteristic expression patterns across genes or conditions. In addition, these eigenvectors can be readily identified by commonly used linear algebra approaches, in particular the singular value decomposition (SVD), coupled with closely integrated normalization steps. The largest several left and right singular vectors of the normalized gene expression matrix are computed and then a final clustering step using k-means, and normalized cuts is applied to the data projected onto the topmost singular vectors. They incorporate different normalizations of genes and conditions in the hope of discarding irrelevant constant background noise.

V. DETERMINISTIC ALGORITHMS

Some deterministic algorithms such as BiMax [16] and OPSM [5] have also been proposed. The BiMax approach, proposed by Prelic et al. is based on a simple, binary data model that assumes only a response/no response level for a gene with respect to an experiment. Accordingly, the expression data has to be discretized before application of the BiMax algorithm, which could cause loss of useful information. The algorithm exactly finds all the inclusion maximal biclusters consisting of genes that jointly respond across a subset of experiments. The number of co-clusters identified by BiMax is exponential in the number of genes and experiments, making it infeasible to store and evaluate all the co-clusters in case of large datasets. The order preserving sub matrix algorithm (OPSM) looks for submatrices in which the expression levels of all the genes induce the same linear ordering of the experiments. This algorithm although very accurate, is designed to identify only a single co-cluster. A recent extension to OPSM finds multiple, overlapping co-clusters in noisy datasets, but is very expensive in the number of features [17].

VI. OTHER ALGORITHMS

Biclustering of gene expression data by tendency is described in [18]. This work focuses on discovering a subset of genes which exhibit similar expression patterns along a subset of conditions in the gene expression matrix. Specifically, they are looking for the Order Preserving clusters (OPCluster), in each of which a subset of genes induce a similar linear ordering along a subset of conditions. The pioneering work of the OPSM model [5], which enforces the strict order shared by the genes in a cluster, is included in our model as a special case. This model is more robust than OPSM because similarly expressed conditions are allowed to form order equivalent groups and no restriction is placed on the order within a group.

Genetic algorithm based methods with local search strategy for identifying overlapped biclusters in gene expression data is presented in [19]. The objective is to find sub-matrices, i.e., maximal subgroups of genes and subgroups of conditions where the genes exhibit highly correlated activities over a range of conditions. Since these two objectives are mutually conflicting, they become suitable candidates for multi-objective modeling. In [19] a novel multi-objective evolutionary biclustering framework is introduced by incorporating local search strategies. A new quantitative measure to evaluate the goodness of the biclusters is developed.

A method on discovering biclusters in gene expression data based on high-dimensional linear geometries is described in [20]. They extend previous work [21] and present a novel perspective for biclustering problem through a geometric interpretation. Such a new perspective to regard biclusters with different coherent patterns as hyperplanes in a high dimensional space, and facilitates the use of any generic plane finding algorithm for detecting them. The geometric viewpoint of their approach provides a unified framework to handle different types of linear patterns simultaneously by matching a specific set of linear geometries. It also reveals the existence of the general linear model, which can unify the additive and the multiplicative models. As a particular realization of their framework, they implemented a Hough Transform-based hyperplane detection algorithm. The experimental results on human lymphoma gene expression dataset show that their algorithm is highly effective for gene expression data biclustering.

SVD based methods has also been used in order to obtain biclusters in gene expression data and also in many potential applications [22, 23]. Applying SVD directly on the data may obtain biclusters, but obtaining efficient biclusters on data is still a challenging problem. Hence in [3] propose modular SVD based method for biclustering in gene expression data. The standard SVD based method may not be very effective under different conditions of gene, since it considers the global information of gene and conditions and represents them with a set of weights. While applying SVD on sub data, local features of genes and conditions can be extracted efficiently in order to obtain better biclusters. In [3] made an attempt by overcoming the aforementioned problem by partitioning a gene expression data into several smaller sub-data and then SVD is applied to each of the sub-data separately. The three main steps involved in that method, named M-SVD Biclustering, are:

1. An original whole pattern denoted by a matrix is partitioned into a set of equally sized sub-data in a non-overlapping way.
2. SVD is performed on each of such sub-data.
3. At last, a single global feature is synthesized by concatenating each sub-data’s.

In order to partition the data, they experimented in two ways. In the first type, they choose rows which are similar by computing mean of the row data. In the second type, they partitioned data in non-overlapping way. After thorough study, they decided to work with second type of partitioning
the data, which leads in better result compared to first type [3].

VII. OTHER APPLICATIONS

We may be interested in simultaneously clustering genes and experimental conditions in bioinformatics applications, simultaneously clustering documents and words in text mining, tokens and contexts in natural language processing, users and movies in recommender systems, as well as other applications.

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 [25] 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.

VIII. OTHER SURVEYS

Madeira and Oliveira [6] are the first to classify many existing numerical biclustering algorithms systematically based on the additive and multiplicative bicluster models. It should be pointed out that some symbolic, coherent evolution or numerical biclusters, such as those produced by cMonkey [24], SAMBA [14] and some statistical criteria, cannot be classified as additive or multiplicative patterns directly. They have presented a comprehensive survey of the models, methods and applications developed in the field of biclustering algorithms. The list of applications presented is by no means exhaustive, and an all-inclusive list of potential applications would be prohibitively long. Another interesting survey on biclustering algorithms is also in [7]. They survey some of the biclustering models and algorithms that were developed for gene expression analysis. Their coverage is not exhaustive, and is biased toward the more practical methods. They explain at least one method from each class of algorithms under development.

IX. CONCLUSION

In this paper we survey on several algorithms in biological data analysis. These algorithm use of biclustering or co-clustering, to cluster row and column of a matrix simultaneously. In analysis of gene expression data mostly use of this approach.

The list of available algorithms is also very extensive, and many combinations of ideas can be adapted to obtain new algorithms that are potentially more effective in particular applications.

Many issues in biclustering algorithm design also remain open and should be addressed by the scientific community. A further interesting direction is to find algorithms for the biclustering problem and use of these algorithms in new applications.

REFERENCES


