Expert Systems With Applications 55 (2016) 520–531

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Biclustering of human cancer microarray data using co-similarity based co-clustering

Syed Fawad Hussain[a,b,*], Muhammad Ramazan[c]

[a] Faculty of Computer Science and Engineering, GIK Institute, Khyber Pakhtunkhwa, Pakistan
[b] Laboratoire TIMC-IMAG, CNRS/UJF 5525, Universite de Grenoble, Saint-Martin-d'Hères, France
[c] INSERM and Université Joseph Fourier, Institut Albert Bonniot, Grenoble, France

ARTICLE INFO

Article history:
Received 11 February 2016
Accepted 16 February 2016
Available online 26 February 2016

Keywords:
Biclustering
Co-clustering
Microarray analysis
Hierarchical clustering

ABSTRACT

Biclustering of gene expression data aims at finding localized patterns in a subspace. A bicluster (sometimes called a co-cluster), in the context of gene expression data, is a set of genes that exhibit similar expression intensity under a subset of experimental features (conditions). Most biclustering algorithms proposed in the literature aim at finding sub-matrices that exhibit some sort of coherence by selecting an initial sub-matrix and iteratively adding or subtracting rows and columns. These algorithms are generally dependent on the initial, hard selection of the gene and condition clusters respectively. In this work, we adapt a recently proposed approach for clustering textual data to find biclusters in gene expression data. Our proposed technique is based on the concept of co-similarity between genes (and between conditions) that exploits weighted higher order paths in a bipartite graph representation of the gene expression data. Therefore, we build statistical relations between genes and between conditions by comparing all genes and conditions before finally extracting biclusters from the data. We show that the proposed technique is able to find meaningful non-overlapping biclusters both on synthetically generated data as well as real cancer data. Our results indicate that the proposed technique is resistant to noise in the data and can successfully retrieve biclusters even in the presence of relatively large amount of noise. We also analyze our results with respect to the discovered genes and observe that our extracted biclusters are supported by biological evidences, such as enrichment of gene functions and biological processes.

1. Introduction

Gene expression data is generated by DNA chips and Microarray techniques involving DNA sequences from two different samples—one of which is a normal tissue sample (used as a reference) and the other is a test sample that usually corresponds to a certain condition that we are interested in, such as a tumorous tissue (Eren, Deveci, Küçüktunç, & Çatalyürek, 2013; Madeira & Oliveira, 2004; Voggenreiter, Bleuler, & Gruissem, 2012). The widespread use of microarray technologies in the past 10 years have provided scientists and researchers with the ability to measure the level of expression of different genes under a given set of different conditions. The information gathered during this process is highly significant to biologists and a host of data analysis methods have been used to analyze the data.

One of the most important visualization techniques for gene expression analysis is clustering gene expression data. In DNA microarray experiments, finding clusters of genes that exhibit similar transcriptional characteristics is instrumental in functional annotation, tissue classification and motif identification. The data collected during such experiments is usually stored in the form of a matrix as described in Table 1 below. Each element in the data matrix is a numeric representation of the expression of a particular gene under a particular experimental condition. Clustering is then performed to identify groups or subsets of genes that may be related to each other. Such groups tend to show similar behavior in their expression levels, such as being up-regulated or down-regulated.

Several clustering techniques on gene expression datasets have been used in the literature to identify biologically relevant groupings of genes and samples (Eren et al., 2013; Voggenreiter et al., 2012). In (Eisen, Spellman, Brown, & Botstein, 1998), for instance, the authors have used clustering as a means to identify subsets of co-regulated genes similarly expressed across all samples. Such clustered genes can be thought of as a group that participates in similar cellular processes. Similarly, (Golub et al., 1999) and

* Corresponding author at: Faculty of Computer Science and Engineering, GIK Institute, Pakistan. Tel.: +923215692930.
E-mail addresses: fawadsyed@gmail.com, fawadhussain@giki.edu.pk (S.F. Hussain), ramzan797@yahoo.com (M. Ramazan).

http://dx.doi.org/10.1016/j.eswa.2016.02.029
0957-4174/

Table 1
Example of raw gene expression values from the colon cancer dataset (Alon et al., 1999).

	Exp 1	Exp 2	Exp 3	Exp 4
Gene 1	8589.4163	5468.2409	4263.4075	4064.9357
Gene 2	9164.2537	6719.5295	4883.4487	3718.1589
Gene 3	3825.705	6970.3614	5369.9688	4705.65
Gene 4	6246.4487	7823.5341	5955.835	3975.5643
Gene 5	3230.3287	3694.45	3400.74	3463.5857

(Alizadeh et al., 2000) applied clustering to cluster samples into homogeneous groups according to their gene expression profiles where each cluster corresponds to a particular phenotype. Several other studies have been performed for extracting similar information, see for instance (Eren et al., 2013; Saber, 2015). All these studies, however, assume that a sample exhibits similar behavior under all conditions and tries to find correlations in a high dimensional space.

In many experiments, however, a given cellular process is expressive only under a sub-set of conditions rather than the entire set. Similarly, a single gene may participate in one or more pathways that may or may not be co-active under all conditions (Cho & Dhillon, 2008; Madeira & Oliveira, 2004). In such cases, microarrays can be used to identify disease-related genes by comparing their expressive levels under normal and affected conditions. Hence, a new family of algorithms, known as biclustering algorithms (or co-clustering algorithms) have been proposed in the literature, particularly in the last decade starting from the seminal work of Cheng and Church (Cheng & Church, 2000). Biclustering algorithms are a well-studied set in the literature of gene expression data analysis because of their strength in finding local patterns. These algorithms have been shown to outperform traditional clustering algorithms in bioinformatics (for instance see (Eren et al., 2013; Prelic et al., 2006; Saber, 2015)). The first algorithm to directly find block patterns in a matrix was proposed by Hartigan (Hartigan, 1972). Cheng and Church (Cheng & Church, 2000) were the first to apply biclustering to gene expression data by using a block correlation metric, termed as Mean Residual Score (MRS), that measures the quality of a discovered bicluster. Since then, several other authors have proposed different algorithms generally based on the MRS measure, such as (Aguilar-Ruiz, 2005; Divina & Aguilar-Ruiz, 2006; Madeira & Oliveira, 2004; Prelic et al., 2006).

A notable drawback of the above mentioned approaches, however, is that its discovery of biclusters is greatly biased toward biclusters showing low variance correlations among constituent samples/conditions. Moreover, as the variance changes by the square of the change in scale, the MRS measure favors correlations over lower scales. Thus, such algorithms are biased toward discovering 'flat' biclusters containing genes with relatively smooth expression levels with low scales (Chakraborty & Maka, 2005). Bisson and Hussain (Bisson & Hussain, 2008) proposed a new measure, called χ-Sim, that measures the co-similarity between rows and between columns as a means of weighting subspaces using higher order paths in a bi-partite graph corresponding to the data values (such as in Table 1). The algorithm has been used to produce biclusters both for text (Bisson & Hussain, 2008), and more recently (Hussain, 2011) for gene expression datasets. Therefore, in this paper, we intend to extend and enhance their work by proposing new strategies both for improving the biclustering task as well as automatically extracting significant biclusters.

We propose different strategies that substantially enhance the χ-Sim algorithm and also empirically demonstrate the effect of these proposed strategies. As a first contribution, we propose a variation of the χ-Sim algorithm that is specifically suited for high dimensional data and study its behavior on different datasets. We also investigate different transformation techniques such as row and column standardization, discretization and binormalization. The χ-Sim algorithm is well suited to deal with sparse, integral data coming from a text corpus that is homogenous. Gene expression data, by contrast, may contain genes at different scales which could affect the computation of the co-similarity measure. Therefore, we adapt the co-similarity measure for biclustering of gene expression data. Secondly, we investigate the effect of noise on the χ-Sim algorithm. Noise can be referred to as impurities resulting either from the experimentation or biological process (de França & Coelho, 2015; Klebanov & Yakovlev, 2007). Real data is rarely perfect and it is important to study the behavior of an algorithm in the presence of noise. To this end, we use several synthetic datasets with known noise implanted into the biclusters and investigate the behavior of the algorithm with increasing levels of noise. Thirdly, we study different clustering techniques in combination with χ-Sim such as single linkage, complete linkage, average linkage and Ward's linkage. Finally, we extend the biclustering technique to select the most discriminating biclusters from amongst a list of possible bicluster combinations.

In order to evaluate the performance of our proposed algorithm on gene expression dataset, we perform several experimentations. Firstly, we compare the biclustering performance of χ-Sim with other biclustering algorithms on publicly available synthetic datasets, specifically designed for this kind of test. As used previously in the literature (Prelic et al., 2006), we use two external measures called the *average cocluster relevance* and the *average cocluster recovery* to assess the performance of these algorithms. We measure the sample clustering using the accuracy measure and also compare our proposed algorithm on several real world human cancer microarray datasets in the literature including the Colon cancer, Leukemia, and Lung cancer datasets. In all the cases, χ-Sim with our proposed strategies result in better accuracy than by using ordinary clustering or other tested biclustering techniques. As a further step, we also investigate χ-Sim in combination with different hierarchical linkage strategies, in particular with Single Linkage, Complete linkage and Ward's linkage. Interestingly and contrary to other observations where single linkage has been shown to perform better with hierarchical clustering on gene expression data, we observe that χ-Sim results in better average accuracy when used in combination with Hierarchical clustering using Ward's linkage on the real datasets. We assess the robustness of χ-Sim to data noise by comparing the average cocluster relevance and average cocluster accuracy at varying levels of noise from 0 to 10%. Since gene clusters do not usually have ground truth to compare against, we investigate the gene clusters by illustrating coherence of the resulting biclusters in a checkerboard structure. Furthermore, we evaluate the enrichment of functional annotations of gene clusters in the context of gene ontology (GO) using the publicly available functional profiling tool, DAVID (Huang, Sherman, & Lempicki, 2009a; Huang, Sherman, & Lempicki, 2009b).

The rest of this paper is organized as follows: In Section 2, we give a brief overview of the literature on biclustering algorithms. Section 3 presents the biclustering technique proposed in this paper and discusses various aspects of the algorithm. In Section 4, we describe the synthetic and real world datasets used for our evaluation as well as the evaluation criteria, while Section 5 presents the experimental results and analysis. Finally, we conclude our work in Section 6.

2. Literature review

A vast amount of literature exists on biclustering of microarray data. Cheng and Church (Cheng & Church, 2000) introduced the first algorithm to bicluster gene expression data. Biclustering, however, has also been used for other applications such as text