Computers in Biology and Medicine 42 (2012) 245–256

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

An effective measure for assessing the quality of biclusters

Federico Divina [a], Beatriz Pontes [b,*], Raúl Giráldez [a], Jesús S. Aguilar-Ruiz [a]

[a] School of Engineering, Pablo de Olavide University, Ctra. Utrera s/n, 41013 Seville, Spain
[b] Department of Computer Science, University of Seville, Avda. Reina Mercedes s/n, 41012 Seville, Spain

ARTICLE INFO

Article history:
Received 17 March 2011
Accepted 26 November 2011

Keywords:
Biclustering
Gene expression data
Shifting and scaling patterns

ABSTRACT

Biclustering is becoming a popular technique for the study of gene expression data. This is mainly due to the capability of biclustering to address the data using various dimensions simultaneously, as opposed to clustering, which can use only one dimension at the time. Different heuristics have been proposed in order to discover interesting biclusters in data. Such heuristics have one common characteristic: they are guided by a measure that determines the quality of biclusters. It follows that defining such a measure is probably the most important aspect. One of the popular quality measure is the *mean squared residue* (MSR). However, it has been proven that MSR fails at identifying some kind of patterns. This motivates us to introduce a novel measure, called *virtual error* (VE), that overcomes this limitation. Results obtained by using VE confirm that it can identify interesting patterns that could not be found by MSR.

1. Introduction

Microarray techniques allow to simultaneously measure the expression level of thousands of genes under different experimental conditions, producing in this way a huge amount of data. Microarray data is widely used in genomic research, and is usually organized in matrices. In such matrices, rows and columns may represent, for instance, experimental conditions and genes, respectively. Thus, an element of such expression matrix stands for the expression level of a given gene under a specific experimental condition.

Different techniques have been used in order to extract information from expression matrices. Among these, clustering has been widely used, with the main goal of finding groups of genes that present a similar variation of expression level under all the experimental conditions [1]. However, relevant genes are not necessarily related to every condition. In other words, genes might be relevant only for a subset of experimental conditions [2]. Thus, clustering should be performed not only on one dimension (genes) but on two dimensions (genes and conditions) simultaneously.

For this reason, biclustering techniques [3] are becoming popular due to the ability of simultaneously grouping both genes and conditions. The first approach applied to microarray analysis was proposed by Cheng and Church [4]. Biclustering bases its essential principle on clustering, from which differs in two main aspects. Considering a microarray of gene expression data, a typical clustering technique would build a set of clusters, where each gene belongs exactly to one single cluster. Nevertheless, many genes may be grouped into diverse clusters (or none of them) depending on their participation in different biological processes within the cell [5]. Another difference is found in the fact that biclustering aims at identifying genes that are co-expressed under a subsets of conditions. This is essential for numerous biological problems, such as the analysis of genes contributing to certain diseases [2], assigning biological functionalities to genes or when the conditions of a microarray are diverse.

Finding significant biclusters in a microarray has been proven to be a NP-hard problem [6] and much more complex than clustering [7]. Consequently, many of the proposed techniques are based on optimization procedures as the search heuristic. The development of a suitable heuristic is a critical factor for discovering interesting biclusters in an expression matrix. In order to guide a search heuristic, it is essential to define a measure or cost function for establishing the quality of bicluster. The use of a suitable measure is a key factor, as it determines the effectiveness of the heuristic. Moreover, such a measure can be used for comparing the performances of different search strategies.

As already stated, Cheng and Church [4] were the first in applying biclustering to microarray data. Their proposal was based on a greedy search heuristic based on a cost function, called *mean squared residue* (henceforth MSR). MSR measures the numerical coherence among the genes in a bicluster. Cheng and Church maximize the volume with an upper bound on MSR, since MSR tends to decrease as volume of bicluster increases. MSR has also been used as part of the cost function in some other works. Gremalschi and

* Corresponding author.
E-mail addresses: fdivina@upo.es (F. Divina), bepontes@us.es (B. Pontes), giraldez@upo.es (R. Giráldez), aguilar@upo.es (J.S. Aguilar-Ruiz).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.11.015

Altun [8] proposed the opposite strategy to Cheng and Church that is to minimize MSR with a lower bound on volume of bicluster. In [9], the authors developed an iterative algorithm for finding a predefined number of biclusters. Bryan et al. [10] applied in their work a simulated annealing heuristic. A greedy strategy was proposed in [11], where the search starts from seed generated with the *k*-means clustering algorithm. Similar strategies were proposed in [12] and in [13], where a particle swap optimization (PSO) technique was used to refine the initial biclusters. A multi-objective PSO approach was proposed by Liu et al. [14], and in another work [15] they proposed a multiple objective ant colony optimization algorithm. An approach based on evolutionary computation was proposed by Divina and Aguilar [7] and Bleuler et al. [16], while other authors [17] based their proposal on fuzzy technology and spectral clustering.

Other biclustering approaches, which are not based on MSR, include the proposal by Tanay et al. [6], based on the use of bipartite graphs and probabilistic techniques. Sheng et al. [18] introduced the use of Gibbs sampling for finding biclusters. Carmona-Saez et al. [19] presented a new data mining technique, based on matrix factorization. Madeira and Oliveira [20] found all relevant biclusters in linear time on the size of the microarray. Ayadi et al. [21] proposed an enumeration algorithm which uses a tree structure to represent different biclusters discovered during the enumeration process. The same authors also proposed a hill climbing strategy [22]. Bicego et al. [23] rely on a probabilistic model, called topic model, to detect groups of highly correlated genes and conditions. Liu and Wang [24] developed a polynomial time algorithm, which searched for optimal biclusters with the maximum similarity score. Finally, Hanczar and Nadif [25,26] try to improve the performances of biclustering algorithms by using the ensemble approach.

Even if MSR has been used in many proposals for finding biclusters, it nevertheless presents some drawbacks that will be discussed in the next section. In this paper, we propose a measure, called *virtual error* (henceforth VE), as a novel cost function for evaluating biclusters based on the concept of behavioural pattern. Gene correlation in a bicluster can be represented by two distinct kind of patterns: shifting and scaling, being both of them formally described by Aguilar [27]. Taking into account the concept of pattern, it is possible to focus the analysis of expression data on the general behaviour that genes exhibit under subsets of conditions, instead of grouping genes with similar expression values. Shifting patterns represent groups of genes following exactly the same trends, i.e., parallel behaviour, but in different range of values. Scaling patterns represent genes in a bicluster fluctuating in unison, without presenting the same differences through the conditions, although conserving a multiplicative factor.

In order to test the effectiveness of VE, we incorporated it in a multi-objective evolutionary biclustering algorithm. Experiments show that VE yields the algorithm at finding interesting biclusters, confirming the validity of our proposal.

This paper is organized as follows. In the next section, we present the main motivations for this work. In Section 3 an analysis of the shifting and scaling patterns is given; we then provide a formal definition of VE in Section 4, followed by a description of the algorithm used in the experiments in Section 5. In Section 6 experimental results obtained from different datasets are presented and discussed. Finally, in Section 7, we summarize the main conclusions.

2. Motivation

Let, from now on, $\mathcal{B}$ be a bicluster containing I conditions and J genes, and let b_{ij} denote the elements of $\mathcal{B}$, where $1 \le i \le I$ and $1 \le j \le J$. Then $\mathcal{B}$ can be represented as follows:

$$\mathcal{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1J} \\ b_{21} & b_{22} & \cdots & b_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ b_{I1} & b_{I2} & \cdots & b_{IJ} \end{pmatrix}$$

where rows are relative to conditions and columns to genes.

The MSR of a bicluster $\mathcal{B}$ is then given by Eq. (1)

$$MSR(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (b_{ij} - \mu_{c_i} - \mu_{g_j} + \mu_{\mathcal{B}})^2 \quad (1)$$

where μ_{c_i} and μ_{g_j} are the means of ith row (condition c_i) and jth column (gene g_j), respectively; and $\mu_{\mathcal{B}}$ is the mean of the whole bicluster.

The lower the MSR, the better the numerical coherence among the genes is and, therefore, the better the quality of a bicluster seems to be. Thus, when the genes of a bicluster $\mathcal{B}$ show exactly the same shape, with the only difference that they started with different initial values, then the MSR of $\mathcal{B}$ is equal to 0 [27].

Nevertheless, biclusters with constant genes, i.e., that present a flat behaviour across all the experimental conditions, will also have MSR equal to 0. The same holds for a bicluster containing only one gene or condition. Thus, MSR equals to 0 does not always identify a good bicluster. Other measures, such as the volume and the gene variance of biclusters may be used in combination to the MSR, in order to reject trivial biclusters.

The volume is the number of rows multiplied by the number of columns ($I \cdot J$), and the gene variance of a bicluster $\mathcal{B}$ is given in Eq. (2)

$$var_{\mathcal{B}} = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} (b_{ij} - \mu_{g_j})^2 \quad (2)$$

If a bicluster presents a high gene variance, it means that its genes exhibit fluctuating trends under the same subset of conditions.

From the above considerations, it becomes evident that a criterion needs to be used in order to establish when the MSR of a bicluster can be considered low. In [4] a user parameter, denoted as δ, is used as threshold: biclusters with MSR higher than δ are rejected. Before applying biclustering, δ needs to be independently set for each dataset [28].

MSR has been proven to be inefficient for finding some kind of biclusters in microarray data, especially when they present strong scaling tendencies. In [27] an in-depth analysis on the consequences of using MSR as a quality measure for searching biclusters is proposed. One of the main conclusions is that MSR is not capable of assessing the quality of biclusters containing shifting trends, as shifting behaviour does not affect the MSR. Moreover, scaling patterns have an undesirable effect for evaluating biclusters: small scaling variations in data lead to great increases of MSR. Therefore, good biclusters may have a score greater than δ.

Fig. 1 shows an example of a bicluster discovered in the Human B-cells dataset [29]. This is a typical visualization of bicluster, where conditions are represented in the *X*-axis, the values of gene expression are represented in the *Y*-axis and each line is a gene. Cheng and Church [4] set δ to 1200 for this dataset. From a visual inspection of the bicluster, it can be seen that it is a quality bicluster, since the genes are highly co-expressed, presenting strong scaling trends. Nevertheless, the MSR for this example is 3470.15, almost three times the value of δ.

These observations motivate us to propose a novel approach for evaluating biclusters, taking into account the scaling behaviour inherent to gene data. This behaviour is more difficult to detect than the shifting one, but it is more probable in nature. Being able to find biclusters containing also scaling patterns