Methods 83 (2015) 18–27

Contents lists available at ScienceDirect

Methods

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

Measuring the quality of linear patterns in biclusters

Shuhua Chen, Juan Liu *, Tao Zeng

School of Computer, Wuhan University, Wuhan, Hubei 430072, China

ARTICLE INFO

Article history:
Received 31 January 2015
Accepted 2 April 2015
Available online 15 April 2015

Keywords:
Biclustering
Gene expression
Linear pattern
Coherence measurement

ABSTRACT

In microarray analysis, biclustering is used to find the maximal subsets of rows and columns satisfying some coherence criteria. The found submatrices are usually called as biclusters. On one hand, different criteria would help to find different types of biclusters, thus the definition of coherence criterion is critical to the biclustering method. On the other hand, qualitative criteria result to qualitative biclustering methods that cannot evaluate the qualities of the biclusters, while quantitative criteria can numerically show how well the mined biclusters and are more useful in real applications. In bioinformatics communities, there are several quantitative coherence measurements for linear patterns proposed. However, they face the problem of weakness in finding all subtypes of linear patterns or sensitivity to the noise. In this work, we introduce a coherence measurement for the general linear patterns, the minimal mean squared error (MMSE), which is designed to handle the evaluation of biclusters with shifting, scaling and the general linear (the mixed form of shifting and scaling) correlations. The experiments on synthetic and real data sets show that the proposed methods is appropriate for identifying significant general linear biclusters.

1. Introduction

Nowadays, microarray techniques play an important role in biological research. The data are usually transformed into a numerical matrix to be analyzed, in which rows refer to genes and columns represents experimental conditions. genes are not necessarily coherently expressed on all conditions, instead they might be co-expressed only on a small subset of conditions such as cells [1] or samples from the same disease subtype [2]. In such case, biclusters are biologically and clinically interesting. Therefore, biclustering is crucial on finding gene subsets (rows) showing similar expression behaviors on subsets of conditions (columns) and many methods have been proposed in literature for gene expression data biclustering.

To do biclustering, one first needs to determine a criterion to judge whether a submatrix can be regarded as a bicluster, then he should design an efficient searching algorithm to find out as many as possible maximal submatrices that satisfy the criterion. There are two kinds of assessment criteria: qualitative and quantitative, accordingly the biclutering methods can be classified as qualitative and quantitative ones. Since the qualitative biclustering methods cannot numerically describe how well the mined biclusters, the quantitative ones are more informative to real applications. Moreover, according to the found pattern types, biclustering algorithms can also be categorized as nonlinear and linear coherent ones. Various techniques have been proposed to identify nonlinear coherent bicluster patterns from gene expression data: (1) two-way hierarchical clustering method (HCL) [3]; (2) the Bayesian-based biclustering (BBC) [4] that are based on a rigorous statical model; (3) the qualitative biclustering algorithm (QUBIC) that can search biclusters in a general form [5]; (4) the iterative signature algorithm (ISA) [6]; (5) the order preserving submatrix algorithm (OPSM) [7]; (6) the xMotif method [8]; (7) Samba method [9]; and (8) the gene shaving (GSH) method [10]. Though originally targeted by non-linear methods [9,7,6,8,11,12], linear measures are still informative in evaluating the quality of discovered biclusters [13].

Biclustering is first introduced by Cheng and Church [14]. They use the Mean Squared Residue (MSR) to measure the shifting patterns in biclusters. The MSR measure has been widely followed by other methods as a coherence measure to identify linear biclusters from gene expression data [14–20]. The innovative idea of MSR is based on the variability of the genes' expression neighborhood with regard to their arithmetic mean—if all the elements in a submatrix are similar, then the mean squared residue is small. However, the MSR-based biclustering methods [14,21] suffer from a major limitation, pointed out by Aguilar-Ruiz [16], that it can only help to identify biclusters exhibiting shifting patterns whereas linear biclusters include not only shifting, but also scaling and even more general linear patterns. Since the scaling or linear patterns are also very important to the biological functions, and

* Corresponding author.
E-mail address: liujuan@whu.edu.cn (J. Liu).

http://dx.doi.org/10.1016/j.ymeth.2015.04.005
1046-2023/

have been convinced by many researches related to gene regulatory pathways [15,22], as well as different kinds of cancers and tumors [23,24] (especially the negative correlated linear patterns), developing biclustering methods for the linear patterns is of significance.

In order to face the challenge of searching scaling and linear patterns, a few methods have been proposed. For example, the p-cluster and δ-cluster methods [25] search scaling patterns indirectly by using logarithm transformation to convert them into normal shifting patterns, they impose a strict assumption on the model thus cannot deal with general linear patterns. Anirban et al. proposed Scaling Mean Squared Residue (SMSR) to specially measure scaling patterns [26], however, this measurement cannot correctly evaluate the negative scaling patterns. Teng et al. used the Average Correlation Value (ACV) to evaluate the quality of biclusters [27], which is not a rigid measurement and sensitive to noise though it can correctly identify most approximate perfect scaling patterns. Pontes et al. suggested to use virtual errors (VE) by computing the difference between the standardized gene expression values and the expected pattern values, to identify some interesting biclusters which are ignored by MSR [28]. However, this measurement still has no ability to perfectly evaluate the negative correlated patterns. Ayadi et al. proposed Average Spearman's rho (ASR) based on Spearman's rank correlation [29] to overcome the weakness of ACV, but it fails to measure the negative correlated patterns. Flores et al. proposed Spearman's biclustering measure(SBM) using absolute value to detect the negative correlated patterns [20]. However, it fails to distinguish a perfect linear pattern from a non-linear pattern with same ranking trends. Nepomuceno et al. have tried to minimize the mean squared residue by nonlinear optimization techniques [30]. Unfortunately, they failed to find the negative correlations due to the constraints on the parameters. Gu et al. proposed a statistical model to describe different kinds of gene expression patterns, but it puts constraints on the overlap among biclusters [4].

From above we can see that there are two common but critical problems to be solved in the above methods: (i) they lack a unified rigid ranking measurement to evaluate the scaling/linear patterns; (ii) they cannot detect a special kind of patterns where negative correlations are involved. Although some newly developed qualitative methods have the ability of detecting the scaling patterns [31,5], even with the negative correlations [5], how to quantitatively measure scaling/linear patterns is still an open question. With these regards, we try to define a generalized, unified measurement not only for shifting patterns, but also for positively and negatively scaling patterns, or even general linear patterns.

In this work, we define *minimal mean squared error* (MMSE) as a coherence measurement to identify biclusters with shifting, scaling, and general linear patterns. By the validation experiments, we show the superiority of our proposed MMSE as compared to other measurements on the measurement of general linear patterns. The experimental results also show MMSE-based biclustering has competitive performance than other traditional clustering and non MSR-based biclustering methods because its found MMSE-biclusters include most numerical and biological significant linear patterns undetected by other methods.

2. Problem definition

Through out the paper, we denote a gene expression profiling (or microarray) data set as a triplet $M = (G, C, l)$, where $G = \{g_1, g_2, \ldots, g_n\}$ is a set of genes (rows), $C = \{c_1, c_2, \ldots, c_m\}$ is a set of conditions (columns), and $l : G \times C \Rightarrow R$ is the level function by which $l(g_i, c_j)$ represents the expression level of gene g_i on condition c_j. Or simply, a gene microarray data set is represented as a matrix below if i stands for gene g_i, column j for condition c_j, where $l_{i,j}$ is short for $l(g_i, c_j)$,

$$\mathbf{M} = \begin{pmatrix} l_{1,1} & l_{1,2} & \ldots & l_{1,m} \\ l_{2,1} & l_{2,2} & \ldots & l_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ l_{n,1} & l_{n,2} & \ldots & l_{n,m} \end{pmatrix}$$

Definition 1 (*bicluster*). Given a gene expression data set $M = (G, C, l)$, let $B = (I, J, d)$ be a triplet, where $I \subseteq G$, $J \subseteq C$, and $d : I \times J \Rightarrow R$ is the level function of B satisfying $d(g_i, c_j) = l(g_i, c_j)$, $\forall (g_i, c_j) \in I \times J$. B is a bicluster *iff* genes in I exhibit a coherent expression behavior (correlated with each other following a specific pattern) across all the conditions in J.

Now that we only consider how to measure the coherence of B instead of the whole gene expression data M, for simplicity, we abbreviate $d(g_i, c_j)$ as $d_{i,j}$, $g_i \in I$ as $i \in I$, and $c_j \in J$ as $j \in J$ without confusion hereafter (gene g_i in ith row, and condition c_j in jth column). Thus, the triplet $B = (I, J, d)$ is equivalent to a submatrix of M:

$$\mathbf{B} = \begin{pmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,|J|} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,|J|} \\ \vdots & \vdots & \vdots & \vdots \\ d_{|I|,1} & d_{|I|,2} & \ldots & d_{|I|,|J|} \end{pmatrix}$$

Definition 2 (*shifting pattern*). A bicluster $B = (I, J, d)$ is said to exhibit a shifting pattern if all of its elements $d_{i,j}$ satisfy the condition:

$$d_{i,j} = \pi_j + \beta_i \tag{1}$$

where π_j is the base value of the jth column, and β_i is the shifting factor for the ith row.

Definition 3 (*scaling pattern*). A bicluster $B = (I, J, d)$ is said to exhibit a scaling pattern if all of its elements $d_{i,j}$ satisfy the condition:

$$d_{i,j} = \alpha_i \pi_j \tag{2}$$

where π_j is the base value of the j_{th} column, and α_i is the scaling factor for the i_{th} row.

To integrate the ideas behind both shifting and scaling patterns, we define linear patterns:

Definition 4 (*linear pattern*). A bicluster $B = (I, J, d)$ is said to exhibit a linear pattern if all of its elements $d_{i,j}$ satisfy the condition:

$$d_{i,j} = \alpha_i \pi_j + \beta_i \tag{3}$$

where π_j is the base value of the jth column, α_i and β_i are the scaling and shifting factors for the ith row respectively. We also call $\pi = \{\pi_1, \pi_2, \ldots, \pi_{|J|}\}$ the base vector, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_{|I|}\}$ the scaling vector, and $\beta = \{\beta_1, \beta_2, \ldots, \beta_{|I|}\}$ the shifting vector of the linear pattern.

Shifting, scaling and linear patterns are all of biological interests. A shifting pattern of a subset of genes reveals that the genes express towards to the same trend, though the curves shift to each