Journal of Biomedical Informatics 57 (2015) 163–180

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

Biclustering on expression data: A review

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

[a] Department of Languages and Computer Systems, University of Seville, Seville, Spain
[b] School of Engineering, Pablo de Olavide University, Seville, Spain

ARTICLE INFO

Article history:
Received 22 January 2015
Revised 22 June 2015
Accepted 30 June 2015
Available online 6 July 2015

Keywords:
Microarray analysis
Gene expression data
Biclustering techniques

ABSTRACT

Biclustering has become a popular technique for the study of gene expression data, especially for discovering functionally related gene sets under different subsets of experimental conditions. Most of biclustering approaches use a measure or cost function that determines the quality of biclusters. In such cases, the development of both a suitable heuristics and a good measure for guiding the search are essential for discovering interesting biclusters in an expression matrix. Nevertheless, not all existing biclustering approaches base their search on evaluation measures for biclusters. There exists a diverse set of biclustering tools that follow different strategies and algorithmic concepts which guide the search towards meaningful results. In this paper we present a extensive survey of biclustering approaches, classifying them into two categories according to whether or not use evaluation metrics within the search method: biclustering algorithms based on evaluation measures and non metric-based biclustering algorithms. In both cases, they have been classified according to the type of meta-heuristics which they are based on.

1. Introduction

Technological advances in genomic offer the possibility of completely sequentialize the genome of some living species. The use of microarray techniques allows to measures the expression levels of thousands of genes under several experimental conditions. Usually, the resulting data is organized in a numerical matrix, called Expression Matrix [1]. Each element of the this data matrix denotes the numerical expression level of a gene under a certain experimental condition. With the development of microarray techniques, the interest in extracting useful knowledge from gene expression data has experimented an enormous increase, since the analysis of this information can allow discovering or justifying certain biological phenomena [2].

Various machine learning techniques have been applied successfully to this context [3]. Clustering techniques aim at finding groups of genes that present a similar variation of expression level under all the experimental conditions. If two different genes show similar expression tendencies across the samples, this suggests a common pattern of regulation, possibly reflecting some kind of interaction or relationship between their functions [1].

Yet despite their usefulness, the use of clustering algorithms has an important drawback, since they consider the whole set of samples. Nevertheless, genes are not necessarily related to every sample, but they might be relevant only for a subset of samples. This aspect is fundamental for numerous problems in the Biomedicine field [4]. Thus, clustering should be simultaneously performed on both dimensions, genes and conditions. Another restriction of the clustering techniques is that each gene must be clustered into exactly one group. However, many genes may belong to several clusters depending on their influence in different biological processes [5]. These drawbacks are solved by biclustering techniques, which have also been widely applied to gene expression data [6–10]. Biclustering was introduced in the 1970s by Hartigan [11], although Cheng and Church [12] were the first to apply it to gene expression data analysis. Other names such as co-clustering, bi-dimensional clustering, two-way clustering or subspace clustering often refer to the same problem formulation.

Tanay et al. [13] proved that biclustering is an NP-hard problem, and therefore much more complex than clustering [14]. Therefore, most of the proposed methods are based on optimization procedures as the search heuristics. The development an effective heuristic as well as the use of a suitable cost function for guiding the search are critical factors for finding significant biclusters in a micriarray. Nevertheless, not all existing biclustering approaches base their search on evaluation measures for biclusters. There exists a diverse set of biclustering tools that follow different strategies and algorithmic concepts which guide the search towards meaningful results.

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

http://dx.doi.org/10.1016/j.jbi.2015.06.028
1532-0464/

This paper provides a review of a large number of biclustering approaches existing in the literature and a classification which separates them into two main categories: biclustering algorithm based on evaluation measures, turn grouped according to their properties type of meta-heuristics in which they are based on; and non metric-based biclustering algorithms, turn grouped attending to their most distinctive property. In both cases we have focused on classical biclustering strategies, thus excluding in this study different specializations existing in the literature, such as biclustering based on a previous matrix binarization or biclustering for temporal series.

Next section presents an unified notation for bicluster representation, and a description of the different kind of expression patterns which biclustering algorithms aim at finding in their solutions. Third and fourth sections survey most important existing biclustering algorithms, based or not on the use of evaluation measures within the search, respectively. In both sections they have been classified according to the type of meta-heuristics in which they have been based on. Finally, a discussion on the methods under study is provided in the last section, together with the main conclusions derived from this work.

2. Definitions

Biclusters are represented in the literature in different ways, where genes can be found either in rows or columns, and different names refer the same expression sub-matrix.

Let, from now on, $\mathcal{B}$ be a bicluster consisting of a set I of $|I|$ genes and a set J of $|J|$ conditions, in which b_{ij} refers to the expression level of gene i under sample j. Then $\mathcal{B}$ can be represented as follows:

$$\mathcal{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1|J|} \\ b_{21} & b_{22} & \dots & b_{2|J|} \\ \vdots & \vdots & \ddots & \vdots \\ b_{|I|1} & b_{|I|2} & \dots & b_{|I||J|} \end{pmatrix}$$

where the gene g_i is the ith row, i.e., $g_i = \{b_{i1}, b_{i2}, \dots, b_{i|J|}\}$, and condition c_j is the jth column, i.e., $c_j = \{b_{1j}, b_{2j}, \dots, b_{|I|j}\}$.

Genes and samples means in biclusters are frequently used in several evaluation measure definitions. We represent these values as b_{iJ} and b_{Ij} referring to the i row (gene) and j column (sample) means, respectively. Furthermore, the mean of all the expression values in $\mathcal{B}$ is referred to as b_{IJ}. Note that these definitions above may alter the original authors' notations in the contributions reviewed in this paper.

2.1. Bicluster taxonomy based on gene expression patterns

Several types of biclusters have been described and categorized in the literature, depending on the pattern exhibited by the genes across the experimental conditions [15]. For some of them it is possible to represent the values in the bicluster using a formal equation. We define the following elements: π represents any constant value for $\mathcal{B}$; $\beta_i (1 \leqslant i \leqslant |I|)$ and $\beta_j (1 \leqslant j \leqslant |J|)$ refer to constant values used in additive models for each gene i or condition j; and $\alpha_i, (1 \leqslant i \leqslant |I|)$ and $\alpha_j, (1 \leqslant j \leqslant |J|)$ correspond to constant values used in multiplicative models for each experimental gene i or condition j. Thus, biclusters can be categorized in the follows types:

- **Constant values.** A bicluster with constant values reveals subsets of genes with similar expression values within a subset of conditions. This situation may be expressed by: $b_{ij} = \pi$.
- **Constant values on rows or columns.** A bicluster with constant values in the rows/columns identifies a subset of genes/conditions with similar expression levels across a subset of conditions/genes. Expression levels might therefore vary from gene to gene or from condition to condition. It can also be expressed either in an additive or multiplicative way:
  - Additive: $b_{ij} = \pi + \beta_i, b_{ij} = \pi + \beta_j$
  - Multiplicative: $b_{ij} = \pi \times \alpha_i, b_{ij} = \pi \times \alpha_j$
- **Coherent values on both rows and columns.** This kind of biclusters identifies more complex relations between genes and conditions, either in an additive or multiplicative way:
  - Additive: $b_{ij} = \pi + \beta_i + \beta_j$
  - Multiplicative: $b_{ij} = \pi \times \alpha_i \times \alpha_j$
- **Coherent evolutions.** Evidence that a subset of genes is up-regulated or down-regulated across a subset of conditions without taking into account their actual expression values. In this situation, data in the bicluster does not follow any mathematical model.

According to the former definitions, it is possible to formally describe two kind of patterns summarizing all the previous situations: shifting and scaling patterns [16]. They have been defined using numerical relations among the values in a bicluster.

A bicluster $\mathcal{B}$ follows a *perfect shifting pattern* if its values can be obtained by adding a constant-condition number β_j to a typical value for each gene (π_i). β_j is said to be the *shifting coefficient* for condition j. Graphically, a perfect shifting pattern gives a parallel behavior of the genes. In this case, the expression values in the bicluster fulfil the following equation: $b_{ij} = \pi_i + \beta_j$.

Similarly, a bicluster follows a *perfect scaling pattern* changing the additive value in the former equation by a multiplicative one. This new term α_j is called the *scaling coefficient*, and represents a constant value for each condition. In this case, the genes do not follow a parallel tendency. Although the genes present the same behavior with regard to the regulation, changes are more abrupt for some genes than for others. The following equation defines whether a bicluster follows a perfect scaling pattern or not: $b_{ij} = \pi_i \times \alpha_j$.

A bicluster may include some of the aforementioned patterns or even both of them, shifting and scaling, at the same time. This kind of pattern corresponds to the most general situation that can be described using a mathematical formula, when a bicluster exhibits coherent values on both rows an columns, for the additive and multiplicative model at the same time. When it is the case, it is said that the bicluster follows a *perfect shifting and scaling pattern*, and its values can be represented by this equation: $b_{ij} = \pi_i \times \alpha_j + \beta_j$. Nevertheless, to visually identify if some bicluster follows a combined pattern is more difficult that to find a single shifting or scaling pattern, since the effects of one have influence on the other.

2.2. Bicluster taxonomy based on structure

It is also interesting classify the biclustering mathods regarding to the way in which rows and columns from the input matrix are incorporated in biclusters. We named this as *Bucluster Structure*. In this sense, we can define the follows structures:

- **Row exhaustive.** Every gene must belong to at least one bicluster, that is, there are no genes not assigned to at least a bicluster.
- **Column exhaustive.** Every condition must belong to at least one bicluster, that is, there are no conditions not assigned to at least a bicluster.