Pattern Recognition 42 (2009) 2578–2588

Contents lists available at ScienceDirect

Pattern Recognition

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

# A probabilistic relaxation labeling framework for reducing the noise effect in geometric biclustering of gene expression data

Hongya Zhao[a,*], Kwok Leung Chan[a], Lee-Ming Cheng[a], Hong Yan[a,b]

[a]Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[b]School of Electrical and Information Engineering, University of Sydney, NSW 2006, Sydney, Australia

ARTICLE INFO

*Article history:*
Received 10 July 2008
Received in revised form 9 February 2009
Accepted 7 March 2009

*Keywords:*
Gene expression data analysis
Clustering
Biclustering
Hough transform
Probabilistic relaxation labeling

ABSTRACT

Biclustering is an important method in DNA microarray analysis which can be applied when only a subset of genes is co-expressed in a subset of conditions. Unlike standard clustering analyses, biclustering methodology can perform simultaneous classification on two dimensions of genes and conditions in a microarray data matrix. However, the performance of biclustering algorithms is affected by the inherent noise in data, types of biclusters and computational complexity. In this paper, we present a geometric biclustering method based on the Hough transform and the relaxation labeling technique. Unlike many existing biclustering algorithms, we first consider the biclustering patterns through geometric interpretation. Such a perspective makes it possible to unify the formulation of different types of biclusters as hyperplanes in spatial space and facilitates the use of a generic plane finding algorithm for bicluster detection. In our algorithm, the Hough transform is employed for hyperplane detection in sub-spaces to reduce the computational complexity. Then sub-biclusters are combined into larger ones under the probabilistic relaxation labeling framework. Our simulation studies demonstrate the robustness of the algorithm against noise and outliers. In addition, our method is able to extract biologically meaningful biclusters from real microarray gene expression data.



## 1. Introduction

Microarray technology is a high-throughput and parallel platform that can provide expression profiling of thousands of genes under different biological conditions, thereby enabling rapid and quantitative analysis of gene expression patterns on a global scale [1,2]. Microarray data can be represented as a matrix. In this paper, we assume that each row of the data matrix corresponds to a gene and each column an experimental condition. In practical applications, the roles of rows and columns may be exchanged and data analysis algorithms need to be changed correspondingly. Each entry in the matrix records the expression level of a gene as a real number, which is usually derived by taking the logarithm of the relative abundance of the mRNA of a gene in a specific condition. By analyzing the matrix, we can learn more about the cellular operation in organisms. However, this analysis is complex due to the large number of genes to be considered, the limited number of conditions, and noise in the data [3,4].

Clustering is widely used in microarray data analysis. While traditional clustering methods, such as the hierarchical and k-means clustering algorithms, are useful in investigating the underlying patterns of gene expression datasets [5–8], they have several limitations. First, most clustering methods only measure the global similarity between expression profiles. Second, all genes and conditions have to be assigned to clusters. In these methods, it is assumed that related genes in a cluster behave similarly across all measured conditions. However, an interesting cellular process for most cases may be involved in a subset of genes co-expressed only under a subset of conditions. Discovering such local expression patterns may be the key to uncovering many genetic pathways that are not apparent otherwise. Therefore, it is highly desirable to move beyond the clustering paradigm, and to develop approaches capable of discovering local patterns in microarray data [4,9–12].

Inspired by Hartigan's so called "direct clustering" [12], biclustering was first introduced to gene expression analysis by Cheng and Church [9]. In general, biclustering refers to the simultaneous clustering on the row and column dimensions of the data matrix [4]. A biclustering scheme that produces gene and condition clusters simultaneously can model the situation in which related genes are considered to be co-regulated under certain conditions, but to behave almost independently under other conditions. Furthermore, a biclustering model can avoid those irrelevant genes that are not active in any experimental conditions under consideration.

* Corresponding author.
*E-mail address:* zhycyz@yahoo.com.cn (H. Zhao).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.03.016

The research literature on biclustering has been booming in recent years. Existing biclustering methods include Cheng and Church's (CC) algorithm [9,13], sequential evolutionary biclustering (SEBI) [14,15], flexible overlapped biclustering algorithm (FLOC) [16], gene shaving [17], order-preserving sub-matrix (OPSM) [18], spectral biclustering [19], coupled two-way clustering (CTWC) [20], statistical-algorithmic method for bicluster analysis algorithm (SAMBA) [21], iterative signature algorithm (ISA) [22,23], xMotif [24], and the fast divide-and-conquer algorithm (Bimax) [25]. Comprehensive surveys of biclustering algorithms can be found in Refs. [4,11]. A systematic comparison of some biclustering methods is made in [25].

In general, existing algorithms perform biclustering by adding or deleting rows and/or columns in the data matrix in optimal ways such that a merit function is improved by the action. A different viewpoint of biclustering can be formulated in terms of the spatial geometric distribution of points in data space. The biclustering problem is tackled as the identification and division of coherent sub-matrices of data matrices into geometric structures (lines or planes) in a multidimensional data space [26]. Such perspective makes it possible to unify the formulations of different types of biclusters and extract them using an algorithm for detecting geometric patterns, such as lines and planes.

Recently, pattern recognition based methods have been developed for data biclustering [26–29]. In these algorithms, the well-known Hough transform (HT) is employed to detect geometric patterns. However, the direct HT-based biclustering algorithm becomes ineffective in terms of both computing time and storage space. To overcome the difficulties, sub-dimension based method has been introduced in the biclustering algorithm. The strategy reduces the computational complexity considerably. For example, the geometrical biclustering algorithm (GBC) only performs the HT in 2D column-pair spaces [28].

After obtaining sub-biclusters in sub-dimensional spaces, we need to merge small sub-biclusters into larger ones. In the GBC, the combination criterion of common genes and conditions can be too strict to form meaningful biclusters of larger sizes. In fact, we find in our studies that the number of genes in the identified biclusters is often small and the outcome is sensitive to the noise. Furthermore, the geometric properties of biclusters are ignored in the combination steps.

To overcome the shortcomings of the GBC, we propose an improved biclustering algorithm within the framework of probabilistic relaxation labeling. Relaxation labeling processes are widely used in many different domains including image processing, pattern recognition, and article intelligence [30,31]. They are iterative procedures that aim to reduce the ambiguity and noise effect to select the best labels for all objects. In the biclustering process of microarray data, many sub-biclusters (sub-matrices) are first detected using the HT. In the next step, the problem is how to merge the small-sized sub-biclusters into larger ones. One of the contributions of our work here is that we have mapped the procedure of merging sub-biclusters into a relaxation framework that can deal with noise and outliers effectively. In the merging step of the proposed algorithm, we consider the expression values in a microarray data matrix as objects and label them based on the distance of points to the detected hyperplanes. According to this criterion, outliers or noisy points with large distances to their corresponding hyperplanes are deleted and the points are close to the hyperplanes are merged into larger sub-biclusters. Thus, consistent and large-sized biclusters can be discovered in this procedure. The details of the criterion are described in Section 4.

The paper is organized as follows. In Section 2, we show that different types of biclusters can be mapped to geometric patterns in the data space. The following two sections present a brief introduction to the HT and the relaxation labeling scheme used in our algorithm. In Section 5, we describe the proposed relaxation-based geometrical biclustering (RGBC) algorithm in details. In Section 6, the characteristics of the algorithm are studied using simulated and real microarray data. Discussions and conclusions are given in Section 7.

## 2. Geometric models of biclusters

In this section, we demonstrate the relations between biclusters and their corresponding hyperplanes from a geometric perspective and we discuss their properties in the complete and sub-dimensional data spaces.

We will work on a gene expression matrix $D_{N\times M}$ with $N$ genes and $M$ experimental conditions. As an example, we demonstrate one matrix in Fig. 1(a) where different intensity values are represented with different gray levels. Traditional clustering methods attempt to group objects (genes or conditions) into different categories to uncover any hidden local patterns embedded in the matrix. However, if we try to cluster $D_{N\times M}$ using all measurements, we would not uncover any useful patterns although they actually exist in $D_{N\times M}$. By relaxing the constraint that related objects must behave similarly across all measurements, such patterns can be uncovered readily as demonstrated in Fig. 1(b), where the rows and columns of $D_{N\times M}$ are appropriately permuted in order to show the specific local pattern clearly.

Biclustering performs clustering in the gene and condition dimensions simultaneously. A bicluster is regarded as a subset of genes that exhibit similar biological functions under a subset of experiment conditions [4]. Denoting the row and column indices of $D_{N\times M}$ as $G=\{g_1,\ldots,g_N\}$ and $C=\{c_1,\ldots,c_M\}$, we have $D=(G,C)\in\Re^{N\times M}$. We define $sB=(X,Y)$ as a sub-matrix of $D$, where $X=\{N_1,\ldots,N_x\}\subseteq G$ and $Y=\{M_1,\ldots,M_y\}\subseteq C$. The sub-matrix $sB$ is called a sub-bicluster if it contains a coherent pattern defined below. A maximal sub-bicluster is called a bicluster if and only if no $sB'$ exists such that $sB\subset sB'$ (that is $X\subset X'$ or $Y\subset Y'$) [29]. Five types of biclusters are reviewed by Madeira and Oliveira [4], representing five coherent patterns in the two dimensional data space, including constant, constant rows, constant columns, additive and multiplicative ones, corresponding to different biological phenomena [4].

Various biclustering algorithms are proposed to identify particular types of biclusters. Most of these algorithms employ data mining techniques to search for the best possible sub-matrices. The general strategy in all these algorithms can be described as permuting rows and/or columns of the data matrix such that an appropriate merit function is improved by the action. Obviously, the form of the merit function depends on the types of bicluster patterns to be uncovered.

In contrast to the existing permutation-based approach, a novel geometric perspective for the biclustering problem is proposed in [26]. In this new viewpoint, sub-matrices become points in the high dimensional data space. Instead of searching for coherent pattern $B$ in $D$ by permutation, the biclustering problem is transformed to the detection of geometric structures formed by spatial arrangement of these data points. This perspective provides a unified formulation for extracting different types of biclusters simultaneously. Furthermore, the geometric view makes it possible to perform biclustering using generic line or plane finding algorithms.

For example, the condition set $Y$ in $B=(X, Y)$ spans a $\|Y\|$-dimensional space, and the expression of every gene in $X$ corresponds to a point in the space. In such data space, five types of biclusters, including constant, constant rows, constant columns, additive and multiplicative ones [4], can be uniquely mapped to five linear structures, such as points, lines or planes. In general, the five types of biclusters can be expressed using the linear equation $\sum_i a_i x_i=0$. For example, a multiplicative bicluster can be represented as $x_i=a_{ij}x_j$. This means that the expression level of a gene involved under one condition is always proportional to the expression level of the gene under another condition. In Fig. 1(c), an example of