Bicluster Analysis for Coherent Pattern Discovery

D

Alan Wee-Chung Liew
School of Information and Communication Technology, Griffith University, Australia

Xiangchao Gan
Max Planck Institute for Plant Breeding Research, Germany

Ngai Fong Law
Hong Kong Polytechnic University, Hong Kong

Hong Yan
City University of Hong Kong, Hong Kong

INTRODUCTION

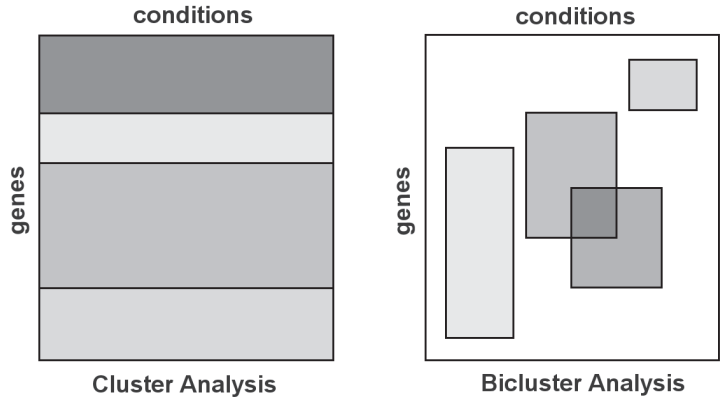
In unsupervised data mining, one is usually interested in discovering groups of data that exhibit certain kind of coherency. A classical technique for unsupervised data partitioning is cluster analysis, where objects are sorted into groups in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Cluster analysis has been applied to many classification problems. In (Wu, Liew, & Yan, 2004), clustering is applied to find natural groupings in the data. In (Borland, Hirschberg, & Lye, 2001), clustering is used for data reduction, where a group of similar objects is summarized by a representative sample in the group. Recently, clustering has been applied extensively in gene expression data analysis. In gene expression data, the objects along the row dimension correspond to genes or some DNA sequence, and the attributes in the column dimension correspond to cDNA microarray experiments or time point samples. Clustering in the row direction, or gene-wise clustering, has been done, for example, on the Yeast gene expression data and human cell (Spellman, Sherlock, Zhang, et al., 1998; Eisen, Spellman, Brown, & Botstein, 1998), whereas clustering in the column direction, or sample-wise clustering, has been done, for example, on cancer type classification (Golub, Slonim, Tamayo, et al., 1999) (Klein, Tu, Stolovitzky, et al., 2001). However, in many real world data, not all attributes of an object are relevant in grouping the objects into meaningful classes. In many cases, some attributes are relevant to only some of the clusters and different clusters may have different relevant subsets of attributes. By relaxing the constraint that related objects must behave similarly across the entire set of attributes, biclustering considers only a relevant subset of attributes when looking for similarity between objects. In this article, we give an overview of the biclustering problem, discuss some common biclustering algorithms, and highlight some interesting applications of biclustering.

BACKGROUND

The goal of biclustering is to find sub-matrices in the dataset, i.e. subsets of objects and subsets of attributes, where the subset of objects exhibits significant homogeneity within the subset of attributes. Figure 1 shows the fundamental difference between clustering and biclustering. Unlike clusters in row-wise or column-wise clustering, biclusters can overlap. In principle, the subsets of attributes for various biclusters can be different. Two biclusters can share some common objects and attributes, and some objects may not belong to any bicluster at all. Due to this flexibility, biclustering has attracted intense interests in the scientific community as a data exploration tool in many fields, ranging from bioinformatics to text mining and marketing.

DOI: 10.4018/978-1-4666-5888-2.ch159

Figure 1. Conceptual difference between cluster analysis (left) and bicluster analysis (right). Different shade of grey denotes different clusters/biclusters, except for the right where it can denote overlapping region of two biclusters.



MODEL OF BICLUSTER PATTERNS

Let a dataset of M objects and N attributes be represented by a rectangular matrix D of M rows and N columns. A bicluster is a subset of rows that exhibit similar behaviors across a subset of columns and vice versa. The bicluster $B=(X, Y)$, therefore, appears as a sub-matrix of D, where the set of row indices X and column indices Y are subsets of M and N, respectively. Biclustering aims to discover a set of biclusters $B_k = (X_k, Y_k)$ such that each bicluster satisfies some notion of homogeneity.

Many bicluster patterns have been proposed in the literature (Zhao, Liew, Wang & Yan, 2012). Some of the most common bicluster patterns are: (a) bicluster with constant values, (b) bicluster with constant values in rows or columns, (c) bicluster with coherent values including additive or multiplicative models, (d) bicluster with coherent evolution. The first three types of biclusters deal with numerical values in the data matrix and try to find subsets of rows and columns with similar behaviors. The bicluster with coherent evolution aims to find coherent patterns, i.e., trends, regardless of the exact numeric values in the data matrix. Figure 2 shows the first three types of numerical biclusters that are hidden in a 6×6 data matrix. Although bicluster with coherent evolution does not deal with numerical values explicitly, the problem can nevertheless be transformed into a numerical one by appropriately quantizing the original data matrix into a binary matrix.

TECHNIQUES FOR BICLUSTER ANALYSIS

Existing biclustering algorithms can be classified into one of the following classes, depending on the bicluster model and the search strategy used. For a comprehensive review, see (Zhao, Liew, Wang, & Yan, 2012).

Distance Based Biclustering

Distance based biclustering uses a distance metric to measure the quality of the biclusters, and performs an iterative search for the biclusters by minimizing the residual sum of squares cost. This class of biclustering algorithms is among the earliest biclustering methods proposed in the literature and is widely used in many applications (Hartigan, 1972). In the well-known d-biclustering algorithm of Cheng and Church's (Cheng & Church, 2000), the following mean squared residue score is minimized

$$H(X,Y) = \frac{1}{|X||Y|} \sum_{i \in X, j \in Y} \left(a_{ij} - a_{iY} - a_{Xj} + a_{XY} \right)^2 \quad (1)$$

where a_{iY}, a_{Xj}, a_{XY}, are the row mean, column mean, and the mean in the submatrix $B=(X, Y)$, respectively. A bicluster is called a d-bicluster if $H(X, Y) £ d$ for some $d > 0$. In Cheng and Church's algorithm,