Available online at www.sciencedirect.com
ScienceDirect
Computers & Operations Research 35 (2008) 2964–2987
computers & operations research
www.elsevier.com/locate/cor

Biclustering in data mining

Stanislav Busygin[a], Oleg Prokopyev[b,*], Panos M. Pardalos[a]

[a]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
[b]Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA

Available online 6 February 2007

Abstract

Biclustering consists in simultaneous partitioning of the set of samples and the set of their attributes (features) into subsets (classes). Samples and features classified together are supposed to have a high relevance to each other. In this paper we review the most widely used and successful biclustering techniques and their related applications. This survey is written from a theoretical viewpoint emphasizing mathematical concepts that can be met in existing biclustering techniques.

Keywords: Data mining; Biclustering; Classification; Clustering; Survey

1. The main concept

Due to recent technological advances in such areas as IT and biomedicine, the researchers face ever-increasing challenges in extracting relevant information from the enormous volumes of available data [1]. The so-called *data avalanche* is created by the fact that there is no concise set of parameters that can fully describe a state of real-world complex systems studied nowadays by biologists, ecologists, sociologists, economists, etc. On the other hand, modern computers and other equipment are able to produce and store virtually unlimited data sets characterizing a complex system, and with the help of available computational power there is a great potential for significant advances in both theoretical and applied research. That is why in recent years there has been a dramatic increase in the interest in sophisticated *data mining* and *machine learning* techniques, utilizing not only statistical methods, but also a wide spectrum of computational methods associated with large-scale optimization, including algebraic methods and neural networks.

The problems of partitioning objects into a number of groups can be met in many areas. For instance, the vector partition problem, which consists in partitioning of n d-dimensional vectors into p parts has broad expressive power and arises in a variety of applications ranging from economics to symbolic computation (see, e.g., [2–4]). However, the most abundant area for the partitioning problems is definitely data mining. Data mining is a broad area covering a variety of methodologies for analyzing and modeling large data sets. Generally speaking, it aims at revealing a genuine similarity in data profiles while discarding the diversity irrelevant to a particular investigated phenomenon. To analyze patterns existing in data, it is often desirable to partition the data samples according to some similarity criteria. This task is called *clustering*. There are many clustering techniques designed for a variety of data types—homogeneous and

[*] Corresponding author.
E-mail addresses: busygin@ufl.edu (S. Busygin), prokopyev@engr.pitt.edu (O. Prokopyev), pardalos@ufl.edu (P.M. Pardalos).

0305-0548/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.cor.2007.01.005

nonhomogeneous numerical data, categorical data, 0–1 data. Among them one should mention hierarchical clustering [5], k-means [6], self-organizing maps (SOM) [7], support vector machines (SVM) [8,9], logical analysis of data (LAD) [10,11], etc. A recent survey on clustering methods can be found in [12].

However, working with a data set, there is always a possibility to analyze not only properties of samples, but also of their components (usually called *attributes* or *features*). It is natural to expect that each associated part of samples recognized as a cluster is induced by properties of a certain subset of features. With respect to these properties we can form an associated cluster of features and bind it to the cluster of samples. Such a pair is called a *bicluster* and the problem of partitioning a data set into biclusters is called a *biclustering* problem.

In this paper we review the most widely used and successful biclustering techniques and their related applications. Previously, there were published few surveys on biclustering [13,14], as well as a Wikipedia article [15]. However, we tried to write this survey from a more theoretical viewpoint emphasizing mathematical concepts that can be found in existing biclustering techniques. In addition this survey discusses recent developments not included in the previous surveys and includes references to public domain software available for some of the methods and most widely used benchmarks data sets.

2. Formal setup

Let a data set of n samples and m features be given as a rectangular matrix $A = (a_{ij})_{m\times n}$, where the value a_{ij} is the expression of ith feature in jth sample. We consider classification of the samples into classes

$$\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_r, \quad \mathscr{S}_k \subseteq \{1, \ldots, n\}, \quad k = 1, \ldots, r,$$
$$\mathscr{S}_1 \cup \mathscr{S}_2 \cup \cdots \cup \mathscr{S}_r = \{1, \ldots, n\},$$
$$\mathscr{S}_k \cap \mathscr{S}_\ell = \emptyset, \quad k, \ell = 1, \ldots, r, \ k \neq \ell.$$

This classification should be done so that samples from the same class share certain common properties. Correspondingly, a feature i may be assigned to one of the feature classes

$$\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_r, \quad \mathscr{F}_k \subseteq \{1, \ldots, m\}, \quad k = 1, \ldots, r,$$
$$\mathscr{F}_1 \cup \mathscr{F}_2 \cup \cdots \cup \mathscr{F}_r = \{1, \ldots, m\},$$
$$\mathscr{F}_k \cap \mathscr{F}_\ell = \emptyset, \quad k, \ell = 1, \ldots, r, \ k \neq \ell,$$

in such a way that features of the class $\mathscr{F}_k$ are "responsible" for creating the class of samples $\mathscr{S}_k$. Such a simultaneous classification of samples and features is called *biclustering* (or *co-clustering*).

Definition 1. A *biclustering* of a data set is a collection of pairs of sample and feature subsets $\mathscr{B} = ((\mathscr{S}_1, \mathscr{F}_1)$ $(\mathscr{S}_2, \mathscr{F}_2), \ldots, (\mathscr{S}_r, \mathscr{F}_r))$ such that the collection $(\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_r)$ forms a partition of the set of samples, and the collection $(\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_r)$ forms a partition of the set of features. A pair $(\mathscr{S}_k, \mathscr{F}_k)$ will be called a *bicluster*.

It is important to note here that in some of the biclustering methodologies a direct one to one correspondence between classes of samples and classes of features is not required. Moreover, the number of sample and feature classes is allowed to be different. This way we may consider not only pairs $(\mathscr{S}_k, \mathscr{F}_k)$, but also other pairs $(\mathscr{S}_k, \mathscr{F}_\ell)$, $k \neq \ell$. Such pairs will be referred to as *co-clusters*. Another possible generalization is to allow *overlapping* of co-clusters.

The criteria used to relate clusters of samples and clusters of features may have different nature. Most commonly, it is required that the submatrix corresponding to a bicluster either is overexpressed (i.e., mostly includes values above average), or has a lower variance than the whole data set, but in general, biclustering may rely on any kind of common patterns among elements of a bicluster.

3. Visualization of biclustering

One popular tool for visualizing data sets is *heatmaps*. A heatmap is a rectangular grid composed of pixels each of which corresponds to a data value. The color of a pixel ranges between bright green or blue (lowest values) and bright red (highest values) visualizing the corresponding data value. This way, if the samples or/and features of the data set are ordered with respect to some pattern in the data, the pattern becomes obvious to observe visually.