Pattern Recognition 44 (2011) 252–264

Contents lists available at ScienceDirect

Pattern Recognition

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

Stability-based validation of bicluster solutions

Youngrok Lee, Jeonghwa Lee, Chi-Hyuck Jun *

Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 24 December 2009
Received in revised form
22 July 2010
Accepted 25 August 2010

Keywords:
Biclustering
Validation
Stability
Resampling

ABSTRACT

Bicluster analysis is an unsupervised learning method to detect homogeneous or uniquely characterized two-way subsets of objects and attributes from a data set. It is useful in finding groups that may not be found by the traditional cluster analysis and in interpreting the groups intuitively, especially for high-dimensional data sets. Because of these advantages, over the last few years, various biclustering algorithms have been developed and applied to bioinformatics and text mining area. However, research into validation of bicluster solutions is rare. We propose a new procedure of validating bicluster solutions by developing a stability index to measure the reproducibility of the solution under variation in the input data set. By generating random resample data sets from the input data set, obtaining bicluster solutions from them, and evaluating the expected agreement of the solutions to the bicluster solution for the original input data set, we quantify the stability of the bicluster solution. Experiments using three artificial data sets and two real gene expression data sets indicate that the proposed method is suitable to validate bicluster solutions.

1. Introduction

Biclustering is a method of extracting significant subsets of objects (or observations) and attributes (or features) from a data set. The method has many names including direct clustering [1], simultaneous clustering, two-way clustering, two-mode clustering, co-clustering, and block clustering, but the term 'biclustering' [2] has been most widely used. Bicluster analysis is similar to conventional one-way cluster analysis (called just cluster analysis) but the two approaches have several important differences. First, biclustering methods detect objects which are similar for only a subset of attributes, whereas clustering methods group objects which are similar over all attributes [29]. For example, in bicluster analysis, if several objects are very similar in 20% of their attributes, they may be grouped together even though they are very different in the remaining 80%. Furthermore, biclustering is based on detection and selection, whereas clustering is usually based on partitioning. This means that, in bicluster analysis, an object or an attribute can be a member of more than one bicluster or of no bicluster. This is a big difference from clustering, which usually assigns an object to exactly one cluster. For these reasons, biclusters tend to be more homogeneous than conventional clusters, and biclustering methods may find groups that cannot be detected by cluster analysis. Especially, biclustering is beneficial to the analysis of high-dimensional data sets because it is relatively free from the curse of dimensionality. As more databases are accumulated in various areas due to the development of data warehousing and computing technology, usefulness of bicluster analysis is increasing.

Bicluster analysis proceeds by four steps – data preprocessing, algorithm design or selection, validation of results, and interpretation – [3], as in the traditional clustering procedure [4,5]. The third step, validation, is very important because interpreting an invalid solution may be a waste of time. However, little attention has been given to quantitative validation of bicluster solutions. Therefore, developing a new validation method of bicluster solutions would be helpful in finding valuable information through bicluster analysis. A suitable validation measure would help the selection of biclustering algorithm and the related parameters.

Many researchers have tried to develop validation measures in cluster analysis. Traditionally, validation indices of cluster solutions have been classified into internal indices and external indices [6,4,7]. However, conventional external and internal indices have their own restriction. External indices cannot be used without prior categorical labels, and internal indices are biased to their own definition of structure of clusters [7]. Over the past few years, a considerable number of studies have been made on stability-based validation for cluster solutions [8–12]. The stability is defined as the consistency of cluster solutions obtained from repeatedly resampled or perturbed data [7]. Stability indices are clearly not external measures because they do not use prior label information. On the other hand, they are quite different from internal indices mentioned above because they are not restricted to a specific structure of clusters. Therefore,

* Corresponding author.
E-mail addresses: lyr1004@postech.ac.kr (Y. Lee), bls83@postech.ac.kr (J. Lee), chjun@postech.ac.kr (C.-H. Jun).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.08.029

stability measures are considered as good alternatives to validate cluster solutions when we cannot get enough information about true labels or structures of true clusters.

In the area of biclustering, only a few studies have been conducted to develop validation methods for bicluster solutions [3]. Most of them have proposed internal and external indices which also have limitation described above. Furthermore, even though several studies have addressed stability in bicluster analysis [13–15], they have focused on the robustness of biclustering algorithms, not on stability of bicluster solutions. As far as we know, a stability-based validation measure has not been introduced to biclustering.

The objective of this research is to develop a novel method for the validation of bicluster solutions based on stability. We propose a resampling-based framework to validate the results of bicluster analysis by evaluating their stabilities. Using this framework, we derive a normalized stability index that measures the reproducibility of a bicluster solution. To demonstrate the effectiveness of the proposed method, experiments are performed with three artificial data sets and two yeast data sets. Also, the issues of algorithm selection and parameter optimization through the proposed method will be discussed.

2. Validation measures

2.1. Notation

Let

$$X = (x_{ij}) \tag{1}$$

be the input data matrix with n objects and m attributes. By grouping objects and attributes simultaneously based on similarities, a bicluster solution of X can be produced. Let M be a bicluster solution which consists of K biclusters and B_k be the kth bicluster of M. Also, let O_k and A_k be the set of objects and attributes belonging to B_k, respectively. Then, B_k and M can be defined as

$$B_k = (O_k \times A_k), \tag{2}$$

$$M = \{B_1, \ldots, B_K\}. \tag{3}$$

The intersection of two biclusters B_i and B_j can be defined as a combination of the intersection of two object sets and the intersection of two attribute sets:

$$B_i \cap B_j = ((O_i \cap O_j) \times (A_i \cap A_j)). \tag{4}$$

However, the union of two biclusters is not defined as a combination of the unions. In addition, the size of a bicluster can be defined as

$$|B_k| = |O_k| \times |A_k|. \tag{5}$$

2.2. Internal index

Internal indices evaluate the similarity between objects or attributes belonging to a bicluster or evaluate the fitness of a bicluster to a specific model. They use information only intrinsic to input data and to a bicluster solution.

We can evaluate the coherence of a bicluster based on how well it is fitted to a model. For example, an additive model defines an element value of object i and attribute j in a bicluster as

$$a_{ij} = a_{Ij} + a_{iJ} - a_{IJ} + r_{ij}, \tag{6}$$

where a_{Ij} is the mean of attribute j in the bicluster, a_{iJ} is the mean of object i in the bicluster, a_{IJ} is the mean of all elements in the bicluster, and r_{ij} is a residue which is not explained by the bicluster [16]. The mean squared residue of a bicluster, which is defined as

$$MSR(B_k) = \frac{1}{|B_k|} \sum_{i \in O_k, j \in A_k} r_{ij}^2 \tag{7}$$

represents the coherence of the bicluster. In ideal cases, $MSR(B_k)$ is equal to zero. We can quantify the quality of M by using the average residue defined as

$$ASR(M) = \frac{1}{K} \sum_{k=1}^{K} MSR(B_k). \tag{8}$$

Another way to evaluate bicluster solutions without external comparison sets is to compare distance between objects or attributes within a bicluster with distance between all objects and attributes in input data. Small distance within a bicluster compared to the average distance of input data indicates that the bicluster is homogeneous and coherent. Γ statistic [6] of clustering has been modified to evaluate bicluster solutions [17,3]. Let $P_{(k)}$ be the proximity matrix of objects within attributes of B_k so that $P_{(k)ij}$ denotes the distance between objects i and j within A_k. Also, let $C_{(k)}$ be the co-membership matrix that $C_{(k)ij} = 0$ if both objects i and j are in O_k, and $C_{(k)ij} = 1$ otherwise. Then, if P_k and C_k are strongly positively correlated, we may tell that O_k is a characteristic object group within A_k. A statistic evaluating O_k can be defined as

$$\Gamma_O(B_k) = \frac{2}{n(n-1)} \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (P_{(k)ij} - \mu_{p_k})(C_{(k)ij} - \mu_{c_k})}{\sigma_{p_k}\sigma_{c_k}}, \tag{9}$$

where $\mu_{p_k}(\mu_{c_k})$ and $\sigma_{p_k}(\sigma_{c_k})$ are the mean and the standard deviation of $P_k(C_k)$, respectively. In the same manner, A_k can be evaluated by $\Gamma_A(B_k)$ where P_k and C_k are constructed on attributes. Then, Γ statistic of M can be defined as

$$\Gamma(M) = \sum_{k=1}^{K} \frac{\Gamma_O(B_k) + \Gamma_A(B_k)}{2K}. \tag{10}$$

Obviously, values of $\Gamma(M)$ depend on the distance measure used to construct the proximity matrix P_k. $\Gamma(M)$ lies in $[-1,1]$, and large $\Gamma(M)$ is preferred.

2.3. External index

External measures are used to compare two bicluster solutions. Suppose that two bicluster solutions exist: M_1 having K_1 biclusters and M_2 having K_2 biclusters which can be represented as

$$M_i = \{B_1^{(i)}, \ldots, B_{K_i}^{(i)}\}, \quad i = 1,2, \tag{11}$$

where $B_j^{(i)}$ denotes the jth bicluster in the ith solution.

Several metrics to evaluate similarity between M_1 and M_2 have been proposed [17–19,3]. Let s be a metric to evaluate the similarity between two biclusters. The following metrics have been defined to compare two biclusters:

$$s_{Prelic}(B_i,B_j) = \frac{|O_i \cap O_j|}{|O_i \cup O_j|}, \tag{12}$$

$$s_{Liu\&Wang}(B_i,B_j) = \frac{|O_i \cap O_j| + |A_i \cap A_j|}{|O_i \cup O_j| + |A_i \cup A_j|}, \tag{13}$$

$$s_{Jaccard}(B_i,B_j) = \frac{|B_i \cap B_j|}{|B_i| + |B_j| - |B_i \cap B_j|}, \tag{14}$$