Computational Statistics and Data Analysis 55 (2011) 1969–1979

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On biological validity indices for soft clustering algorithms for gene expression data

Han-Ming Wu *

Department of Mathematics, Tamkang University, Taipei County 25137, Taiwan

ARTICLE INFO

Article history:
Received 5 October 2009
Received in revised form 29 September 2010
Accepted 13 December 2010
Available online 21 December 2010

Keywords:
Biological validity indices
Fuzzy clustering
Fuzzy cluster validity
Fuzzy c-means
Gene expression
Microarray data analysis
Soft clustering

ABSTRACT

Unsupervised clustering methods such as K-means, hierarchical clustering and fuzzy c-means have been widely applied to the analysis of gene expression data to identify biologically relevant groups of genes. Recent studies have suggested that the incorporation of biological information into validation methods to assess the quality of clustering results might be useful in facilitating biological and biomedical knowledge discoveries. In this study, we generalize two bio-validity indices, the biological homogeneity index and the biological stability index, to quantify the abilities of soft clustering algorithms such as fuzzy c-means and model-based clustering. The results of an evaluation of several existing soft clustering algorithms using simulated and real data sets indicate that the soft versions of the indices provide both better precision and better accuracy than the classical ones. The significance of the proposed indices is also discussed.

1. Introduction

Unsupervised clustering methods have been widely applied to the analysis of gene expression data to identify biologically relevant groups of genes. Algorithms such as K-means, hierarchical clustering and self-organizing maps (SOM) are the popular *hard* clustering methods, which assign each gene to only one cluster. However, this restriction may not be appropriate for analyzing gene expression profiles because a single gene might be involved in multiple functional categories. In contrast, a *soft* clustering algorithm, such as fuzzy c-means, assigns one gene to multiple clusters according to their degrees of membership and thus gives more information on gene multi-functionalities (Dembele and Kastner, 2003).

To assess the quality and reliability of the clusters produced by a clustering algorithm, a variety of cluster validity indices have been proposed such as the Dunn index (Dunn, 1973), the adjusted Rand index (Hubert and Arabie, 1985), the silhouette width (Rousseeuw, 1987), the figure of merit (FOM) (Yeung et al., 2001), Hennig's stability index (Hennig, 2007) (all primarily intended for hard clusters), and the Xie–Beni index (Xie and Beni, 1991) and Qiu and Joe's separation index (Qiu and Joe, 2006) (primarily for fuzzy clusters). Some of these indices are the most popular statistical indices for gene expression data analysis. Refer to Handl et al. (2005) for an overview of cluster validation measures for post-genomic data. On the other hand, recent studies have suggested that the incorporation of biological information, such as functional and/or curated annotations, into validation methods might be useful for supporting biological and biomedical discoveries. For example, Bolshakova et al. (2005) presented a knowledge-driven approach for assessing the cluster validity based on similarities extracted from gene ontology (GO). Park et al. (2005) used a genetic algorithm for fuzzy clustering and prior knowledge of experimental data for their evaluation. Datta and Datta (2006) made use of the biological information along with gene

* Tel.: +886 2 2621 5656x3147; fax: +886 2 2620 9916.
E-mail address: hmwu@mail.tku.edu.tw.

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.12.003

expression data and proposed two indices, the biological homogeneity index (BHI) and the biological stability index (BSI), for the biological evaluation of clustering algorithms. In summary, these indices attempt to produce clustering results with good statistical properties, such as compactness, well-separatedness, connectedness and stability, and also attempt to provide more biologically relevant results.

For soft clustering algorithms, a common way to biologically evaluate the resulting class memberships is to select the class label with the highest membership value so that the existing indices can be applied. However, bio-indices that are designed specifically for hard clustering methods may not be sufficient when applied to the soft clustering of genes since some genes may belong to several statistical clusters. A biological evaluation index that considers the class membership is necessary for the correct evaluation of soft clustering algorithms. In the present study, we generalize the BHI and the BSI to quantify the abilities of soft clustering algorithms to produce biologically meaningful clusters using a reference set of functional classes. The indices investigated are the soft biological homogeneity index (SBHI) and the soft biological stability index (SBSI). To the best of our knowledge, no biological validation indices have been proposed thus far for soft clustering algorithms.

We evaluated the performances of several existing soft clustering algorithms in R (R Development Core Team, 2006), including fuzzy c-means clustering (Bezdek, 1981), fuzzy c-shell clustering (Dave, 1996), model-based clustering (Fraley and Raftery, 2002) and consensus clustering (Monti et al., 2003), on two simulated and three gene expression data sets and identified the optimal algorithm for each number of clusters. A biological reference set for the annotated genes of the relevant species was obtained from the gene ontology database. The proposed indices are helpful for selecting the optimal algorithm from a class of soft clustering algorithms for the given data sets. We provided an R code for computing these indices, whereas the function fclustIndex in package **e1071** (Dimitriadou et al., 2006) has several statistical validation measures for fuzzy clusters.

This paper is structured as follows. Section 2 introduces some soft clustering methods to be evaluated, which are available in R. The classical biological validation indices for hard clusters, the BHI and the BSI, are given in Section 3. Section 4 describes the generalized indices, the SBHI and the SBSI, for soft clustering algorithms and discusses their significance. Section 5 provides examples using two simulated and three gene expression data sets. We then conclude the paper in Section 6.

2. Soft clustering methods

In this section, we give a brief description of several existing soft clustering methods and their availability in R. These algorithms will be evaluated by the proposed indices. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the data set to be analyzed, consisting of n data points in p-dimensional space, and let $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ be a set of K cluster centers. Let $U = \{u_{ij}\}$ be an $n \times K$ fuzzy partition matrix where u_{ij} is the membership degree of a point $\mathbf{x}_i$ in cluster j. Let m be a fuzziness index (often $m = 2$).

2.1. Fuzzy c-means clustering (FCM)

Fuzzy c-means clustering (Bezdek, 1981) is the most widely used soft clustering algorithm. It minimizes the following objective function:

$$J_K = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m \|\mathbf{x}_i - \mathbf{c}_k\|^2. \tag{1}$$

The fuzzy clustering is conducted through an alternating process between the membership u_{ik} and the fuzzy cluster centers $\mathbf{c}_k$:

$$u_{ik}^m = \left(\sum_{j=1}^{K} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_k\|}{\|\mathbf{x}_i - \mathbf{c}_j\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \qquad \mathbf{c}_k = \frac{\sum_{i=1}^{n} u_{ik}^m \cdot \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^m}. \tag{2}$$

The iteration stops when a fixed number of iterations is reached or $\max_{ik} |u_{ik}^{(t+1)} - u_{ik}^{(t)}| < \epsilon$, where ϵ is a termination tolerance, and t is the iteration step. The fuzzy c-means algorithm is available in the **e1071** package as the function cmeans.

Kaufman and Rousseeuw (1990) proposed another fuzzy clustering algorithm, called fanny, which aims to minimize the following objective function:

$$\sum_{k=1}^{K} \left(\sum_{i,j}^{n} u_{ik}^m u_{jk}^m \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right) \Big/ \left(2 \sum_{j=1}^{n} u_{jk}^m \right).$$

The fanny algorithm is available in the R package **cluster**.

2.2. Fuzzy c-shell clustering

The first algorithm proposed for fuzzy c-shell clustering was introduced by Dave (1996) for the recognition of circle contours. Each cluster in the algorithm is characterized by its center $\mathbf{v}$ and radius r. Fuzzy c-shell clustering is conducted by