Pattern Recognition Letters 32 (2011) 221–234

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Validity index for clusters of different sizes and densities

Krista Rizman Žalik *, Borut Žalik

University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, SI-2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 13 February 2009
Available online 18 September 2010
Communicated by T.K. Ho

Keywords:
Clustering
k-Means clustering
Unsupervised classification
Validity index

ABSTRACT

Cluster validity indices are used to validate results of clustering and to find a set of clusters that best fits natural partitions for given data set. Most of the previous validity indices have been considerably dependent on the number of data objects in clusters, on cluster centroids and on average values. They have a tendency to ignore small clusters and clusters with low density. Two cluster validity indices are proposed for efficient validation of partitions containing clusters that widely differ in sizes and densities. The first proposed index exploits a compactness measure and a separation measure, and the second index is based an overlap measure and a separation measure. The compactness and the overlap measures are calculated from few data objects of a cluster while the separation measure uses all data objects. The compactness measure is calculated only from data objects of a cluster that are far enough away from the cluster centroids, while the overlap measure is calculated from data objects that are enough near to one or more other clusters. A good partition is expected to have low degree of overlap and a larger separation distance and compactness. The maximum value of the ratio of compactness to separation and the minimum value of the ratio of overlap to separation indicate the optimal partition. Testing of both proposed indices on some artificial and three well-known real data sets showed the effectiveness and reliability of the proposed indices.

1. Introduction

Clustering is one of the most important tasks in data analysis. It has been used for decades in image processing and pattern recognition. Clustering is a process of dividing a set of given data into groups, or clusters, such that all data in the same group are similar to each other, while data from different clusters are dissimilar (Hartigan, 1975; Jain et al., 1999). The conventional (hard or crisp) clustering methods put each data object to exactly one cluster. Fuzzy sets introduce the idea of allowing a membership function that defines the membership values of all data objects to all clusters, and the development of fuzzy clustering methods. The result of any fuzzy clustering algorithm is a partition of data objects into k clusters $(C_1, C_2, \ldots, C_k)$, of a given data set X consisting of n data objects $X = \{x_1, \ldots, x_n\}$. How strong each data object x_j belongs to the ith cluster C_i is described with a membership value u_{ij}. The result of any fuzzy clustering algorithm is a partition matrix $U(X)$ with size $k \times n$: $U = [u_{ij}]$, $i = 1, \ldots, k$ and $j = 1, \ldots, n$. In the crisp partitioning of the data, the following condition holds: $u_{ij} = 1$ if $x_j \in C_i$, otherwise $u_{ij} = 0$. In fuzzy clustering: $u_{ij} \in [0; 1]$ and each membership value u_{ij} denotes the grade of membership of the jth element to the ith cluster. The sum of all membership values for each data object is 1.

A wide variety of clustering algorithms have been proposed for different applications (Jain and Dubes, 1988). The *k-means* algorithm (Hartigan, 1975) is by far the most popular clustering method. Each cluster is represented by the mean (or weighted average) of its data objects, the cluster centroid. The sum of discrepancies between an object and its centroid, expressed through an appropriate distance, is used as the objective function. The *k*-means tries to minimize the total intra-cluster variance or the squared error function J.

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - v_i)^2 \tag{1}$$

where k denotes number of clusters (C_i, $i = 1, 2, \ldots, k$) x_j is the data object and v_i is the centroid or mean object of all the objects x_j from the cluster C_i.

Fuzzy *c-means* (FCM) is a clustering method which allows one piece of data to belong to two or more clusters. This method developed by Dunn (1973) and improved by Bezdek (1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} u_{ij}^m (x_j - v_i)^2; \quad 1 \le m \le \infty \tag{2}$$

where u_{ij} is the degree of membership of x_i in the cluster C_i, x_j is the jth of d-dimensionally measured data, v_i is the d-dimensionally measured centroid of the cluster C_i.

* Corresponding author. Tel.: +386 2 251 81 80.
E-mail address: krista.zalik@uni-mb.si (K.R. Žalik).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.08.007

The presence of large variability in cluster geometric shapes, densities and sizes and the number of clusters that cannot always be known a priori, are major difficulties when clustering. Many clustering algorithms (also *k*-means and fuzzy *c*-means) require the user to predefine the number of clusters before the clustering process. However, it is sometimes impossible to know the number of clusters in advance. The clustering results depend on the choice regarding number of clusters. Determining the appropriate number of clusters and the validity of the obtained partitioning are two fundamental problems in clustering. Finding the optional number of clusters that best fits the natural partition for given data set is difficult, since for the same data set several partitions exists depending on the level of details. Validity indices are often used for accessing the optimal number of clusters. This requires a clustering algorithm to be executed several times, with a different number of clusters in each run. Another alternative to identifying the correct number of clusters is to improve the optimization function and discover the number of clusters dynamically during execution of the clustering algorithm that satisfies the new optimization function (Žalik, 2008). A new function is required, because an objective function of *k*-means clustering, that summarizes any discrepancies between an object and its centroid, monotonically decreases with any increase in cluster numbers. Therefore, it cannot be used as an objective function for determining the correct number of clusters.

Cluster analysis contributes to engineering applications only when cluster validity is measured. Cluster validity indices have been widely used to validate partitions produced by clustering algorithms. Validity indices are also often used for accessing the optimal number of clusters. The partitioning that optimizes the considered index is selected as the final result. Most of validation indices proposed during last decades have focused on compactness and separation. Separation is a measure of clusters' isolation from each other and compactness is a measure of closeness of data objects within a cluster. A low value of variance is indicator of closeness. Members of each cluster should be as close to each other as possible and clusters should be widely separated. Most popular validity measures have the tendency to ignore clusters with low density and are not efficient in validation of partitions having different sizes and densities.

The main objective of our research was to design a cluster validity index that is suitable for the validation of partitions having different sizes and densities. Two new cluster validity indices are proposed. First proposed index uses ratio assessment between the two main cluster properties: separation and compactness. Distances between the closest pairs of cluster centroids and the sizes of clusters are taken into account. The second index bases on the ratio between overlap and separation. Both suggested indices do not ignore clusters with low density and small clusters. Experimental results on artificial and well-known real-life data sets indicate that both indices are effective.

The remainder of this paper is organized as follows. In Section 2, we review several popular validity indices. Two clustering validity indices are then proposed in Section 3. Section 4 presents simulation results. Conclusions are drawn in Section 5.

2. Cluster validity indices

Many cluster validity indices have been developed for evaluating quality of partitions with the goal to find optimal partitioning that consists of compact and well-separated clusters (Berry and Linoff, 1996; Halkidi and Vazirgiannis, 2008). Fuzzy cluster validity measures use the membership degrees produced by corresponding fuzzy clustering algorithms. The classical validity indices evaluate the properties of crisp structure imposed on the data by the clustering algorithm. Crisp clustering means having non-overlapping partitions.

Dunn proposed (Dunn, 1974) validity index for crisp clustering. Let there be a data set with n data objects $X = \{x_j;\ j = 1, \ldots, n\}$ partitioned into k clusters $(C_1, C_2, \ldots, C_k)$; each cluster has a centroid v_i $(i = 1, 2, \ldots, k)$. The Dunn's measure DI is defined as

$$DI = \min_{1\le i\le k}\left\{\min_{1\le j\le k, j\neq i}\left\{\frac{\delta(C_i, C_j)}{\max_{1\le l\le k}\Delta(C_l)}\right\}\right\} \tag{3}$$

$$\delta(C_i, C_j) = \min\{d(x_i, x_j) | x_i \in C_i, x_j \in C_j\} \tag{4}$$

$$\Delta(C_l) = \max\{d(x_i, x_j) | x_i, x_j \in C_l\} \tag{5}$$

where d is a distance function and C_i is the set whose elements are assigned to the ith cluster. The main disadvantage of the Dunn's measure is its high computational complexity as k increases. The objective is to maximize DM index for achieving proper clustering and the optimal number of clusters.

Davies–Bouldin's index DB (Davies and Bouldin, 1970) differs from Dunn's index by using the average error for each cluster. DB index is the ratio of cluster scatter $S_{i,q}$ of cluster C_i to cluster separation. Between-cluster scatter $S_{i,q}$ of cluster C_i is defined as

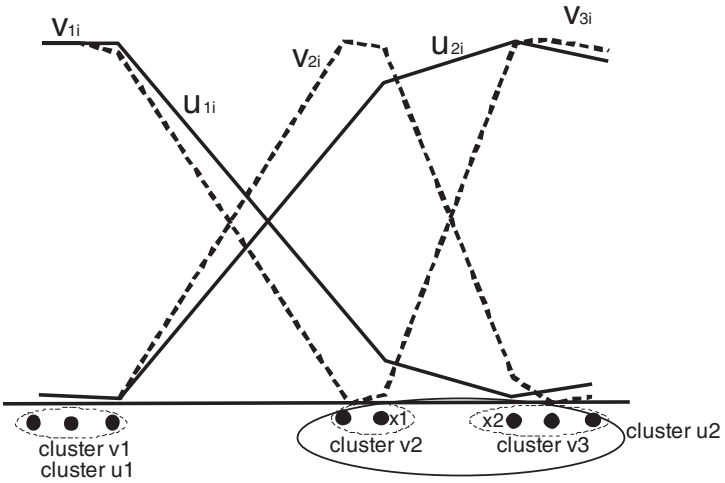

Fig. 1. Data set of eight data objects partitioned into two and three clusters.