Applied Soft Computing 22 (2014) 47–59

Contents lists available at ScienceDirect

Applied Soft Computing

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

Robust validity index for a modified subtractive clustering algorithm

Horng-Lin Shieh*

St. John's University, Taiwan

ARTICLE INFO

Article history:
Received 11 January 2013
Received in revised form 18 February 2014
Accepted 1 May 2014
Available online 10 May 2014

Keywords:
Partition index
Robust
Subtractive clustering (SC) algorithm
Validity index

ABSTRACT

A novel robust validity index is proposed for subtractive clustering (SC) algorithm. Although the SC algorithm is a simple and fast data clustering method with robust properties against outliers and noise; it has two limitations. First, the cluster number generated by the SC algorithm is influenced by a given threshold. Second, the cluster centers obtained by SC are based on data that have the highest potential values but may not be the actual cluster centers. The validity index is a function as a measure of the fitness of a partition for a given data set. To solve the first problem, this study proposes a novel robust validity index that evaluates the fitness of a partition generated by SC algorithm in terms of three properties: compactness, separation and partition index. To solve the second problem, a modified algorithm based on distance relations between data and cluster centers is designed to ascertain the actual centers generated by the SC algorithm. Experiments confirm that the preferences of the proposed index outperform all others.

1. Introduction

Clustering algorithms are widely used to group objects based on attributes of data that describe the objects and their relation to one another. As such, a clustering algorithm aims to partition the data into groups; data with similar attributes are partitioned into one cluster and differentiated from data in other groups [1,2]. Clustering algorithms have been successfully used in data mining, pattern recognition, function approximation, machine learning, and system modeling.

The major clustering algorithms proposed in the literature can be classified as partitioning-based algorithms, hierarchical algorithms, density-based clustering algorithms, grid-based algorithms, model-based algorithms, and conceptual clustering algorithms [3].

The subtractive clustering (SC) algorithm proposed by Chiu [4] is a density-based clustering algorithm based on the density of data points (*i.e.*, "potential values") in the feature space. The core concept of the SC algorithm is to find the regions in the feature space with the highest density of data points. The point with the highest potential value is selected as the center of a cluster. The potential of data points within a prescribed radius are then removed, and the algorithm finds the point with the next highest potential value. This procedure is repeated until a predefined criterion is met.

The SC algorithm is widely used in various real applications. Chen and Gaob [5] adopted SC algorithm for an adaptive network based fuzzy inference system (ANFI) to estimate the train station parking (TSP) error in urban rail transit. In Ref. [6], Bilgin et al. proposed an unsupervised hyperspectral image segmentation method that used a novel subtractive-clustering-based similarity segmentation approach and a novel method of cluster validation by one-class support vector (SV) machine. In Ref. [7], Chen adopted particle swarm optimization (PSO) techniques to obtain appropriate parameter settings for subtractive clustering (SC) and integrated the adaptive-network-based fuzzy inference system (ANFIS) model to construct a model for predicting business failures. In Ref. [8], a fuzzy classifier based on simulated annealing (SA) and subtractive clustering was used to optimize a fuzzy inference system for classification tasks. In Ref. [9], a subtractive based fuzzy inference system is introduced to estimate the potato crop parameters like biomass, leaf area index, plant height and soil moisture. In Ref. [10], Amina et al. proposed a wavelet neural network for prediction of electricity consumption of the power system of the Greek Island of Crete and the SC algorithm had been applied to the definition of fuzzy rules.

Unlike the *K*-means, which require iterations of several epochs, the SC algorithm requires only one pass of the training data. However, the SC algorithm only roughly estimates the cluster centers, since the cluster centers obtained are located at some data points. Moreover, since no cluster validity is used, the clusters produced may not accurately represent the clusters [11]. To overcome the disadvantage of SC algorithm, the solution proposed in this study is to use a robust validity index for a modified SC algorithm.

* Tel.: +886 953 262 773.
E-mail address: shieh@mail.sju.edu.tw

http://dx.doi.org/10.1016/j.asoc.2014.05.001
1568-4946/

This paper is organized as follows: Section 2 surveys validity indexes proposed in the literature and proposes a new robust validity index for SC algorithm. A modified SC algorithm for evaluating cluster centers is introduced in Section 3. Section 4 presents the experiment results, and Section 5 discusses the results.

2. Robust validity index for data clustering

2.1. Overview of validity indexes

The cluster validity index is widely used to evaluate partition fitness in clustering algorithms. To measure the qualities of the partitions provided by the output, a *validity index* assigns a value to the output of the clustering algorithm. The validity index for finding an optimal c, denoted c^*, that adequately describes the data structure is the most intensively studied topic in cluster validity. Several popular validity indexes are reviewed in Refs. [1,12,13].

An important quality of a clustering algorithm is the association between data points and cluster centers. The membership function is used to measure the strength of the association. If the data set contains c clusters, then each data point has c memberships which represent the close degree between the data point to the cluster center. For each data point, if one of the membership values of a particular data point belonged to a cluster is larger than the others, then this point is identified as being an element of that cluster [14]. Therefore, information about membership degrees can be represented by a single number indicating the fitness of the data point as classified by the clustering algorithms.

The literature suggests that the two key factors in validity indexes of data clustering algorithms are compactness and separation. *Compactness* refers to the cohesion *i.e.*, the concentration of data in each cluster. For a data clustering algorithm, high compactness indicates a good partition. Notably, compactness increases with cluster number and is highest when each datum forms a cluster. In contrast, *separation* is a measure of coupling between clusters. Low coupling indicates that the clusters have a weak relationship but that the partition is good. The validity index for measuring the goodness of partitions can be designed to consider both of these factors:

$$\textit{Validity index}\,(c) = \frac{\textit{Compactness}}{\textit{Separation}}, \tag{1}$$

or

$$\textit{Validity index}(c) = \textit{Compactness} - \textit{Separation}, \tag{2}$$

where $2 \le c \le c_{\max}$ is the number of clusters. A reasonable approach is to evaluate the maximal value of Eq. (1) or Eq. (2) in searching for the optimal cluster number c^* of a given data set. The most common validity indexes are reviewed below.

The partition coefficient (*PC*) index proposed by Bezdek [15] in 1974 was the first index for the fuzzy c-means algorithm (FCM) algorithm. It indicates the amount of *overlap* between clusters obtained by FCM. Let μ_{ik} be the membership value of data x_i belonging to cluster k. The *PC* index is defined as:

$$PC(c) = \frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n}(\mu_{ik})^2, \tag{3}$$

and

$$\sum_{k=1}^{c}\mu_{ik} = 1. \tag{4}$$

In a clustering algorithm for partitioning a data set, the cluster number is optimal when $PC(c)$ is maximal. The core concept of the *PC* index is that the cluster number is optimal when the cluster partition is least ambiguous. However, its disadvantage is that it only considers the fuzzy membership degree μ_{ik} for each cluster; *i.e.*, it does not consider structure.

Another index proposed by Bezdek is the partition entropy (*PE*) index [15,16], which is defined as

$$PE(c) = -\frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\log_a(\mu_{ik}), \tag{5}$$

where a is the base of the logarithm and the range of the $PE(c)$ is $[0,\log_a(c)]$. To find the optimal c^*, the author solves $\arg\min_{2\le c\le c_{\min}} PE(c)$ to produce the best clustering performance for a data set. Like the *PC* index, the disadvantage of the *PE* index is that it obtains the minimum value for each hard partition. Both indices only evaluate fuzziness and do not consider the data structure of the clusters.

In 1996, Dave defined a modified partition coefficient (*MPC*) [17] for the c-shell clustering algorithm. The *MPC* was defined as

$$MPC(c) = 1 - \frac{c}{c-1}(1 - PC(c)) \tag{6}$$

The *MPC*, which has a value of $[0,1]$, is a normalized version of the *PC* index. When $PC(c) = 1/c$, $MPC = 0$ and $PC(c) = 1$, $MPC = 1$. Therefore, it has the same disadvantage as the *PC* index.

The Xie and Beni (*XB*) index [18], which combines compactness and separation properties, is defined as

$$XB(c) = \frac{\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^2||x_i - v_k||^2}{n \times \min_{i\ne j}(||v_i - v_j||)}, \tag{7}$$

where v_i and v_j represent the centers of cluster i and j, respectively. In Eq. (7), the numerator indicates the compactness of the fuzzy partition while the denominator indicates the strength of the separation between clusters [19]. A small value in the numerator of the *XB* index represents a high compactness while a high value in the denominator denotes good cluster separation. Hence, the lowest $XB(c)$, $2 \le c \le c_{\max}$ indicates the optimal cluster number for a data set X.

The Fukuyama and Sugeno (*FS*) index [20] is another index which combines the properties of compactness and separation measurements. The *FS* index is defined as follows:

$$FS(c) = \sum_{i=1}^{n}\sum_{k=1}^{c}(\mu_{ik})^m||x_i - v_k||^2 - -\sum_{k=1}^{c}\left\{\left[\sum_{i=1}^{n}(\mu_{ik})^2\right]\cdot||v_k - \bar{v}||\right\}, \tag{8}$$

where $1 < m < \infty$, and $\bar{v} = 1/c\sum_{i=1}^{c} v_i$ represents the mean of the cluster centroids. In Eq. (8), the first term represents the geometrical compactness of the clusters, and the second term indicates the separation between the clusters. When m is 2, the first term of the *FS* index equals the numerator of the *XB* index. The second term of the *FS* index is used for separation measurements. Cluster number is optimized by solving $\arg\min_{2\le c\le n-1} FS(c)$.

The *I*-index proposed by Maulik and Bandyopadhyay [21] in 2002 is defined as

$$I(c) = \left(\frac{1}{c}\times\frac{E_1}{E_c}\times D_c\right)^2. \tag{9}$$

Here,

$$E_c = \sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}||x_i - v_j||,$$

$$D_c = \max_{i,j=1}^{c}||v_i - v_j||.$$