Expert Systems with Applications 37 (2010) 6567–6578

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Interval competitive agglomeration clustering algorithm

Jin-Tsong Jeng [b], Chen-Chia Chuang [a,*], C.W. Tao [a]

[a] Department of Electrical Engineering, National Ilan University, 1, Sec. 1, Shen-Lung Road, I-Lan 260, Taiwan
[b] Department of Computer Science and Information Engineering National Formosa University, No. 64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan

ARTICLE INFO

Keywords:
Symbolic interval-values data
Interval competitive agglomeration clustering algorithm
Fuzzy *c*-means clustering algorithm

ABSTRACT

In this study, an interval competitive agglomeration (ICA) clustering algorithm is proposed to overcome the problems of the unknown clusters number and the initialization of prototypes in the clustering algorithm for the symbolic interval-values data. In the proposed ICA clustering algorithm, both the Euclidean distance measure and the Hausdorff distance measure for the symbolic interval-values data are independently considered. Besides, the advantages of both hierarchical clustering algorithm and partitional clustering algorithm are also incorporated into the ICA clustering algorithm. Hence, the ICA clustering algorithm can be fast converges in a few iterations regardless of the initial number of clusters. Moreover, it is also converges to the same optimal partition regardless of its initialization. Experiments with simply data sets and real data sets show the merits and usefulness of the ICA clustering algorithm for the symbolic interval-values data.

1. Introduction

Clustering, also known as unsupervised classification, is a process by which a data set is divided into different clusters such that elements of the same cluster are as similar as possible and elements of different clusters are as dissimilar as possible. Most existing clustering algorithms can be classified into the following two categories: hierarchical clustering algorithm and partitional clustering algorithm (Gordon, 1999; Jain, Murty, & Flynn, 1999). The hierarchical clustering procedures provide a nested sequence of partitions with a graphical representation known as the dendrogram. The partitional clustering procedures generate a single partition (as opposed to a nested sequence) of the data in an attempt to recover the natural grouping present in the data. Prototype-based clustering algorithms are the most popular class of the partitional clustering algorithm. In the prototype-based clustering algorithms, each cluster is represented by a prototype, and the sum of distances from the feature vectors to the prototypes is usually used as the objective function.

In the clustering analysis, the patterns to be grouped are usually represented as a vector of the quantitative or the qualitative measurements where each column represents a variable. Each pattern takes a single value for each variable. However, this model is too restrictive to represent complex data. In order to take into the account variability and/or the uncertainty inherent to the data, variables must assume sets of categories or intervals, possibly even with frequencies or weights. These kinds of data have been mainly studied in symbolic data analysis (SDA). The aim of SDA is to provide the suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by the multi-valued variables, where the cells of the data table contain sets of categories, intervals, or weight (probability) distributions (Billard & Diday, 2003; Bock & Diday, 2000).

The SDA provides a number of clustering methods for the symbolic data. These methods differ in the type of the considered symbolic data, in their cluster structures and/or in the considered clustering criteria. With the hierarchical methods, an agglomerative approach has been introduced that forms composite symbolic objects using a join operator whenever mutual pairs of the symbolic objects are selected for agglomeration based on minimum dissimilarity (Gowda & Diday, 1991) or maximum similarity (Gowda & Diday, 1992). In Ichino and Yaguchi (1994), authors defined generalized Minkowski metrics for mixed feature variables and presents dendrograms obtained from the application of standard linkage methods for the data sets containing the numeric and the symbolic feature values. In Gowda and Ravi (1995a, 1995b), the divisive and agglomerative algorithms for the symbolic data based on the combined usage of similarity and dissimilarity measures are proposed. These proximity measures are defined on the basis of the position, span and content of symbolic data. In Chavent (1998), author proposes a divisive clustering method that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterization of each cluster in the hierarchy. In Gowda and Ravi (1999), a hierarchical clustering algorithm for the symbolic data based on the gravitational approach is also proposed. The

* Corresponding author.
E-mail addresses: tsong@nfu.edu.tw (J.-T. Jeng), ccchuang@niu.edu.tw (C.-C. Chuang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.129

agglomerative clustering algorithms based on the similarity (Guru, Kiranagi, & Nagabhushan, 2004) and dissimilarity functions (Guru & Kiranagi, 2005) are introduced, respectively.

A number of authors have addressed the problem of non-hierarchical (i.e. partitional) clustering algorithms for the symbolic data. In Diday and Brito (1989), a transfer algorithm is used to partition a set of symbolic objects into clusters that described by the weight distribution vectors. In Ralambondrainy (1995), the classical k-means clustering algorithm is extended in order to manage data characterized by the numerical and the categorical variables. In Gordon (2000), an iterative relocation algorithm is used to partition a set of symbolic objects into classes so as to minimize the sum of the description potentials of the classes. In Verde, De Carvalho, and Lechevallier (2001), a dynamic clustering algorithm for the symbolic data is proposed. In Bock (2002), authors has proposed several clustering algorithms for the symbolic data described by interval variables, based on a clustering criterion and has thereby generalized similar approaches in the classical data analysis. In Chavent and Lechevallier (2002), authors proposed a dynamic clustering algorithm for the interval data where the class representatives are defined by an optimality criterion based on a modified Hausdorff distance. In Souza and De Carvalho (2004), authors proposed partitioning clustering methods for the interval data based on the city-block distances, also considering the adaptive distances. In De Carvalho, Souza, Chavent, and Lechevallier (2006), an adequacy criterion based on the adaptive Hausdorff distance is introduced into the partitioning clustering algorithm for the symbolic interval-values data. Recently, the fuzzy c-mean clustering algorithm is extended to deal with the symbolic interval-values data (De Carvalho, 2007). Moreover, this algorithm is superior to the previous results.

Although the hierarchical clustering algorithm and partitional clustering algorithm successfully applied into the various SDA applications, some of drawbacks are existed (Frigui & Krishnapuram, 1997; Jain & Dubes, 1988). The main disadvantage of the hierarchical clustering procedures is that they consider only local neighbors when merging/splitting clusters and they cannot incorporate a priori knowledge about the global shape or size of clusters. In the partitional clustering algorithm, the reasonable initialization and the number of clusters are hardly determined. Hence, an interval competitive agglomeration (ICA) clustering algorithm is proposed to overcome the above problems. In the ICA clustering algorithm, the concepts of the competitive agglomeration (CA) clustering algorithm (Frigui & Krishnapuram, 1997) are extended to deal with the symbolic interval-values data. Moreover, the advantages of an ICA clustering algorithm are also liked to the CA clustering algorithm. The ICA clustering algorithm minimizes a fuzzy prototype-based objective function iteratively, so that it can be used to find clusters of various shapes. The objective function is designed so that it inherits the advantages of hierarchical clustering. Additionally, the ICA clustering algorithm starts by partitioning the data set into a large number of small clusters. As the algorithm progresses, adjacent clusters compete for the symbolic interval-values data and the clusters that lose the competition gradually become depleted and vanish. Thus, the final partition is taken to have the "optimal" number of clusters from the view of the objective function. Moreover, the final result is far less sensitive to initialization and local minima. In the proposed ICA clustering algorithm, two distance measure of the symbolic interval-values data are independently considered. One is the Euclidean distance measure that used in the interval fuzzy c-means (IFCM) clustering algorithm (De Carvalho, 2007). Another is the Hausdorff distance measure that also used in De Carvalho et al. (2006). The good properties of the CA clustering algorithm for the crisp-values data are also shown in the ICA clustering algorithm for the symbolic interval-values data. Experiment results show the merits and the usefulness of the ICA clustering algorithm.

The organization of the rest of the paper is as follows. In Section 2, an IFCM clustering algorithm is briefly introduced. In Section 3, an ICA clustering algorithm is proposed and discussed. The simulation results are shown in Section 4. Finally, the conclusions are summarized in Section 5.

2. Interval fuzzy *c*-means (IFCM) clustering algorithm (De Carvalho, 2007)

This algorithm is an extension of the standard fuzzy c-means clustering algorithm that furnishes a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable squared Euclidean distance between the vectors of symbolic interval-values data. An IFCM clustering algorithm is stated as follows.

Let $X = \{\vec{x}_k | k = 1, \ldots, n\}$ be a set of n vectors in an n-dimensional feature space with coordinate axis labels $(x^1, \ldots, x^j, \ldots, x^p)$. Each pattern k is represented as vector of intervals $\vec{x}_k = \left(x_k^1, \ldots, x_k^j, \ldots, x_k^p\right)$ where $x_k^j = \left[a_k^j, b_k^j\right]$ with $a_k^j \leqslant b_k^j$. Let $G = (\vec{g}_1, \ldots, \vec{g}_i, \ldots, \vec{g}_c)$ represent a c-tuple of prototypes each of which characterizes one of the c clusters. The prototype $\vec{g}_i$ can be also represented as a vector of intervals $\left(g_i^1, \ldots, g_i^j, \ldots, g_i^p\right)$ where $g_i^j = \left[\alpha_i^j, \beta_i^j\right]$ with $\alpha_k^j \leqslant \beta_k^j$. An IFCM clustering algorithm minimizes the following objective function:

$$W^1(G, U, X) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^2 \phi(\vec{x}_k, \vec{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^2 \sum_{j=1}^{p} \left[\left(a_k^j - \alpha_i^j\right)^2 + \left(b_k^j - \beta_i^j\right)^2\right], \quad (1)$$

subject to

$$\sum_{i=1}^{c} u_{ik} = 1, \quad \text{for } k = 1, \ldots, n. \quad (2)$$

In (1), ϕ is the square of Euclidean distance measuring the dissimilarity between the vectors of the symbolic interval-values data, u_{ik} is the membership degree of pattern k in ith cluster, $U = [u_{ik}]$ is a $c \times n$ matrix called a constrained fuzzy c-partition matrix.

To minimize the objective function in (1) with respect to U, the Lagrange multipliers method is applied and obtained as

$$J^1(G, U, X) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^2 \sum_{j=1}^{p} \left[\left(a_k^j - \alpha_i^j\right)^2 + \left(b_k^j - \beta_i^j\right)^2\right] - \sum_{k=1}^{n} \lambda_k \left(\sum_{i=1}^{c} u_{ik} - 1\right). \quad (3)$$

Then, G is fixed and solve

$$\frac{\partial J^1}{\partial u_{st}} = 2u_{st} \sum_{j=1}^{p} \left[\left(a_t^j - \alpha_s^j\right)^2 + \left(b_t^j - \beta_s^j\right)^2\right] - \lambda_t = 0, \quad \text{for } s = 1, \ldots, c \quad \text{and } t = 1, \ldots, n \quad (4)$$

to obtain an updating equation for the memberships u_{st}. Thus, Eq. (4) can be rewritten as

$$u_{st} = \frac{\lambda_t}{2\sum_{j=1}^{p}\left[\left(a_t^j - \alpha_s^j\right)^2 + \left(b_t^j - \beta_s^j\right)^2\right]}, \quad \text{for } s = 1, \ldots, c \quad \text{and } t = 1, \ldots, n.(5)$$

Substituting (5) into (2), λ_t is obtained as