Expert Systems with Applications 37 (2010) 6942–6947

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A fuzzy *c*-means clustering algorithm based on nearest-neighbor intervals for incomplete data

Dan Li *, Hong Gu, Liyong Zhang

School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

Keywords:
Clustering
Fuzzy *c*-means
Incomplete data
Nearest-neighbor intervals

ABSTRACT

Partially missing data sets are a prevailing problem in clustering analysis. In this paper, missing attributes are represented as intervals, and a novel fuzzy *c*-means algorithm for incomplete data based on nearest-neighbor intervals is proposed. The algorithm estimates the nearest-neighbor interval representation of missing attributes by using the attribute distribution information of the data sets sufficiently, which can enhances the robustness of missing attribute imputation compared with other numerical imputation methods. Also, the convex hyper-polyhedrons formed by interval prototypes can present the uncertainty of missing attributes, and simultaneously reflect the shape of the clusters to some degree, which is helpful in enhancing the robustness of clustering analysis. Comparisons and analysis of the experimental results for several UCI data sets demonstrate the capability of the proposed algorithm.

1. Introduction

The fuzzy *c*-means (FCM) algorithm (Bezdek, 1981) is a useful tool for clustering, which partitions a real *s*-dimensional dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^s$ into several clusters to describe an underlying structure within the data, and has been extensively used in pattern recognition and data mining. However, in pattern classification applications, many datasets suffer from incompleteness, i.e. a dataset X can contain vectors that are missing one or more of the attribute values, as a result of failure in data collection, measurement errors, missing observations, random noise, etc. and FCM is not directly applicable to such incomplete datasets.

The problem of doing pattern recognition with incomplete data can be traced back to the 1960s, when Sebestyen (1962) introduced an approach based on probabilistic assumptions. Subsequently the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) was used to handle incomplete data and probabilistic clustering (McLachlan & Basford, 1988). In 1998, several methods were proposed for handling missing values in FCM (Miyamoto, Takata, & Umayahara, 1998). One basic strategy, imputation, replaces the missing values by the weighted averages of the corresponding attributes. Another approach, discarding/ignoring, ignores the missing values and calculates the distances from the remaining coordinates. In 2001, Hathaway and Bezdek proposed other strategies to continue the FCM clustering of incomplete data (Hathaway & Bezdek, 2001). One simple strategy (whole data strategy, WDS) removes all sample data that include missing values from the dataset, but the strategy is not desirable because the elimination brings a loss of information. Another method uses the partial distance strategy (PDS), which calculates partial distances using all available attribute values, and scales this quantity by the reciprocal of the proportion of components used. Two further methods proposed by Hathaway and Bezdek (2001) belong to the imputation method, which involve computations to replace the missing values with estimation based on the available information. The optimal completion strategy (OCS) views the missing values as an optimization problem and imputes missing values in each iteration to find better estimates. The nearest prototype strategy (NPS) replaces missing values with the corresponding attributes of the nearest prototype. Besides the above methods, by taking into account the information why data are missing, Timm, Doring, and Kruse (2004) developed a fuzzy clustering algorithm extended from the Gath and Geva algorithm. Hathaway and Bezdek (2002) used triangle inequality-based approximation schemes to cluster incomplete relational data, and Honda and Ichihashi (2004) partitioned the incomplete datasets into several linear fuzzy clusters by extracting local principal components.

In this paper, by adopting the idea of nearest-neighbor rule, a novel fuzzy *c*-means algorithm for incomplete data based on nearest-neighbor intervals (FCM-NNI) is proposed. Firstly, because of the uncertainty of missing attributes, missing attributes are represented by nearest-neighbor intervals (NNI) based on the nearest-neighbor information, which are more robust than the numerical values obtained by imputation methods mentioned above. Secondly, the clustering problem can be thus viewed as clustering

* Corresponding author. Tel.: +86 411 82965258.
E-mail addresses: ldan@dlut.edu.cn (D. Li), guhong@dlut.edu.cn (H. Gu), zhlyad@163.com (L. Zhang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.03.028

for interval-valued data, which will result in interval cluster prototypes rather than point prototypes. Therefore, the convex hyper-polyhedrons formed by interval prototypes in the attribute space, as a kind of cluster prototype with more complicated geometrical structure, can present the uncertainty of missing attributes, and at the same time reflect the shape of the clusters to some degree, thus validating the robustness of clustering pattern with more accurate clustering results.

This paper is organized as follows. Section 2 presents a short description of the FCM algorithm and FCM clustering algorithm for interval-valued data (IFCM) based on clustering objective function minimization. The nearest-neighbor interval representation of missing attributes and the novel FCM-NNI algorithm are introduced in Section 3. Section 4 presents clustering results of several UCI data sets and a comparative study of our proposed algorithm with various other methods for handling missing values in FCM. Finally, conclusions are drawn in Section 5.

2. FCM clustering algorithm for interval-valued data

2.1. Fuzzy c-means algorithm

The fuzzy c-means (FCM) algorithm partitions a set of complete data $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^s$ into c-(fuzzy) clusters by minimizing the clustering objective function

$$J(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2, \tag{1}$$

with the constraint of

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

where $\boldsymbol{x}_k = [x_{1k}, x_{2k}, \ldots, x_{sk}]^T$ is an object datum, and x_{jk} is the jth attribute value of $\boldsymbol{x}_k$; $\boldsymbol{v}_i$ is the ith point cluster prototype, $\boldsymbol{v}_i \in \mathbb{R}^s$, and let the matrix of cluster prototypes $\boldsymbol{V} = [v_{ji}] = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c] \in \mathbb{R}^{s \times c}$ for convenience; u_{ik} is the membership that represents the degree to which $\boldsymbol{x}_k$ belongs to the ith cluster, $\forall i, k$: $u_{ik} \in [0,1]$, and let the partition matrix $\boldsymbol{U} = [u_{ik}] \in \mathbb{R}^{c \times n}$ for convenience; m is a fuzzification parameter, $m \in (1, \infty)$; and $\|\cdot\|_2$ denotes Euclidean norm.

FCM uses the Lagrange multiplier method, and let the Lagrange function be

$$J_a(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2 + \sum_{k=1}^{n} \lambda_k \left(\sum_{i=1}^{c} u_{ik} - 1 \right), \tag{3}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_n]^T$ is Lagrange multiplier, and the necessary conditions for minimizing (1) with the constraint of (2) are the update equations as follows (Bezdek, 1981):

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}^m}, \quad \text{for } i = 1, 2, \ldots, c \tag{4}$$

and

$$u_{ik} = \left[\sum_{t=1}^{c} \left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|_2^2}{\|\boldsymbol{x}_k - \boldsymbol{v}_t\|_2^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \text{for } i = 1, 2, \ldots, c \quad \text{and} \quad k = 1, 2, \ldots, n. \tag{5}$$

The procedure of FCM can be described as follows.

Step 1: Choose m, c and ε, where $\varepsilon > 0$ is a small positive constant; then initialize the partition matrix $\boldsymbol{U}^{(0)}$.

Step 2: When the iteration index is l ($l = 1, 2, \ldots$), calculate the matrix of cluster prototypes $\boldsymbol{V}^{(l)}$ using (4) and $\boldsymbol{U}^{(l-1)}$.

Step 3: Update the partition matrix $\boldsymbol{U}^{(l)}$ using (5) and $\boldsymbol{V}^{(l)}$.

Step 4: If $\forall i,\ k : \max |u_{ik}^{(l)} - u_{ik}^{(l-1)}| < \varepsilon$, then stop and get the partition matrix $\boldsymbol{U}$ and the matrix of cluster prototypes $\boldsymbol{V}$; otherwise set $l = l + 1$ and return to Step 2.

2.2. FCM clustering algorithm for interval-valued data (IFCM)

Let $\overline{X} = \{\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_n\}$ be an s-dimensional interval-valued data set to be partitioned into c-(fuzzy) clusters, where $\bar{\boldsymbol{x}}_k = [\bar{x}_{1k}, \bar{x}_{2k}, \ldots, \bar{x}_{sk}]^T$, $\forall j,\ k : \bar{x}_{jk} = [x_{jk}^-, x_{jk}^+]$. The IFCM algorithm minimizes the objective function

$$J_I(\boldsymbol{U}, \overline{\boldsymbol{V}}) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{v}}_i\|_2^2, \tag{6}$$

with the constraint of (2), where $\bar{\boldsymbol{v}}_i$ is the ith interval cluster prototype, and let the matrix of interval cluster prototypes $\overline{\boldsymbol{V}} = [\bar{v}_{ji}] = [\bar{\boldsymbol{v}}_1, \bar{\boldsymbol{v}}_2, \ldots \bar{\boldsymbol{v}}_c]$, where $\bar{v}_{ji} = [v_{ji}^-, v_{ji}^+], \forall i = 1, 2, \ldots c,\ j = 1, 2, \ldots, s$. The Euclidean distance between $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{v}}_i$ is defined as

$$\|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{v}}_i\|_2 = [(\boldsymbol{x}_k^- - \boldsymbol{v}_i^-)^T (\boldsymbol{x}_k^- - \boldsymbol{v}_i^-) + (\boldsymbol{x}_k^+ - \boldsymbol{v}_i^+)^T (\boldsymbol{x}_k^+ - \boldsymbol{v}_i^+)]^{\frac{1}{2}}, \tag{7}$$

where

$$\boldsymbol{x}_k^- = [x_{1k}^-, x_{2k}^-, \ldots, x_{sk}^-]^T, \quad \boldsymbol{x}_k^+ = [x_{1k}^+, x_{2k}^+, \ldots, x_{sk}^+]^T,$$
$$\boldsymbol{v}_i^- = [v_{1i}^-, v_{2i}^-, \ldots, v_{si}^-]^T, \quad \boldsymbol{v}_i^+ = [v_{1i}^+, v_{2i}^+, \ldots, v_{si}^+]^T.$$

The necessary conditions for minimizing (6) with the constraint of (2) are the update equations as follows (Yu & Fan, 2004):

$$\boldsymbol{v}_i^- = \frac{\sum_{k=1}^{n} u_{ik}^m \boldsymbol{x}_k^-}{\sum_{k=1}^{n} u_{ik}^m}, \quad \text{for } i = 1, 2, \ldots, c. \tag{8}$$

and

$$\boldsymbol{v}_i^+ = \frac{\sum_{k=1}^{n} u_{ik}^m \boldsymbol{x}_k^+}{\sum_{k=1}^{n} u_{ik}^m}, \quad \text{for } i = 1, 2, \ldots, c, \tag{9}$$

And if $\exists k, h, 1 \leqslant k \leqslant n, 1 \leqslant h \leqslant c, \forall j : \bar{x}_{jk} \subseteq \bar{v}_{jh}$, that is, $\bar{\boldsymbol{x}}_k$ is within the convex hyper-polyhedron formed by $\bar{\boldsymbol{v}}_h$, then $\bar{\boldsymbol{x}}_k$ can be considered to belong fully to the hth cluster with membership 1, and belong to the other clusters with membership 0. Thus

$$u_{ik} = \begin{cases} 1, & i = h \\ 0, & i \neq h \end{cases} \quad \text{for } i = 1, 2, \ldots, c, \tag{10}$$

else

$$u_{ik} = \left[\sum_{t=1}^{c} \left(\frac{\|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{v}}_i\|_2^2}{\|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{v}}_t\|_2^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \text{for } i = 1, 2, \ldots, c. \tag{11}$$

The procedure of IFCM is similar to that of FCM, and hence is omitted here.

3. Fuzzy c-means algorithm for incomplete data based on nearest-neighbor intervals

3.1. Nearest-neighbor intervals determination

Recently, the use of nearest-neighbor (NN) based techniques has been proposed for imputation of missing values. A simple NN imputation method is to substitute the missing attribute by the corresponding attribute of the nearest-neighbor (Stade, 1996). And in another popular approach, k-nearest-neighbor imputation (Acuna & Rodriguez, 2004), missing attributes are supplemented by the mean value of the attribute in the k-nearest-neighbors. Subsequently, many similarity measures other than Euclidean distance are introduced in searching for nearest-neighbors (Huang & Zhu, 2002; Huang, 2006). All the approaches mentioned above develop