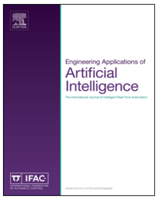
Engineering Applications of Artificial Intelligence 53 (2016) 41–52

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



A modified Fuzzy *k*-Partition based on indiscernibility relation for categorical data clustering

Iwan Tri Riyadi Yanto[a], Maizatul Akmar Ismail[b], Tutut Herawan[b]

[a] *Department of Information System, University of Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Department of Information Systems, University of Malaya, 50603 Pantai Valley, Kuala Lumpur, Malaysia*

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form
29 November 2015
Accepted 12 January 2016
Available online 11 April 2016

Keywords:
Clustering
Categorical data
Fuzzy *k*-Partition
Indescernibility relation

ABSTRACT

Categorical data clustering has been adopted by many scientific communities to classify objects from large databases. In order to classify the objects, Fuzzy *k*-Partition approach has been proposed for categorical data clustering. However, existing Fuzzy *k*-Partition approaches suffer from high computational time and low clustering accuracy. Moreover, the parameter maximize of the classification likelihood function in Fuzzy *k*-Partition approach will always have the same categories, hence producing the same results. To overcome these issues, we propose a modified Fuzzy *k*-Partition based on indiscernibility relation. The indiscernibility relation induces an approximation space which is constructed by equivalence classes of indiscernible objects, thus it can be applied to classify categorical data. The novelty of the proposed approach is that unlike previous approach that use the likelihood function of multivariate multinomial distributions, the proposed approach is based on indescernibility relation. We performed an extensive theoretical analysis of the proposed approach to show its effectiveness in achieving lower computational complexity. Further, we compared the proposed approach with Fuzzy Centroid and Fuzzy *k*-Partition approaches in terms of response time and clustering accuracy on several UCI benchmark and real world datasets. The results show that the proposed approach achieves lower response time and higher clustering accuracy as compared to other Fuzzy *k*-based approaches.

1. Introduction

Clustering is a fundamental problem that frequently arises in a broad variety of fields such as pattern recognition, image processing, machine learning and statistics (Haixia and Zheng, 2009; Jain et al., 1999). It can be defined as a process of partitioning a given data set of multiple attributes into groups. The *k*-means algorithm (MacQueen, 1967) is the most popular among clustering algorithms developed to date because of its effectiveness and efficiency in clustering large data sets. However, *k*-means clustering algorithm fails to handle data sets with categorical attributes because it can only minimize a numerical cost function. As a result, Huang (Huang, 1998) proposed the *k*-modes clustering method that removes the numeric-only limitation of the *k*-means algorithm. Since then major improvements have been made in *k*-modes algorithms including new dissimilarity measures to the *k*-modes clustering (He et al., 2005; Ng et al., 2007; San et al., 2004) and a fuzzy set based *k*-modes algorithm (Huang, 1999; Wei et al., 2009). To improve the efficiency of fuzzy *k*-modes, Kim et al. (2004) [10] proposed a technique using Fuzzy Centroid (FC) approach. On the base of a different construction on categorical data, Umayahara and Miyamoto (2005) proposed another fuzzy approach for clustering documents data.

The Fuzzy *c*-mean (FCM) clustering algorithm (Kim et al., 2004) and its variants for clustering numerical (Khalilia et al., 2014; Leski, 2004), symbolic (De Carvalho, 2007; Dobosz and Duch, 2010) and categorical data (Huang, 1999, 1998; Parmar et al., 2007; Yang et al., 2008) are non-parametric approaches which are based on the least sum of squared errors within-clusters. Yang et al. (2008) proposed Fuzzy *k*-Partititon (F*k*P) algorithm which is a parametric approach based on the likelihood function of multivariate multinomial distributions. The F*k*P can also be referred to a Fuzzy-based clustering algorithm for categorical data. However, almost all fuzzy categorical data clustering algorithms mentioned above represent data set in the binary values. Moreover, in F*k*P algorithm we observed that the maximized parameter of the classification likelihood function in the same categories always have the same results. Another issue with the aforesaid approaches is that they tend to have high computational time and low clusters purity. This indicates that an approach that does not suffer from high computational time and low clusters purity is needed.

E-mail addresses: yanto.itr@is.uad.ac.id (I.T.R. Yanto), maizatul@um.edu.my (M.A. Ismail), tutut@um.edu.my (T. Herawan).

http://dx.doi.org/10.1016/j.engappai.2016.01.026
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

In this paper, we propose a modified Fuzzy *k*-Partition based on indiscernibility relation for categorical data clustering. The indiscernibility relation induces an approximation space which is constructed by equivalence classes of indiscernible objects. The indiscernibility relation is intended to express fact that due to the lack of knowledge we are unable to discern some objects by just employing the available information. The indiscernibility relation induces an approximation space made of equivalence classes of indiscernible objects. Thus, the indiscernibility relation can be applied to the categorical data without representing data in the binary values. In summary, this paper makes the following contributions:

- A modified Fuzzy *k*-Partition approach based on indiscernibility relation for categorical data clustering is proposed.
- A correctness of proof and related algorithm of proposed approach are presented.
- Theoretical comparative analysis in term of computational complexity between the proposed approach with others Fuzzy *k*-based approaches is presented.
- Comparison from experiment results on bechmark and real world data sets between the proposed approach with others Fuzzy *k*-based approaches in terms of response time and clustering purity are presented.

The rest of the paper is organized as follows. Section 2 describes related works on Fuzzy-based categorical data clustering. Section 3 describes the proposed approach based on the indiscernibility and fuzzy set concept, followed by its correctness, proposed algorithm and its computational complexity. Section 4 describes the experiment results on benchmark and real world datasets. Finally, we conclude our work in Section 5.

2. Fuzzy-based categorical data clustering

Recently, fuzzy-based clustering has been widely focused by many scholars and some significant results have been achieved in the theoretical and practical aspects. In this section, we review related works of two Fuzzy-based categorical data clustering approaches i.e. Fuzzy Centroid and Fuzzy *k*- Partition.

2.1. Fuzzy Centroid

The Fuzzy *k*-modes proposed by Huang (1998) is the most used algorithm for numerical data and there are several extensions of FCM (Yang et al., 2008). For clustering data, hard and fuzzy *k*-modes algorithms using simple matching dissimilarity measure (Huang, 1999). Let $Y = y_1, y_2, \ldots, y_I$ be a set of categorical data and let each data be defined by a set of attributes $A_1, \ldots, A_J$ with $y_i = \left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right)$, for $i = 1, 2, \ldots, I$. Each attribute A_j describes a domain of values denoted by $DOM(A_j) = \left\{a_j^1, \ldots, a_j^{L_j}\right\}$, where L_j is the number of categories of the attribute A_j, for $j = 1, 2, \ldots, J$. Suppose that $v_k = (v_{k1}, v_{k2}, \ldots, v_{kJ})$ is the centroid of the *k*-th cluster where each v_{kj} is coded by $(v_{kj1}, v_{kj2}, \ldots, v_{kjL_j})$ for $k = 1, 2, \ldots, K$, and $j = 1, 2, \ldots, J$ with $v_{kjl} = 1$ and $v_{kjl'} = 0$ for $l' \neq l, 1 \leq j \leq J, 1 \leq l', l \leq L_j$ if $v_{kj} = a_j^l$. The matching dissimilarity measure by Huang (1998, 1999) is defined as follows

$$d(y_i, v_k) = \sum\nolimits_{j=1}^{J} \delta(y_{ij}, v_{kj}), \tag{1}$$

where

$$\delta\left(y_{ij}, v_{kj}\right) = \begin{cases} 0 & if \quad y_{ij} = v_{kj} \\ 1 & if \quad y_{ij} \neq v_{kj} \end{cases}$$

The minimize objective function of fuzzy *k*-modes (Huang, 1998) is as follows

$$H_m(\mu, v) = \sum\nolimits_{i=1}^{I} \sum\nolimits_{k=1}^{K} \mu_{ik}^m d(y_i, v_k), \tag{2}$$

subject to

$$\sum\nolimits_{k=1}^{K} \mu_{ik} = 1, \quad \text{for} \quad i = 1, 2, \ldots, I,$$

where m is the fuzziness index. The update equations for hard *k*-modes are as follow:

$$\mu_{ik} = \begin{cases} 1 & if \quad d(y_i, v_k) = \min\limits_{1 \leq k' \leq K} d(y_i, v_k') \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum\nolimits_{i=1}^{I} \mu_{ik} y_{ijl} = \max\limits_{1 \leq l' \leq L} \sum\nolimits_{i=1}^{I} \mu_{ik} y_{ijl}{}' \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Huang (1999) extended the hard *k*-mode to Fuzzy *k*-modes. Using the objective (2), the update equation for fuzzy *k*-modes using objective function (2) is as follows

$$\mu_{ik} = \frac{1}{\sum\nolimits_{k'=1}^{K} \left[\frac{d(y_i, v_k)}{d(y_i, v_{k'})}\right]^{\frac{1}{m-1}}} \tag{5}$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum\nolimits_{i=1}^{I} \mu_{ik}^m y_{ijl} = \max\limits_{1 \leq l' \leq L} \sum\nolimits_{i=1}^{I} \mu_{ik}^m y_{ijl}{}' \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The use of hard centroids can give rise to the artifacts. For example, although the Fuzzy *k*-modes algorithm efficiently handles categorical data sets, it uses a hard centroid representation for categorical data in a cluster. The use of hard rejection of data can lead to misclassification in the region of doubt (Yang et al., 2008).

Kim et al. (2004) improved the performance of fuzzy *k*-modes by changing hard centroids to Fuzzy Centroid with $\tilde{v}_{kj} = \left(\tilde{v}_{kj1}, \ldots, \tilde{v}_{kjL_j}\right)$, for $k = 1, 2, \ldots K$ and $j = 1, 2, \ldots, J$, where $\tilde{v}_{kjl} \in [0, 1]$ and $\sum\nolimits_{i=1}^{L_j} \tilde{v}_{kjl} = 1$. The minimize objective function of Fuzzy Centroid is as follows

$$H_m(\mu, v) = \sum\nolimits_{i=1}^{I} \sum\nolimits_{k=1}^{K} \mu_{ik}^m d(y_i, \tilde{v}_k), \tag{7}$$

subject to

$$\sum_{k=1}^{K} \mu_{ik} = 1, i = 1, 2, \ldots, I.$$

$$\sum_{l=1}^{L_j} \tilde{v}_{kjl} = 1.$$

The distance measure with the centroid updates equations which are given as following equation:

$$d(y_i, \tilde{v}_k) = \sum_{j=1}^{J} \delta(y_{ij}, \tilde{v}_{kj}) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} \left(1 - y_{ijl}\right) \tilde{v}_{kjl},$$

$$\tilde{v}_{kjl} = \frac{\sum\nolimits_{i=1}^{I} \mu_{ik}^m \bullet y_{ijl}}{\sum\nolimits_{i=1}^{I} \mu_{ik}^m}. \tag{8}$$

The update equation of memberships can be obtained as follows

$$\mu_{ik} = \frac{1}{\sum\nolimits_{k'=1}^{K} \left[\frac{d(y_i, \tilde{v}_k)}{d(y_i, \tilde{v}_{k'})}\right]^{\frac{1}{m-1}}}. \tag{9}$$

Both of the Fuzzy *k*-modes with hard centorid and Fuzzy Centroid approach are non-parametric approaches. The algorithms