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/

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 $l'\neq l, 1\le j\le J, 1\le l'\le L_j$ if $v_{kj}=a_j^l$. The matching dissimilarity measure by Huang (1998,·1999) is defined as follows

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

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

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

subject to

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

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

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

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

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

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

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

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 $\sum_{i=1}^{L_j}\tilde{v}_{kjl}=1$. The minimize objective function of Fuzzy Centroid is as follows

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

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:

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

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

The update equation of memberships can be obtained as follows

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

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

use the dissimilarity functional based on the least total within cluster matching dissimilarity. This selection implies, in essence, the assumption of data organized into spherical clusters (Bryant and Williamson, 1978; Chatzis, 2011).

2.2. Fuzzy k-Partition

The Fuzzy k-Partition model proposed by Yang et al. (2008) is another alternative approach for categorical data clustering. It is based on the likelihood function of multivariate multinomial distribution. The approach is operated on a data set Y composed of I observations of J discrete attribute with only one of a finite number (say L_j) of value categories for the attribute j. The model uses the indicator function $z_1, z_2, \ldots, z_k$ for each partition $P = P_1, P_2, \ldots, P_K$ of Y into K classes as mutually disjoint sets $P_1, P_2, \ldots, P_K$ where $P_1 \cup P_2 \cup \ldots \cup P_K = Y$ such that $z_k(y) = 1$ if $y \in P_k$ and otherwise, $z_k(y) = 0$ for all y in Y,$k = 1, 2, .., K$. This is known as clustering data into K classes using z and termed a hard k-partition of Y. The review of the model is given as follows:

For each attribute j in individual i, let the values be represented by Y_{ijl} with a set of L_j binary random attributes where y_{ijl} is a realization of Y_{ijl} with

$$Y_{ijl} = y_{ijl}, \text{ for } i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J, \text{ and } l = 1, 2, \ldots, L_j$$

Thus, y_{ijl} has a binary value, that is, y_{ijl} has value 0 or 1. Consider Y_i, for $i = 1, 2, \ldots, I$ to be a random sample of size I from a multivariate multinomial distribution $f(y, \lambda)$. Let $P = P_1, P_2, \ldots, P_K$ be a partition of Y. A classification joint distribution function $Y_1, Y_2, \ldots, Y_I$ based on the partition can be written as $\Pi_{k=1}^{K} \Pi_{y_i \in P_k} f_k(y_i, \lambda_k)$ which is co-called Classification Maximum Likelihood (CML) approach (Bryant and Williamson, 1978; Scott and Symons, 1971; Symons, 1981). Consider the extension the indicator function $z_{ik} = z_k(y_i)$ to be function $\mu_{ik} = \mu_k(y_i)$ assuming in the interval [0,1] such that $\sum_{k=1}^{K} \mu_{ik} = 1$ for $i = 1, 2, \ldots, I$. In [12], μ is called a Fuzzy k-Partition of the data set Y that had been used for fuzzy clustering (Bezdek, 2013; Wu and Yang, 2002; Yang, 1993). By increasing the power of the fuzziness power of μ_{ik} to μ_{ik}^m, the extension of maximizing the log likelihood CML procedure as described in (Leski, 2004) can be written as follows:

$$\text{Maximize } J_m(\mu, \lambda) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m \ln f_k(y_i, \lambda_k) \tag{10}$$

$$\text{Subject to } \sum_{k=1}^{K} \mu_{ik} = 1, \text{for} i = 1, 2, \ldots, I \text{with} \mu_{ik} \in [0, 1],$$

where $m > 1$ is a fixed constant as an index of fuzziness. The optimization for $J_m(\mu, \lambda)$ is by choosing a Fuzzy k-Partition and an estimate λ to maximize $J_m(\mu, \lambda)$.

Consider $f_k(y, \lambda_k)$ as a multivariate multinomial distribution with

$$f_k(y; \lambda_k) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{jl}} where \sum_{l=1}^{l_j} \lambda_{kjl} = 1 \forall k, j, \tag{11}$$

where λ_{kjl} is a probability of value l for the jth attribute by individual i with the kth extreme profile, i.e $P(Y_{ijl} = 1 | Y_i$ in k class$) = \lambda_{kjl}$. By replacing $f_k(y, \lambda_k)$ with the above multivariate multinomial distribution, the model can be written as

$$\begin{aligned} J_m(\mu, \lambda) &= \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m ln \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{ijl}} \\ &= \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m \sum_{j=1}^{J} \sum_{l=1}^{l_j} ln(\lambda_{kjl})^{y_{ijl}} \end{aligned} \tag{12}$$

The maximization of the Fuzzy k-Partition objective function $J_m(\mu, \lambda)$ can be obtained by updating the equation as follows:

$$\lambda_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^m \bullet y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^m} \tag{13}$$

$$\mu_{ik} = \left[\sum_{s=1}^{K} \left(\frac{\sum_{j=1}^{J} \sum_{l=1}^{L_j} ln(\lambda_{kjl})^{y_{ij}}}{\sum_{j=1}^{J} \sum_{l=1}^{L_j} ln(\lambda_{sjl})^{y_{ij}}} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{14}$$

Fuzzy k-Partition is a parametric approach based on the likelihood function of multivariate multinomial distribution. It can improve the accuracy of the clusters. However, since the Fuzzy k-Partition has more complicated computation, it may spend more running time than that the Fuzzy Centroid approach. Meanwhile, in Fuzzy Centroid, the categorical data must be represented as binary random attributes. Thus, it tends to have high computational time.

In the following section, we present an alternative Fuzzy-based categorical data clustering which is based on indiscernibility relation. We will illustrate how that our proposed Fuzzy indescernibility approach will produce better results in terms of lower response time and higher cluster purity as compared to Fuzzy Centroid and Fuzzy k-Partition approaches.

3. Proposed modified Fuzzy *k*-Partition approach

In this section, we introduce modified Fuzzy k-Partition approach. Its algorithm is presented along with necessary preliminary information.

3.1. Indiscernibility relation

In this section, we reviewed some definitions with regard to indiscernibility relation. The concept of an indiscernibility relation comes from the fact that two instances in an information system can have similar attribute-value. In rough set theory (Pawlak, 1982), data are often presented as a finite table (later we called an *information system*), where columns of which are labeled by *attributes*, rows by *objects* of interest and entries of the table are *attribute values* (Pawlak, 1992). Formally, an *information system* is defined as a 4-tuple (quadruple) $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \cup_{a \in A} V_a$, V_a is the domain (value set) of attribute a, $f : U \times A \rightarrow V$ is a total function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, which is called information (knowledge) function.

The indiscernibility relation of objects in information system is intended to express the fact that due to the lack of knowledge, we are unable to discern some objects employing the available information. Therefore, we are unable to deal with just a single object. Nevertheless, we have to consider classes of indiscernible (similar) objects. The following definition precisely describes the notion of indiscernibility relation between two objects (Pawlak and Skowron, 2007).

Definition 1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ are said to be B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$.*

From Definition 1, it is clear that every subset of A induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute B, denoted by $IND(B)$, is an equivalence relation which is reflexive, symmetric, and transitive. It is well known that, an equivalence relation induces unique partition. The partition of U induced by $IND(B)$ in an information system $S = (U, A, V, f)$ denoted by U/B and the equivalence class in the partition U/B containing $x \in U$,denoted by $[x]_B$. Based on the notion of indiscernibility relation above, in the following subsection we present the proposed modified Fuzzy k-Partition based on indiscernibility relation.

3.2. Fuzzy k-Partition based on indiscernibility relation

In this section, we present the proposed approach called fuzzy indescernibility based (FID), which we refer to as indiscernible set in every attribute to represent the categorical data Y. The main proposed approach is replacing the binary data y_{ijl} as a realization of Y_{ijl} with the equivalence classes in the indiscernible relation of the original categorical data Y. In this sub-section, we introduce several rudimentary used in the proposed approach.

Definition 2. Consider the categorical data Y which can be represented by the information system $S=(U,A,V,f)$, the row of data Y represents a finite set of object $U=[y_1,y_2,\ldots,y_I]$ and the column as a finite set of attribute $a_j=[a_1,a_2,\ldots,a_J]$. Suppose that $a_j\in A, V(a_j)$ have l-different values, say γ_{jl} for $l=1,2,\ldots,L_j$. Let $X\left(a_j=\gamma_{jl}\right)$ be a subset of the objects having have l-different values of attribute a_j. The data Y_{ijl} can be represent by

$$y_{ijl}=\begin{cases}1 & y_{ij}=\gamma_{jl}\\ 0 & y_{ij}\neq\gamma_{jl}\end{cases},\text{for } j=1,2,\ldots,L_j$$

Obviously, for $i=1,2,..,I$, the above equation is equivalent to $y_{ijl}.y_{i\in X(a_j=\gamma_{jl})}=1,\text{for} j=1,2,\ldots,L_j$

and the logarithm value of λ_{kjl} in equation (12) can be represented as follows

$$\ln(\lambda_{kjl})^{y_{ijl}}=\begin{cases}0 & i\notin X\left(a_j=\gamma_{jl}\right)\\ \ln(\lambda_{ijk}) & i\in X\left(a_j=\gamma_{jl}\right)\end{cases},\text{for } j=1,2,\ldots,L_j$$

or

$$\ln(\lambda_{kjl})^{y_{ijl}.y_{i\in X(a_j=\gamma_{jl})}}=\ln\left(\lambda_{(i\in X(a_j=\gamma_{jl}))jk}\right),\text{for}=1,2,\ldots L_j.$$

We can see that the probability of value λ_{kjl} which will give contribution in the objective function in Eq. (12) is only object in each the equivalence class in the partition U/a_j , $a_j\subset A$ containing $x\in U$, that is $[x]_{a_j}$. Otherwise the result is 0. In other word, it can be said that the data which have the same category or in the same equivalence class have the same probability of value λ_{kjl}.

Definition 3. Let the information system $S=(U,A,V,f)$. Suppose $a_j\in A, V(a_j)$ have l-different values, say γ_{jl}, for $l=1,2,\ldots,l_j$. Let $X\left(a_j=\gamma_{jl}\right)$, be a subset of the objects having have l-different values of attribute a_j. The maximize parameter λ_{kjl} of the initial cluster k, for object i belong to the set $X\left(a_j=\gamma_{jl}\right)$, denoted$\lambda^k_{\frac{U}{a_j}}$, is defined by $\lambda^k_{\frac{U}{a_j}}=\left\{\lambda_{ijk}\middle| i\in X\left(a_j=\gamma_{jl}\right)\right\},forl=1,2,\ldots,L_j$

The maximization of the objective function $J_m(\mu,\lambda)$ can be rewritten as follows:

$$J_m(\mu,\lambda)=\sum_{i=1}^{I}\sum_{k=1}^{K}\mu_{ik}^m\sum_{j=1}^{J}\ln\left(\lambda^k_{\frac{U}{a_j}}\right) \tag{15}$$

by the contrains

$$\mu_{ik}\geq 0,\sum_{k=1}^{K}\mu_{ik}=1, \tag{16}$$

$$\lambda^k_{\frac{U}{a_j}}\geq 0,\sum_{j=1}^{J}\lambda^k_{\frac{U}{a_j}}=1. \tag{17}$$

The maximition of the objective function in Eq. (15) is based on the indiscernibility relation in Definition 1. Only the equivalence class that has contribution is used to maximize the function. The optimum can be obtained by setting the first derivatives of the Lagrangian J_m with respect to the all parameter to be 0.

Proposition 1. Let an information system $S=(U,A,V,f)$. Suppose $a_j\in A$, for $V(a_j)$ have l-different values, say γ_{jl}, for $l=1,2,\ldots,l_j$. If $X\left(a_j=\gamma_{jl}\right)$ be a subset of the objects having have l-different values of attribute a_j, then μ_{ik} and $\lambda^k_{\frac{U}{a_j}}$ are local maximum for $J_m(\mu,\lambda)$ only if

$$\lambda^k_{U/a_j}=\frac{\sum_{i\in\{X(a_j=\gamma_{jl})\}}\mu_{ik}^m}{\sum_{i=1}^{I}\mu_{ik}^m},\ for\ \ l=1,2,\ldots,l_j \tag{18}$$

and

$$\mu_{ik}=\left[\sum_{s=1}^{k}\left[\frac{\sum_{j=1}^{J}\ln(\lambda^k_{\frac{U}{a_j}})}{\sum_{j=1}^{J}\ln(\lambda^s_{\frac{U}{a_j}})}\right]^{\frac{1}{m-1}}\right]^{-1} \tag{19}$$

Proof. The problem occur maximizing $J_m(\mu,\lambda)$ with respect to μ_{ik} *and* $\lambda^k_{\frac{U}{a_j}}$under contrains of (16) and (17). Using Lagrangian multiplier method, the problem is equivalent to maximizing

$$J_m(\mu,\lambda)=\sum_{i=1}^{I}\sum_{k=1}^{K}\mu_{ik}^m\sum_{j=1}^{J}\ln\left(\lambda^k_{\frac{U}{a_j}}\right)-w_1\left(\sum_{k=1}^{K}\mu_{ik}-1\right)-w_2\left(\sum_{j=1}^{J}\lambda^k_{\frac{U}{a_j}}-1\right) \tag{20}$$

The necessary conditions for this problem are

$$\sigma_{ii}=\frac{\beta_{ii}+S_i^2-S_i}{S_i^2} \tag{21}$$

$$\frac{\partial J_m(\mu,\lambda,w_1,w_2)}{\partial\lambda^k_{\frac{U}{a_j}}}=\frac{\sum_{i=1}^{i}\mu_{ik}^m}{\lambda^k_{\frac{U}{a_j}}}-w_2=0 \tag{22}$$

$$\frac{\partial J_m(\mu,\lambda,w_1,w_2)}{\partial w_1}=\sum_{k=1}^{K}\mu_{ik}-1=0 \tag{23}$$

$$\frac{\partial J_m(\mu,\lambda,w_1,w_2)}{\partial w_2}=\sum_{j=1}^{J}\lambda^k_{\frac{U}{a_j}}-1=0 \tag{24}$$

From (21), we obtain

$$m\mu_{ik}^{m-1}=\frac{w_1}{\sum_{j=1}^{J}\ln\left(\lambda^k_{U/a_j}\right)}$$

$$\mu_{ik}=\left(\frac{w_1}{m\sum_{j=1}^{J}\ln\left(\lambda^k_{U/a_j}\right)}\right)^{\frac{1}{m-1}} \tag{25}$$

Subtituting (25) into (23),

$$\sum_{k=1}^{K}\left(\frac{w_1}{m\sum_{j=1}^{J}\ln\left(\lambda^k_{U/a_j}\right)}\right)^{\frac{1}{m-1}}-1=0$$

$$\left(\frac{w_1}{m}\right)^{\frac{1}{m-1}}=\frac{1}{\sum_{k=1}^{K}\left(\frac{1}{\sum_{j=1}^{J}\ln\left(\lambda^k_{U/a_j}\right)}\right)^{\frac{1}{m-1}}} \tag{26}$$

Substituting (26) into (25), we get (19)

$$\mu_{ik}=\left[\sum_{s=1}^{k}\left[\frac{\sum_{j=1}^{J}\ln(\lambda^k_{\frac{U}{a_j}})}{\sum_{j=1}^{J}\ln(\lambda^s_{\frac{U}{a_j}})}\right]^{\frac{1}{m-1}}\right]^{-1}.$$

And also, from (22), we have

$$\lambda^k_{\frac{U}{a_j}}=\frac{\sum_{i=1}^{i}\mu_{ik}^m}{w_2}. \tag{27}$$

Substituting (27) into (24),

$$\sum_{j=1}^{J}\frac{\sum_{i=1}^{i}\mu_{ik}^{m}}{w_2}=1$$

$$w_2=\sum_{j=1}^{J}\sum_{i=1}^{i}\mu_{ik}^{m}. \tag{28}$$

Substituting (28) into (27),

$$\lambda_{\frac{U}{a_j}}^{k}=\frac{\sum_{i=1}^{i}\mu_{ik}^{m}}{\sum_{j=1}^{J}\sum_{i=1}^{i}\mu_{ik}^{m}}. \tag{29}$$

Based on Definition 3, (29) can be obtained as in (18)

$\lambda_{U/a_j}^{k}=\frac{\sum_{i\in\{X(a_j=\gamma_{jl})\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}$; $l=1,2,\ldots,l_j$.

The proposition applies the indiscernibility relation to Definitions 2 and 3 to analyze the categorical data. The indiscernibility relation is intended to express the 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. Fig. 1 shows the pseudo-code of the proposed algorithm.

To clearly depict the idea of the proposed algorithm, we illustrate an example from a given Boolean data set adopted from (Pawlak, 1999). The following table is a modified information system from example 3 as in (Pawlak, 1999).

From Table 1, there are 5 instances with 4 categorical attributes. Based on Definition 1 and each of the attribute, there are four partitions of U induced by indescernibility relation on each attribute, i.e.

$X(a=No)=1,2,5$
$X(a=Yes)=3,4$
$\frac{U}{a}=\{\{1,2,5\},3,4\}$
$X(b=Bad)=1$
$X(b=Good)=2,3,4,5$
$\frac{U}{b}=\{\{1\},2,3,4,5\}$
$X(c=Low)=1,2,3,4$
$X(c=High)=5$
$\frac{U}{c}=\{\{12,3,4\},5\}$
$X(d=small)=1$
$X(d=Medium)=3,4$
$X(d=Large)=2,5$
$\frac{U}{d}=\{\{1\},\{3,4\},2,5\}$

Fuzzy indescernibility based Algorithm
Input: Categorical data set
Output: Clusters

Begin
1. Compute the equivalent classes using indescernibility relation
2. Compute the random initial μ_{ik}
3. Repeat
   a. Update λ_{U/a_j}^{k} applying equation (18)
   b. Update μ_{ik} applying equation (19)
4. Until μ_{ik} estimates to be stabil ($|J_{m}^{it}{}_{m}(\mu,\lambda)-J_{m}^{it-1}{}_{m}(\mu,\lambda)|<tol$ or $\|\mu_{ik}^{it}-\mu_{ik}^{it-1}\|<tol$) or given maximum number of iteration.

End

Fig. 1. Proposed algorithm.

Table 1
A modified information system.

U/A	A	B	c	D
1	No	Bad	Low	Small
2	No	Good	Low	Large
3	Yes	Good	Low	Medium
4	Yes	Good	Low	Medium
5	No	Good	High	Large

Table 2
The random initial of membership function.

i	$k=1$	$k=2$
1	0.2	0.8
2	0.6	0.4
3	0.8	0.2
4	0.3	0.7
5	0.4	0.6
$\sum$	2.3	2.7

Given the random initial of membership functions described in Table 2 as follows:

For $j=1$ that is attribute a, there are 2 different ($L_j=2$) values. Based on Eq. (16), the maximize parameter $\lambda_{\frac{U}{a_j}}^{k}$; for $k=1$ can be obtained as follows:

$$\lambda_{\{1,2,5\}}^{1}=\frac{\sum_{i\in\{X(a=No)\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{\sum_{i\in\{1,2,5\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{0.2+0.6+0.4}{2.3}=0.5217$$

$$\lambda_{\{3,4\}}^{1}=\frac{\sum_{i\in\{X(a=Yes)\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{\sum_{i\in\{3,4\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{0.8+0.3}{2.3}=0.4783,$$

and the maximize parameter $\lambda_{\frac{U}{a_j}}^{k}$; for $k=2$ can be obtained as follows:

$$\lambda_{\{1,2,5\}}^{2}=\frac{\sum_{i\in\{X(a=No)\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{\sum_{i\in\{1,2,5\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{0.8+0.4+0.6}{2.7}=0.6667$$

$$\lambda_{\{3,4\}}^{2}=\frac{\sum_{i\in\{X(a=Yes)\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{\sum_{i\in\{3,45\}}\mu_{ik}}{\sum_{i=1}^{I}\mu_{ik}}=\frac{0.2+0.7}{2.7}=0.333.$$

Following the same procedure, the maximization parameters of all attribute in Table 1 can be summarized in Table 3 (for $k=1$) and Table 4 (for $k=2$) as follows:

The new membership function of fuzzy k-Partition based on indescernibility can be obtained by using Eq. (17). If given the fuzziness index of $m=1.4$, then the new membership function for $i=1$ in the first cluster ($k=1$) is given below:

$$\mu_{11}=\left[\sum_{s=1}^{2}\left[\frac{\sum_{j=1}^{J}\ln\left(\lambda_{\frac{U}{a_j}}^{1}\right)}{\sum_{j=1}^{J}\ln\left(\lambda_{\frac{U}{a_j}}^{s}\right)}\right]^{\frac{1}{1.4-1}}\right]^{-1}$$

$$=\left[\left[\frac{-5.7263}{-5.7263+(-3.0896)}\right]^{\frac{1}{1.4-1}}\right]^{-1}=0.0021$$

Following the same procedure, the new membership functions are computed. The calculations results are summarized in Table 5.

Based on the membership function in Table 5 (after 17 iterations), the obtained clusters are {1, 2, 5} as the first cluster and {3, 4} as the second cluster.

Table 3
The maximization parameter in Table 1 for 1st cluster.

i	$\lambda^1_{\frac{U}{a}}$	$\ln(\lambda^1_{\frac{U}{a}})$	$\lambda^1_{\frac{U}{b}}$	$\ln(\lambda^1_{\frac{U}{b}})$	$\lambda^1_{\frac{U}{c}}$	$\ln(\lambda^1_{\frac{U}{c}})$	$\lambda^1_{\frac{U}{d}}$	$\ln(\lambda^1_{\frac{U}{d}})$	$\sum_{j=1}^{J}\ln(\lambda^1_{\frac{U}{a_j}})$
1	0.5217	−0.6506	0.0870	−2.4423	0.8261	−0.1911	0.0870	−2.4423	−5.7263
2	0.5217	−0.6506	0.9130	−0.0910	0.8261	−0.1911	0.4348	−0.8329	−1.7655
3	0.4783	−0.7376	0.9130	−0.0910	0.8261	−0.1911	0.4783	−0.7376	−1.7572
4	0.4783	−0.7376	0.9130	−0.0910	0.8261	−0.1911	0.4783	−0.7376	−1.7572
5	0.5217	−0.6506	0.9130	−0.0910	0.1739	−1.7492	0.4348	−0.8329	−3.3237

Table 4
The maximization parameter in Table 1 for 2nd cluster.

i	$\lambda^2_{\frac{U}{a}}$	$\ln(\lambda^2_{\frac{U}{a}})$	$\lambda^2_{\frac{U}{b}}$	$\ln(\lambda^2_{\frac{U}{b}})$	$\lambda^2_{\frac{U}{c}}$	$\ln(\lambda^2_{\frac{U}{c}})$	$\lambda^2_{\frac{U}{d}}$	$\ln(\lambda^2_{\frac{U}{d}})$	$\sum_{j=1}^{J}\ln(\lambda^2_{\frac{U}{a_j}})$
1	0.6667	−0.4055	0.2963	−1.2164	0.7778	−0.2513	0.2963	−1.2164	−3.0896
2	0.6667	−0.4055	0.7037	−0.3514	0.7778	−0.2513	0.3704	−0.9933	−2.0014
3	0.3333	−1.0986	0.7037	−0.3514	0.7778	−0.2513	0.3333	−1.0986	−2.7999
4	0.3333	−1.0986	0.7037	−0.3514	0.7778	−0.2513	0.3333	−1.0986	−2.7999
5	0.6667	−0.4055	0.7037	−0.3514	0.2222	−1.5041	0.3704	−0.9933	−3.2542

In the next section, we perform experiment with the proposed algorithm based on benchmark and realworld data. We also compare the results obtained with two other Fuzzy k-based approaches in terms of computational time and cluster purity.

4. Experiment results

In this section, we compare the proposed approach with the Fuzzy Centroid and fuzzy k-Partition approaches based on computational complexity, estimation parameters, response time and clustering accuracy. The responses time are calculated by the time needed to execute the algorithm by computer and the clustering accuracies are analyzed using internal criteria and external criteria.

In the experiment, the proposed approach and other two fuzzy k-based approaches are implemented in MATLAB version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total main memory is 2G and the operating system is Windows 8. In this section, we have two different clustering experimentations from real datasets.

4.1. Computational complexity

The computation complexity of the proposed algorithm will be discussed and compared with two existing Fuzzy k-based algorithms. From the following theoretical analysis, a conclusion can be drawn that the proposed algorithm achieved lower computational complexity as compared to (Huang, 1998; Yang et al., 2008).

4.1.1. Computational complexity of Fuzzy Centroid

The time complexity required mainly depends on the updates of the Fuzzy Centroid v_{kjl} and partition matrix μ_{ik} in each iteration. The computational costs of updating the Fuzzy Centroid and partition matrix are $O(KIM)$and $O(KIJ)$, respectively. Thus, the overall complexity for Fuzzy Centroid algorithm is $O(KI(M+J)t)$, where t is number of iteration, k is the number of cluster, I is the number of data, J is the number of attributes, and $M=\sum_{j=1}^{J} L_j$. Similary, the overall computational complexity for Fuzzy Centroids algorithm is $O(2KIMt)$.

Table 5
The new membership functions.

	1st iteration		After 17 iterations	
i	μ_{i1}	μ_{i2}	μ_{i1}	μ_{i2}
1	0.0021	0.9979	0.0140	0.9860
2	0.7780	0.2220	0.0055	0.9945
3	0.9906	0.0094	1.0000	0.0000
4	0.9906	0.0094	1.0000	0.0000
5	0.4474	0.5526	0.0051	0.9949

4.1.2. Computational complexity of Fuzzy k-Partition

The computational complexity of the Fuzzy k-Partition is calculated at each iteration based on two parts, parameters λ_{kjl} and fuzzy partition μ_{ik}, with $O(KIM)$ and $O(KIM)$, respectively, so that the complexity is $O(2KIMt)$.

4.1.3. Computational complexity of the proposed approach

The proposed approach need $O(IJ)$ to contruct equivalence classes $\frac{U}{a_j}$ based on indicernibility and the complexity at each iteration from two parts, parameter $\lambda^k_{\frac{U}{a_j}}$ and fuzzy partition μ_{ik} are $O(KM)$ and $O(KIM)$, respectively. Thus, the computational complexity for the proposed approach is the polynomial of $O(KM(I+1)t+IJ)$.

The following table presents the comparative analysis result in terms of computational complexity.

From Table 6, we can see that the proposed Fuzzy indescernibility approach has the smallest computational complexity.

4.2. Parameter estimation

In this section, the algorithms are implemented to estimate the parameters of multivariate multinomial mixtures where the data points are assumed from the mixture distribution $f(y,\lambda)$.

$$f(y,\lambda)=\sum_{k=1}^{K}\alpha_k f_k(y,\lambda)\text{ with }f_k(y,\lambda)=\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\lambda_{kjl})^{y_{jl}}. \quad (30)$$

The mixing proposition α_k are estimated by $\alpha_k=\sum_{i=1}^{I}\frac{g_{ik}}{I}$, $\alpha_k=\sum_{i=1}^{I}\frac{\mu_{ik}}{I}$, $k=1,\ldots,K$, where g_{ik} and μ_{ik} are the final output from the Fuzzy Centroid, Fuzzy k-Partition and Fuzzy indescernibility, respectively. We use the numerical data drawn from multivariate binomial mixtures distribution as in (Yang et al., 2008). The algorithms are implemented to estimate the parameters of four-

Table 6
A comparison of computational complexity.

Algorithms	Computational complexity
Fuzzy Centroid [4]	$O(2KIMt)$
Fuzzy k-Partition [12]	$O(2KIMt)$
Fuzzy indescernibility	$O(KM(I+1)t+IJ)$

attribute binomial mixture of two classes using random samples drawn from the mixture distribution $f(y,\lambda)$ as in (31)

$$f(y,\lambda)=\sum_{k=1}^{2}\alpha_k f_k(y,\lambda_k)=\alpha B(1,\lambda_{11})B(1,\lambda_{12})B(1,\lambda_{13})B(1,\lambda_{14}) +(1-\alpha)B(1,\lambda_{21})B(1,\lambda_{22})B(1,\lambda_{23})B(1,\lambda_{24}), \quad (31)$$

where $B(1,p)$ is a Bernoulli distribution. The combinations of three different mixing α propositions and three different λ values considered for mixture model (31) are given as follows

$$\alpha_1=(0.1\ 0.9),\qquad \alpha_2=(0.4\ 0.6),\qquad \alpha_3=(0.7\ 0.3)$$

$$\lambda_a=\begin{pmatrix}0.9 & 0.7 & 0.4 & 0.3\\ 0.4 & 0.3 & 0.5 & 0.8\end{pmatrix},\lambda_b=\begin{pmatrix}0.3 & 0.4 & 0.5 & 0.6\\ 0.6 & 0.5 & 0.4 & 0.3\end{pmatrix}$$

The algorithms are implemented for random sample from mixture distribution of all combinations under eight different fuzifiers $m\in[1.11.9]$ and random initial value as $g_{ik}=\mu_{ik}$. The responses time (RT) and Mean Square Error (MSE) of parameters between the estimates and true parameters for these eight fuzifiers are computed. Table 7 shows the results and according to Table 7, the average MSEs of parameters using the Fuzzy indescernibility and Fuzzy k-Partition algorithms were almost lower than those of Fuzzy Centroid. The average responses times of Fuzzy indescernibility are also lower than those of Fuzzy k-Partition and Fuzzy Centroid. These results indicate that Fuzzy indescernibility possesses more accuracy and efficiency than Fuzzy k Partition and Fuzzy Centroid Table 8.

4.3. Experiment on real datasets

This sub-section explains and discusses the experimental results of the proposed approach. The main focus of the experiments is on the performance measurement of the proposed approach in which execution time and accuracy are used as parameters. For comparisons, the clusters purity is commonly used as a measure to test the quality of clustering accuracy. The purity of a cluster as described in Parmar et al. (2007) is defined as in (32):

$$\text{Purity}(i)=\frac{\text{the number of data occuring in both the } i\text{th cluster and its corresponding class}}{\text{the number of data in the data set}}$$

$$\text{Overall Purity}=\frac{\sum_{i=1}^{\#\text{ of cluster}}\text{Purity}(i)}{\#\text{ of cluster}} \quad (32)$$

According to the above measure, a higher value of overall purity indicates a better clustering result, with perfect clustering yielding a value of 100% (Gibson et al., 2000). Other external used to analyze the cluster is Rand Measure. The adjusted Rand index (Hubert and Arabie, 1985) is the corrected-for-chance version of the Rand index that computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. The Adjusted Rand Index is as in (33)

$$RI=\frac{\sum_{i=1}^{m}\sum_{j=1}^{K}\binom{n_{ij}}{2}-\binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{K}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{m}\binom{n_{i.}}{2}+\sum_{j=1}^{K}\binom{n_{.j}}{2}\right]-\binom{n}{2}^{-1}\sum_{i=1}^{m}\binom{n_{i.}}{2}\sum_{j=1}^{K}\binom{n_{.j}}{2}}, \quad (33)$$

where n_{ij} represents the number of objects that are in predefined class i and cluster j, $n_{i.}$ indicates the number of objects in a priori class i, $n_{.j}$ indicates the number of objects cluster j, and n is the total number of objects in the data set.

Davies Bouldin index and Dunn index are used to assess the quality of clustering algorithms based on internal criterion. Davies Bouldin index attempts to minimize the average distance between each cluster and the one most similar to it (Davies and Bouldin, 1979). It is defined as in (34)

$$DB=\frac{1}{K}\sum_{k=1}^{K}\max_{k\neq m}\left(\frac{\sigma_k+\sigma_m}{d(c_k,c_m)}\right) \quad (34)$$

where K is the number of clusters, σ_k is the average distance of all elements in cluster k and $d(c_k,c_m)$ is the distance between cluster k and cluster m. The clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion. Dunn's Validity Index (Dunn, 1974) attempts to identify those cluster sets that are compact and well separated. The Dunn's validation index can be calculated with the following formula in (35):

$$Dn=\min_{1\le k\le K}\left(\min_{k+1\le m\le K}\left(\frac{d(c_k,c_m)}{\max_{1\le n\le k}d'(n)}\right)\right) \quad (35)$$

where $d(c_i,c_j)$ represents the inter cluster distance between cluster k and cluster m. It may be any number of distance measure, such as the distance between the centroids of the cluster. $d'(n)$ is called the intra cluster distance of cluster n that may be measured in variety mays, such as the maximal distance between any pair of element in cluster n. In the experiment, all distance are calculated using Hamming distance.

We elaborate the three approaches through the UCI benchmark datasets as follow:

a. Zoo data set which is comprised of 101 objects, where each data point represents information of an animal in terms of 18 categorical attributes.
b. Soybean data set contains 47 instances and 35 categorical attributes.
c. Balloon dataset which contains 20 instances and 4 categorical attributes.
d. Breast Cancer dataset which contains 699 instances and 9 attributes.
e. Tic-tac-toe dataset which contains 958 instances and 9 attributes.
f. Monk dataset which contains 432 instances and 6 attributes.
g. Spect dataset which contains 187 instances and 922 attributes.
h. Car dataset which contains 1728 instances and 6 attributes.

All approachs are run partially given one initial membership function μ_{ik}. The matrix initial membership μ_{ik} is a random matrix input for all approaches satisfying the constraint (16). Generally, the random matrix μ_{ik} can be obtained by generating a random matrix I number of object, K number of clusters and then divided by the sum rows on each column, respectively. From 100 times implementation of all approaches for the Zoo, Breast cancer, Ballon, Soybean, Tic-tac-toe, Monk, Spect and Car datasets in varying fuzziness index i.e. $m\in[1.1,2.0]$ and then the average accuracy rates are calculated. The results show that the best average accuracy of Fuzzy k-Partition and proposed Fuzzy indescernibility approaches in all case is almost the same, and it is better than that Fuzzy Centroid for Zoo, Breast cancer, Ballon, and Soybean datasets. The improvement is summarized in Table 9. All the approaches have no significant difference in term of accuracy rate for Tic-tac-toe, Monk, Spect and Car datasets.

Table 7
MSE and response time for tests of all combination of α and λ.

α_1, λ_a	**Fuzzy Centroid**			**Fuzzy *k* Partition**			**Fuzzy Indescernible**		
m	λ	α	RT	λ	α	RT	λ	α	RT
1.1	0.3621	0.0004	0.1880	0.2424	0.0533	0.1710	0.2424	0.0533	0.0470
1.2	0.3000	0.5600	0.2180	0.2308	0.3299	0.2500	0.2308	0.3299	0.0620
1.3	0.1754	0.0011	0.4840	0.2403	0.0920	0.3120	0.2403	0.0920	0.1560
1.4	0.1735	0.5285	0.2660	0.1930	0.3307	0.6240	0.1930	0.3307	0.4050
1.5	0.2141	0.2099	0.4520	0.2058	0.3379	0.5780	0.2058	0.3379	0.3900
1.6	0.2160	0.1600	0.3590	0.2091	0.3478	0.4370	0.2091	0.3478	0.2650
1.7	0.2086	0.1600	0.2660	0.2100	0.0434	0.4360	0.2100	0.0434	0.2660
1.8	0.2075	0.1600	0.2180	0.2170	0.3294	0.4370	0.2170	0.3294	0.2650
1.9	0.2404	0.1600	0.2490	0.2090	0.0188	0.2810	0.2090	0.0188	0.1250
Average	0.2331	0.2155	0.3000	0.2175	0.2092	0.3918	0.2175	0.2092	0.2201
α_1, λ_b									
1.1	0.3291	0.0009	0.1720	0.2683	0.0156	0.1720	0.2683	0.0156	0.0310
1.2	0.3067	0.6525	0.2030	0.2711	0.4423	0.2030	0.2711	0.4423	0.0620
1.3	0.3043	0.6277	0.2490	0.2748	0.4150	0.2810	0.2748	0.4150	0.1090
1.4	0.2418	0.5922	0.2650	0.2709	0.3861	0.2810	0.2709	0.3861	0.1410
1.5	0.2753	0.0036	0.2650	0.2789	0.0162	0.3430	0.2789	0.0162	0.1720
1.6	0.2739	0.5157	0.2810	0.2486	0.4947	0.2650	0.2486	0.4947	0.1090
1.7	0.3239	0.1595	0.4680	0.2672	0.0093	0.3120	0.2672	0.0093	0.1560
1.8	0.3304	0.1600	0.3900	0.2607	0.5244	0.2650	0.2607	0.5244	0.0940
1.9	0.3147	0.1600	0.2500	0.2838	0.0141	0.3430	0.2838	0.0141	0.1870
Average	0.3000	0.3191	0.2826	0.2694	0.2575	0.2739	0.2694	0.2575	0.1179
α_2, λ_a									
1.1	0.2027	0.1883	0.1880	0.2108	0.1184	0.2020	0.2108	0.1184	0.0630
1.2	0.3413	0.0560	0.2020	0.2211	0.0011	0.2340	0.2211	0.0011	0.0940
1.3	0.1685	0.1865	0.4050	0.2001	0.0740	0.3280	0.2001	0.0740	0.1720
1.4	0.1942	0.1621	0.2810	0.2337	0.0757	0.6090	0.2337	0.0757	0.4360
1.5	0.1999	0.0111	0.4520	0.1898	0.0693	0.5770	0.1898	0.0693	0.3750
1.6	0.2110	0.0100	0.3120	0.2270	0.0077	0.6240	0.2270	0.0077	0.4210
1.7	0.2286	0.0100	0.2800	0.2576	0.0121	0.3750	0.2576	0.0121	0.2030
1.8	0.2086	0.0100	0.2190	0.2041	0.0088	0.3900	0.2041	0.0088	0.2490
1.9	0.2124	0.0100	0.1870	0.2023	0.0749	0.3770	0.2023	0.0749	0.2210
Average	0.2186	0.0715	0.2807	0.2163	0.0491	0.4129	0.2163	0.0491	0.2482
α_2, λ_b									
1.1	0.3859	0.2754	0.1870	0.2779	0.1559	0.1870	0.2779	0.1559	0.0310
1.2	0.2752	0.1117	0.2020	0.2700	0.0380	0.2030	0.2700	0.0380	0.0630
1.3	0.3137	0.2418	0.2810	0.2722	0.1116	0.2800	0.2722	0.1116	0.1250
1.4	0.2230	0.2480	0.2500	0.2567	0.1241	0.3280	0.2567	0.1241	0.1720
1.5	0.2748	0.2569	0.2180	0.2725	0.1874	0.2500	0.2725	0.1874	0.0940
1.6	0.2994	0.1125	0.4530	0.2622	0.1487	0.3430	0.2622	0.1487	0.1720
1.7	0.3275	0.1358	0.4370	0.2714	0.2045	0.2650	0.2714	0.2045	0.1090
1.8	0.3240	0.0100	0.2810	0.2571	0.1679	0.3280	0.2571	0.1679	0.1560
1.9	0.3522	0.0100	0.3120	0.2677	0.2073	0.2970	0.2677	0.2073	0.1250
Average	0.3084	0.1558	0.2912	0.2675	0.1495	0.2757	0.2675	0.1495	0.1163
α_3, λ_a									
1.1	0.2119	0.3185	0.1870	0.2043	0.1480	0.1870	0.2043	0.1480	0.0470
1.2	0.3658	0.3002	0.2180	0.2570	0.1090	0.2500	0.2570	0.1090	0.0930
1.3	0.1658	0.2907	0.3590	0.2453	0.0683	0.3900	0.2453	0.0683	0.2030
1.4	0.1963	0.2365	0.2810	0.2317	0.1179	0.5770	0.2317	0.1179	0.3900
1.5	0.2091	0.0864	0.4680	0.2182	0.1381	0.4210	0.2182	0.1381	0.2180
1.6	0.2266	0.0400	0.4990	0.2357	0.1761	0.4220	0.2357	0.1761	0.2340
1.7	0.2391	0.0400	0.2970	0.2217	0.0008	0.4210	0.2217	0.0008	0.2340
1.8	0.2128	0.0400	0.2490	0.1945	0.1587	0.3590	0.1945	0.1587	0.1870
1.9	0.2245	0.0400	0.2340	0.2181	0.1703	0.3280	0.2181	0.1703	0.1710
Average	0.2280	0.1547	0.3102	0.2252	0.1208	0.3728	0.2252	0.1208	0.1974
α_3, λ_b									
1.1	0.3770	0.3656	0.1720	0.2814	0.2025	0.1870	0.2814	0.2025	0.0150
1.2	0.3772	0.3757	0.2020	0.2731	0.2116	0.2810	0.2731	0.2116	0.0630
1.3	0.2418	0.3752	0.2960	0.2652	0.2104	0.2650	0.2652	0.2104	0.1100
1.4	0.2552	0.3304	0.2960	0.2763	0.1895	0.4210	0.2763	0.1895	0.3120
1.5	0.2566	0.3132	0.2190	0.2545	0.2312	0.2810	0.2545	0.2312	0.1090
1.6	0.2955	0.0009	0.4990	0.2832	0.0041	0.4060	0.2832	0.0041	0.2180
1.7	0.3106	0.0400	0.5000	0.2959	0.2253	0.3270	0.2959	0.2253	0.1560
1.8	0.3292	0.0400	0.4360	0.2782	0.2696	0.3120	0.2782	0.2696	0.1410
1.9	0.3326	0.0400	0.3160	0.2748	0.0157	0.2980	0.2748	0.0157	0.1410
Average	0.3084	0.2090	0.3262	0.2758	0.1733	0.3087	0.2758	0.1733	0.1406

However, the proposed Fuzzy indescernibility approach has lower executing time due to less computation required as shown in Table 10. For example, for Monk dataset, the executing time for fuzzy indescernibility based is 0.0472 s, while the executing times for Fuzzy *k*-Partition is 0.1323 s and Fuzzy Centroid is 0.4959 s. Thus in this case, the proposed approach improves the executing time of fuzzy *k*-Partition on the average up to 49.72%. Figs. 2 and 3 show the average of Index value based on internal evaluation using Davies Bouldin and external evaluation using adjusted rand index, respectively. From the graph it is clear that we are getting

Table 8
Average MSE and respone time of parameters for all tests.

	Fuzzy Centroid			Fuzzy *k* Partition			Fuzzy Indescernible		
	λ	α	RT	λ	α	RT	λ	α	RT
α_1, λ_a	0.2331	0.2155	0.3000	0.2175	0.2092	0.3918	0.2175	0.2092	0.2201
α_1, λ_b	0.3000	0.3191	0.2826	0.2694	0.2575	0.2739	0.2694	0.2575	0.1179
α_2, λ_a	0.2186	0.0715	0.2807	0.2163	0.0491	0.4129	0.2163	0.0491	0.2482
α_2, λ_b	0.3084	0.1558	0.2912	0.2675	0.1495	0.2757	0.2675	0.1495	0.1163
α_3, λ_a	0.2280	0.1547	0.3102	0.2252	0.1208	0.3728	0.2252	0.1208	0.1974
α_3, λ_b	0.3084	0.2090	0.3262	0.2758	0.1733	0.3087	0.2758	0.1733	0.1406

Table 9
The accuracy improvement of Fuzzy indescernibility to Fuzzy Centroid and Fuzzy *k*-Partition.

	Fuzzy Centroid	Fuzzy *k*-Partition	Fuzzy indescernibility	Improvement
Breast cancer	0.9255	0.9717	0.9717	4.99 %
Zoo	0.832	0.8996	0.8996	8.13 %
Balloon	0.8333	1	1	20.00 %
Soybean	0.9720	1	1	2.88 %
Average of limprovement				9 .00 %

Table 10
The response time improvement of Fuzzy indescernibility to Fuzzy Centroid and Fuzzy *k*-Partition.

	Fuzzy Centroid	Fuzzy *k*-Partition	Fuzzy Indescernibility	Improvement (%)
Breast cancer	0.9237	2.0070	0.5417	73.01
Zoo	0.9986	1.0713	0.9020	15.80
Balloon	0.4296	1.8337	1.3440	26.71
Soybean	0.7985	0.0585	0.0480	17.88
Tic-tac-toe	0.6520	1.2325	0.2481	79.87
Monk	0.4959	0.1323	0.0472	64.34
Spect	0.8108	0.3206	0.2095	34.65
Car	0.7035	0.7037	0.1021	85.49
Average of improvement				49.72

best performance using fuzzy indescernibility in terms of the Dunn's Davies Bouldin index and Rand index.

In the following section, we present further on applicability of the proposed Fuzzy indiscrnibility approach on other real world dataset. The first data set is studies anxiety dataset and the second data is supplier management dataset.

4.3.1. Studies anxiety dataset

The nature, degree, and persistence of stress are particularly important in anxiety disorders and other related psychiatric conditions (Bystritsky and Kronemyer, 2014). In this case, the studies anxiety refers to anxiety condition during study. High level of anxiety is perceived to relate to the low performances in academic. Such anxiety can interfere with students'performance on exam (Harris and Coy, 2003; McCraty, 2003). The study's anxiety data used were taken from a survey aimed to identify cause of study anxiety among university students. The respondents are 770 students which consist of 395 males and 375 females (Yanto et al., 2012). We take two results of studies anxiety from (Yanto et al., 2012) i.e. mathematic anxiety and social anxieties. There are five attributes of mathematics anxiety; Felt mathematic is difficult subject (DS), Always fail in mathematic (FM), Always writing down while in mathematic class (WD), Anxious if do not understand (DU), Lost of interest in mathematic (Li). We run all the approaches

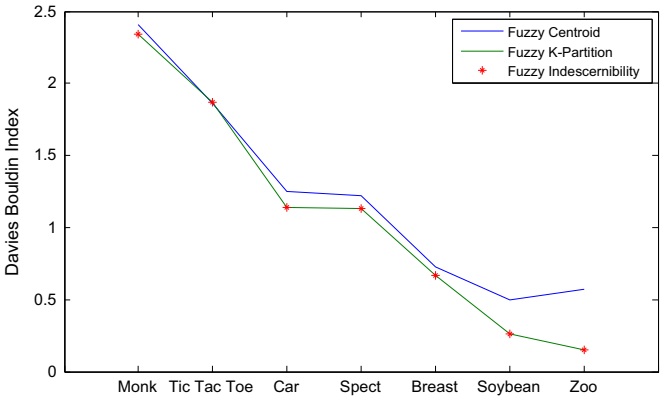

Fig. 2. Davies Bouldin Index for benchmark data sets.

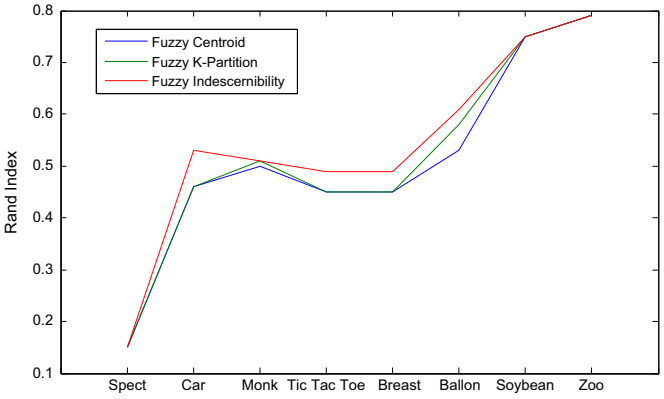

Fig. 3. Adjusted Rand index of benchmarks data sets.

for mathematics anxiety dataset in two different numbers of clusters. The experiment we define nine diferent number of clusters i.e 2–10 clusters with $m = 1.3$. The Davies Bouldin Index and Dunn Index are shown in Figs. 4 and 5, respectively. From the graph it is clear that we are getting best performance using fuzzy indescernibility and Fuzzy *k*-Partition in terms of the Dunn's validity index and Davies Bouldin index. However, from Fig. 6, we can see that the fuzzy indescernibility achieves lower of response time than Fuzzy *k*-Partition.

There are seven attributes in Social anxiety dataset i.e. Problem with peers (PP), Uncomfortable hostel (UH), Problem with roommate (PR), Home-sick (HS), Uncomfortable with the campus environment (CE), Racial diversity (RD), Difficult to study because of many roommates (MR). We run all the approaches for social anxiety dataset in nine experiments. In the experiment, we define two to ten clusters with $m = 1.3$. The Davies Bouldin Index and Dunn Index obtained are shown in Figs. 7 and 8, respectively. From both Figure, we can see that the the Fuzzy indescernibility and

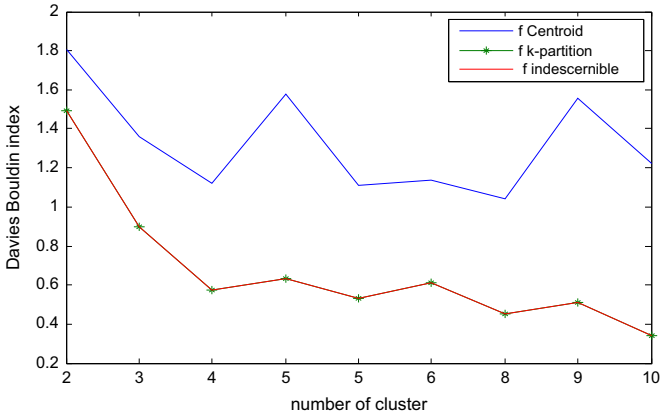

Fig. 4. Davies Bouldin index for mathematics anxiety data set.

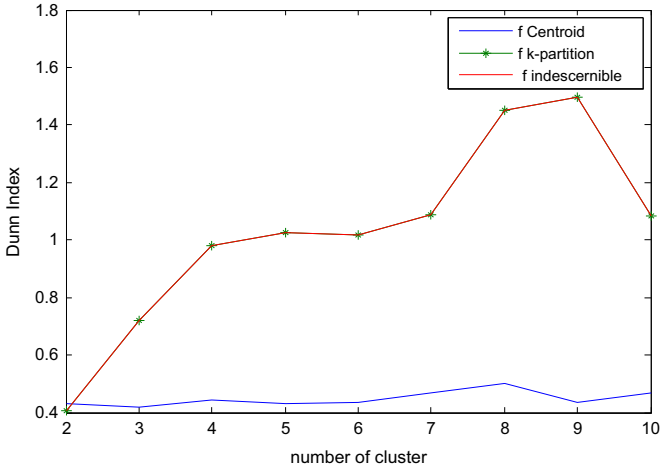

Fig. 5. Dunn Index for mathematics anxiety data set.

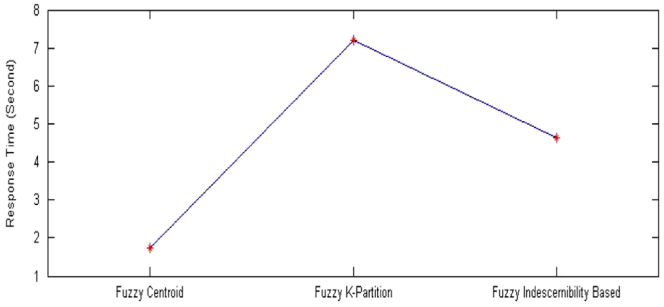

Fig. 6. Response time comparison for mathematics anxiety data set.

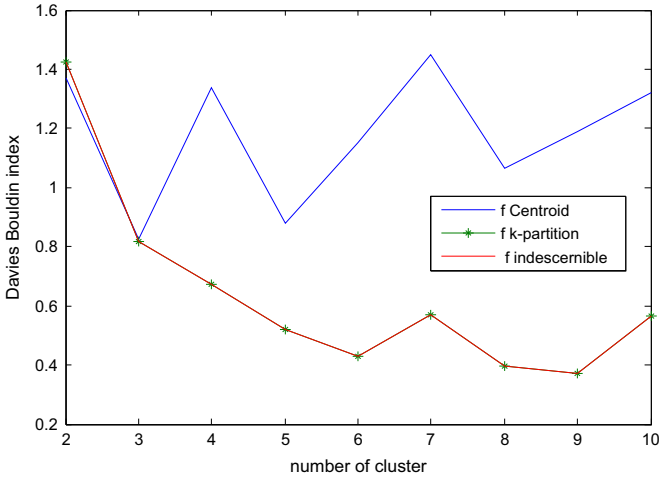

Fig. 7. Davies Bouldin Index for Social anxiety dataset.

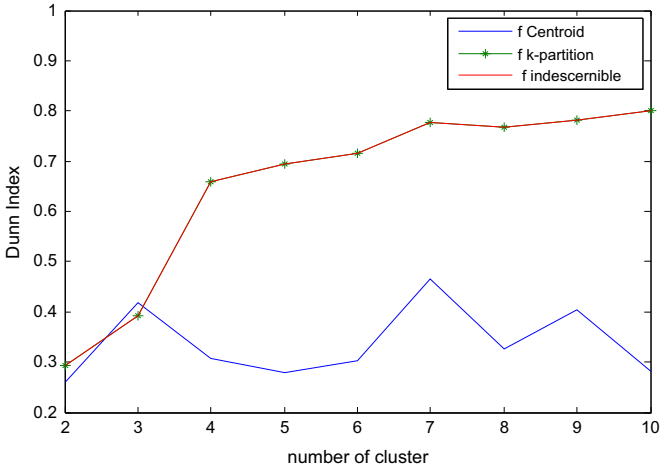

Fig. 8. Dunn Index for Social anxiety dataset.

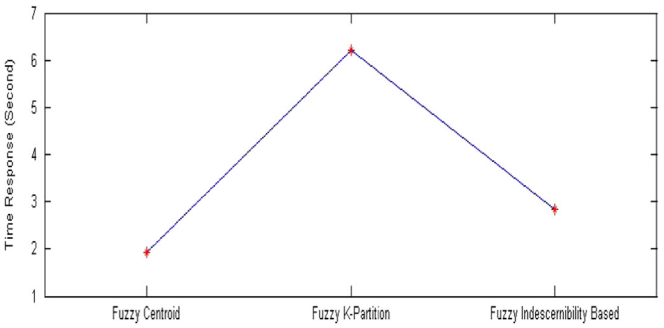

Fig. 9. Response time comparison for Social anxiety dataset.

Fuzzy *k*-Partition obtained better performance than that the Fuzzy Centroid in terms of the Dunn's validity index and Davies Bouldin index. However, from Fig. 9, we can see that the fuzzy indescernibility achieves lower of response time than Fuzzy *k*-Partition.

4.3.2. *Supplier base management data set*

A supplier base management dataset is obtained from (Herawan et al., 2010; Liu and Jiang, 2010) as shown in Table 11. From the dataset, there are 27 suppliers with 7 quantitative attributes, namely Quality system outcome (QSO), Claims (CL), Quality improvement (QI), Response to claims (RC), On-time delivery (OD), Internal audit (IA), and Data administration (DA).

We run all the approaches for Suplier base management dataset in nine experiments. In the experiment, we define two to ten clusters with $m = 1.3$. The Davies Bouldin Index, Dunn Index obtained and the response time are shown in Figs. 10–12, respectively. Figs. 10 and 11 show that the Fuzzy indescernibility and Fuzzy *k*-Partition obtained better performance than that the Fuzzy Centroid in terms of the Dunn's validity index and Davies Bouldin index. However, in Fig. 12, the proposed Fuzzy indescernibility approach achieves lower of response time than Fuzzy *k*-Partition.

5. Conclusion

In this paper, we studied the categorical data clustering with emphasizes on fuzzy-based approaches. This is the first study that

Table 11
The Discretized supplier dataset.

Supplier	QSO	CL	QI	RC	OD	IA	DA
S1	0	0	0	0	0	0	0
S2	0	1	1	0	0	0	0
S3	0	0	1	1	1	0	1
S4	0	0	0	0	0	1	0
S5	0	0	0	0	0	2	2
S6	1	1	0	0	0	0	3
S7	0	1	1	0	0	1	3
S8	0	1	1	0	0	1	3
S9	1	0	0	1	1	2	1
S10	0	0	0	1	1	1	3
S11	0	0	1	1	1	0	2
S12	0	0	1	1	1	1	1
S13	0	0	1	1	1	1	2
S14	0	0	1	1	1	1	2
S15	0	0	1	1	1	1	4
S16	0	0	1	1	1	2	2
S17	1	0	0	0	0	0	0
S18	1	1	0	0	0	1	3
S19	1	1	0	0	0	0	2
S20	0	0	1	1	1	0	1
S21	1	0	0	1	1	1	3
S22	1	0	0	0	0	1	1
S23	1	0	1	1	1	1	2
S24	1	0	1	1	1	2	1
S25	1	0	1	1	1	2	4
S26	1	0	1	1	1	0	0
S27	1	0	0	0	0	0	0

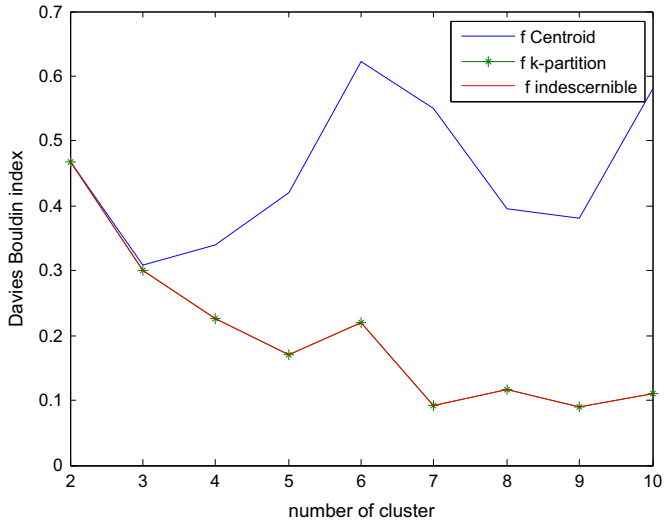

Fig. 10. Davies Bouldin Index for Suplier base management dataset.

proposed fuzzy-based categorical data clustering approach using indiscernibility relation. We have succesfully proposed an alternative algorithm of our modified Fuzzy *k*-Partition approach. Although several algorithms exist that address the issues concerning fuzzy-based categorical data clustering, none of the previous algorithms provide lower response time and higher clusters purity. We presented comparative analysis of the proposed approach theoretically on computational complexity to two well known Fuzzy approaches and it is shown that the proposed approach achieved lower complexity. Furthermore, we carried out experiments on benchmark and real world data sets to show the performance of our proposed Fuzzy indiscernibility approach in terms of response time and clusters purity. The result showed that the proposed Fuzzy indiscernibility approach out performed the two well known Fuzzy approaches in terms of lower response time and higher clusters purity up to 49.72 % and 9 %, respectively.

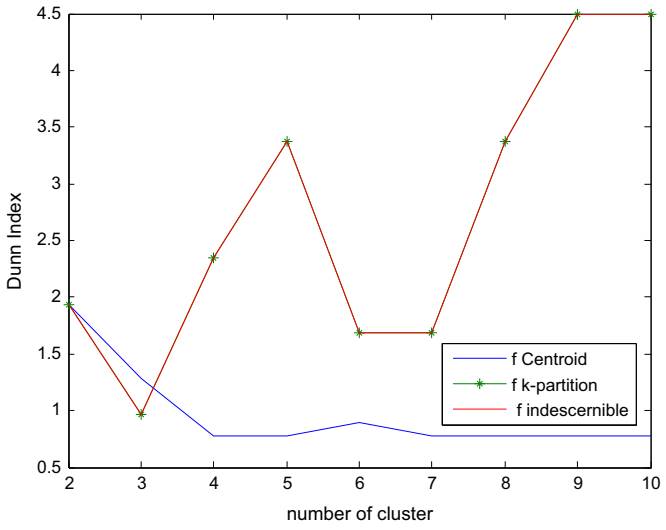

Fig. 11. Dunn Index for Suplier base management dataset.

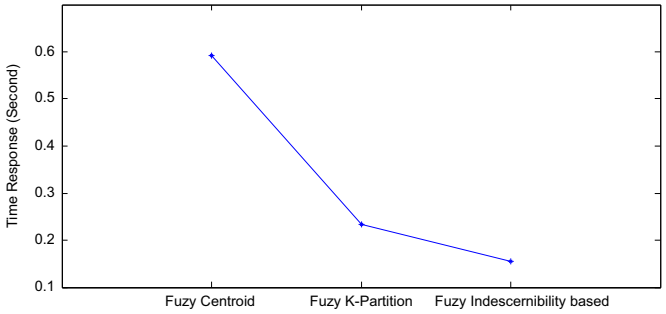

Fig. 12. Response time for Suplier base management dataset.

Acknowledgement

This work is supported by University of Malaya High Impact Research Grant no. vote UM.C/625/HIR/MOHE/SC/13/2 from Ministry of Higher Education Malaysia. The authors thank to H.K. Leang for providing all benchmark datasets.

References

Bezdek, J.C., 2013. Pattern Recognition with Fuzzy Objective Function Algorithms. Springer Science & Business Media.

Bryant, P., Williamson, J.A., 1978. Asymptotic behaviour of classification maximum likelihood estimates. Biometrika 65 (2), 273–281. http://dx.doi.org/10.1093/biomet/65.2.273.

Bystritsky, A., Kronemyer, D., 2014. Stress and anxiety: counterpart elements of the stress/anxiety complex. Psychiatric Clinics North Am. 37 (4), 489–518. http://dx.doi.org/10.1016/j.psc.2014.08.002.

Chatzis, S.P., 2011. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert Syst. Appl. 38 (7), 8684–8689. http://dx.doi.org/10.1016/j.eswa.2011.01.074.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2), 224–227.

De Carvalho, F., de, A.T., 2007. Fuzzy c-means clustering methods for symbolic interval data. Pattern Recognit. Lett. 28 (4), 423–437. http://dx.doi.org/10.1016/j.patrec.2006.08.014.

Dobosz, K., Duch, W., 2010. Understanding neurodynamical systems via Fuzzy Symbolic Dynamics. Neural Netw. 23 (4), 487–496. http://dx.doi.org/10.1016/j.neunet.2009.12.005.

Dunn†, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern. 4 (1), 95–104. http://dx.doi.org/10.1080/01969727408546059.

Gibson, D., Kleinberg, J., Raghavan, P., 2000. Clustering categorical data: an approach based on dynamical systems. VLDB J. Int. J. Very Large Data Bases 8 (3–4), 222–236. http://dx.doi.org/10.1007/s007780050005.

Haixia, X.U., Zheng, T., 2009. An optimal spectral clustering approach based on Cauchy-Schwarz Divergence, 18(1).

Harris, H.L., Coy, D.R., 2003. Helping Students Cope with Test Anxiety. ERIC Digest. ERICCounseling and Student Services Clearing House. ERIC Counseling and Student Services Clearinghouse, University of North Carolina at Greensboro, 201 Ferguson Building, P.O. Box 26170, Greensboro, NC 27402-6170. Tel: 336-334-4114; Tel: 800-414-9769 (Toll Free); Fax: 336-334-4116; e-mail: ericcas-s@uncg.ed. Retrieved from ⟨http://eric.ed.gov/?id=ED479355⟩.

He, Z., Deng, S., Xu, X., 2005. Improving K-modes algorithm considering frequencies of attribute values in mode. Comput. Intell. Secur. 3801, 157–162. http://dx.doi.org/10.1007/11596448_23.

Herawan, T., Tri, I., Yanto, R., Deris, M.M.A.T., 2010. ROSMAN : rough set approach for clustering supplier base MANagement. Int. J. Biomed. Soft Comput. Human Sci. 16 (2), 105–114.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowl. Discov. 2 (3), 283–304. http://dx.doi.org/10.1023/A:1009769707641.

Huang, M.K.N., 1999. A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Syst. 7 (4), 446–452.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2 (1), 193–218. http://dx.doi.org/10.1007/BF01908075.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323. http://dx.doi.org/10.1145/331499.331504.

Khalilia, M.A., Bezdek, J., Popescu, M., Keller, J.M., 2014. Improvements to the relational fuzzy c-means clustering algorithm. Pattern Recognit. 47 (12), 3920–3930. http://dx.doi.org/10.1016/j.patcog.2014.06.021.

Kim, D.-W., Lee, K.H., Lee, D., 2004. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognit. Lett. 25 (11), 1263–1271. http://dx.doi.org/10.1016/j.patrec.2004.04.004.

Leski, J.M., 2004. Fuzzy c-ordered-means clustering. Fuzzy Sets and Syst. . http://dx.doi.org/10.1016/j.fss.2014.12.007

Liu, W., Jiang, L., 2010. A clustering algorithm FCM-ACO for supplier base management. In: Cao, L., Feng, Y., Zhong, J. (Eds.), Advanced Data Mining and Applications SE-10, 6440. Springer, Berlin Heidelberg, pp. 106–113. http://dx.doi.org/10.1007/978-3-642-17316-5_10.

MacQueen J.B., 1967. Kmeans some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. ⟨http://doi.org/citeulike-article-id:6083430⟩.

McCraty, R., 2003. Enhancing Emotional, Social, and Academic Learning With Heart Rhythm Coherence Feedback. Research Center, Institute of HeartMath, Boulder Creek, CA.

Ng, M.K., Li, M.J., Huang, J.Z., He, Z., 2007. On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Trans. Pattern Anal. Machine Intell. 29 (3), 503–507. http://dx.doi.org/10.1109/TPAMI.2007.53.

Parmar, D., Wu, T., Blackhurst, J., 2007. MMR: an algorithm for clustering categorical data using Rough Set Theory. Data Knowl. Eng. 63 (3), 879–893. http://dx.doi.org/10.1016/j.datak.2007.05.005.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inf. Sci. 11 (5), 341–356. http://dx.doi.org/10.1007/BF01001956.

Pawlak, Z., 1992. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Norwell, MA, USA, Retrieved from⟨http://dl.acm.org/citation.cfm?id=531580#⟩.

Pawlak, Z., 1999. Rough classification. Int. J. Human-Comput. Stud. 51 (2), 369–383. http://dx.doi.org/10.1006/ijhc.1983.0315.

Pawlak, Z., Skowron, A., 2007. Rudiments of rough sets. Inf. Sci. 177 (1), 3–27. http://dx.doi.org/10.1016/j.ins.2006.06.003.

San, O.M., Van-Nam, H., Nakamori, Y., 2004. An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. … 14 (2), 241–247, Retrieved from⟨http://zbc.uz.zgora.pl/Content/2572/12san.pdf⟩.

Scott, A.J., Symons, M.J., 1971. Clustering methods based on likelihood ratio criteria. Biometrics 27 (2), 387–397. http://dx.doi.org/10.2307/2529003.

Symons, M.J., 1981. Clustering criteria and multivariate normal mixtures. Biometrics 37 (1), 35–43. http://dx.doi.org/10.2307/2530520.

Umayahara, K., Miyamoto, S., Nakamori, Y., 2005. Formulations of fuzzy clustering for categorical data kazutaka umayahara. Inf. Control 1 (1), 83–94.

Wei, M.W.M., Xuedong, X.H.X., Zhibo, C.Z.C., Haiyan, Z.H.Z., Chunling, W.C.W., 2009. Multi-agent reinforcement learning based on bidding. In: 1st International Conference on Information Science and Engineering (ICISE), vol. 20(3). doi: 10.1109/ICISE.2009.763.

Wu, K.L., Yang, M.S., 2002. Alternative c-means clustering algorithms. Pattern Recognit. 35 (10). http://dx.doi.org/10.1016/S0031-3203(01)00197-2.

Yang, M.S., Chiang, Y.H., Chen, C.C., Lai, C.Y., 2008. A fuzzy k-partitions model for categorical data and its comparison to the GoM model. Fuzzy Sets Syst. 159 (4), 390–405. http://dx.doi.org/10.1016/j.fss.2007.08.012.

Yang, M.-S., 1993. A survey of fuzzy clustering. Math. Comput. Model. 18 (11), 1–16. http://dx.doi.org/10.1016/0895-7177(93)90202-A.

Yanto, I.T.R., Vitasari, P., Herawan, T., Deris, M.M., 2012. Applying variable precision rough set model for clustering student suffering study's anxiety. Exp. Syst. Appl. 39 (1), 452–459. http://dx.doi.org/10.1016/j.eswa.2011.07.036.