Available online at www.sciencedirect.com

SciVerse ScienceDirect

Fuzzy Sets and Systems ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

FUZZY sets and systems

www.elsevier.com/locate/fss

A novel fuzzy clustering algorithm with between-cluster information for categorical data

Liang Bai[a,b], Jiye Liang[a,*], Chuangyin Dang[b], Fuyuan Cao[a]

[a] *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China*

[b] *Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong*

Received 8 October 2010; received in revised form 13 March 2012; accepted 21 June 2012

Abstract

In this paper, we present a new fuzzy clustering algorithm for categorical data. In the algorithm, the objective function of the fuzzy *k*-modes algorithm is modified by adding the between-cluster information so that we can simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation. For obtaining the local optimal solutions of the modified objective function, the corresponding update formulas of the membership matrix and the cluster prototypes are strictly derived. The convergence of the proposed algorithm under the optimization framework is proved. On several real data sets from UCI, the performance of the proposed algorithm is studied. The experimental results illustrate that the algorithm is effective and suitable for categorical data sets.

Keywords: Fuzzy clustering; The fuzzy *k*-modes algorithm; Optimization objective function; Categorical data

1. Introduction

Cluster analysis is a branch in statistical multivariate analysis and unsupervised machine learning which has extensive applications in various domains, including financial fraud, medical diagnosis, image processing, information retrieval and bioinformatics. The goal of clustering is to group a set of objects into clusters so that the objects in the same cluster have high similarity but are very dissimilar with objects in other clusters. Therefore, data clustering can help us to gain insight into the distribution of data. Various types of clustering algorithms have been developed in the literature (e.g., [25] and references therein). Recently, increasing attention has been paid to clustering categorical data, since this task is of great practical relevance in several fields ranging from statistics to psychology [1–3,16,31].

There are a number of challenges in clustering categorical data (which were introduced in [10,12]). First, the lack of an inherent order on the domains of the individual attributes prevents the definition of a notion of similarity, which measures resemblance between categorical data objects. Furthermore, for numerical data, the prototype of a cluster often consists of the mean of the objects in each attribute domain of the cluster, which is used to represent the cluster.

[*] Corresponding author. Tel.: +86 351 7010066.
E-mail addresses: sxbailiang@126.com (L. Bai), ljy@sxu.edu.cn (J. Liang), mecdang@cityu.edu.hk (C. Dang), cfy@sxu.edu.cn (F. Cao).

0165-0114/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fss.2012.06.005

However, it is infeasible to compute the mean for categorical values. This implies that the techniques used in clustering numerical data are not directly applicable to categorical data. Therefore, it is widely recognized that designing clustering techniques to tackle categorical data is very important for many applications.

Several algorithms for categorical data have been reported [4–6,8,13,15,18,21,22]. Among them, the *k*-modes clustering algorithm [21,22] is one of the most efficient clustering methods, which was proposed by Huang [21] in 1997. This algorithm is an extension of the *k*-means clustering algorithm [29] by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method for updating modes in the clustering process to minimize the clustering cost function. These treatments have removed the numeric-only limitation of the *k*-means algorithm, which enable the *k*-means clustering process to effectively cluster large categorical data sets from real world databases. Furthermore, a fuzzy version of the *k*-modes clustering algorithm has been reported in [23], where each pattern is allowed to have memberships in all clusters rather than just a distinct membership to a single cluster. The membership matrix provides more information to help the users to decide the core and boundary objects of clusters. Such information is extremely useful in applications [17,32,36], such as data mining in which the uncertain boundary objects are sometimes more interesting than objects which can be clustered with certainty. Lee et al. [28] introduced a generalization of the *k*-modes type clustering algorithm with fuzzy *p*-mode prototypes. A fuzzy *p*-mode cluster prototype at a categorical attribute is expressed as a list of *p* categorical values that have larger frequencies than others in the cluster.

The fuzzy *k*-modes clustering algorithm begins with an initial set of cluster prototypes and uses the alternating minimization method to solve a non-convex optimization problem in finding cluster solutions [25]. However, in the clustering process, the update formulas of the membership matrix and cluster prototypes are based on the within-cluster information only, i.e., the within-cluster compactness. The between-cluster information, i.e., the between-cluster separation, is not considered, which often results in the clustering results with weak between-cluster separation. For example, while computing the membership of an object to a cluster, we only consider the distance between the object and the cluster prototype and overlook the overlapping degrees between the cluster and other clusters. If the cluster has weak separation with other clusters, we should reduce the memberships of the objects in the boundary area between the cluster and other clusters to it. When selecting a categorical value from an attribute domain to represent a given cluster, we only consider the frequency of the categorical value within the cluster. However, the representability of the attribute value in this cluster is likely to be overestimated because other clusters also contain this value with high frequency. A detailed analysis on the importance of the between-cluster information will be provided in Section 2.

On the basis of the above idea, we will integrate the within-cluster and between-cluster information to update the membership matrix and cluster prototypes, which can effectively produce the clustering results with high within-cluster similarity and low between-cluster similarity. In this paper, the major contributions are as follows:

- Both the within-cluster and between-cluster information is employed to develop a new optimization objective function, which is used to derive a novel fuzzy clustering algorithm.
- The updating formulas of the membership matrix and cluster prototypes are derived, and the convergence of the proposed algorithm under the optimization framework is proved.
- The performance of the proposed algorithm is investigated by using several real data sets from UCI.

The rest of this paper is organized as follows. A detailed review of the fuzzy *k*-modes algorithm is presented in Section 2. In Section 3, a new objective function based on the between-cluster information is proposed to evaluate the between-cluster separation. In Section 4, the new fuzzy *k*-modes algorithm is proposed and analyzed. Section 5 illustrates the performance of the proposed algorithm. Finally, a concluding remark is given in Section 6.

2. The fuzzy *k*-modes clustering algorithm

In [22], Huang et al. provided the notations of categorical data which are introduced as follows: Let $U=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ be a set of n objects, $A=\{a_1,a_2,\ldots,a_m\}$ be a set of m attributes and D_{a_j} be the domain of attribute a_j for $1\le j\le m$. Here, we only consider two general types of the attribute domains, numerical and categorical, and assume that other types used in database systems can be mapped to one of these two types. A numerical domain consists of real numbers. A domain D_{a_j} is defined as categorical if it is finite and unordered, i.e., $D_{a_j}=\{a_j^{(1)},a_j^{(2)},\ldots,a_j^{(n_j)}\}$ where n_j is the number of categories of attribute a_j for $1\le j\le m$. For any $1\le p\le q\le n_j$, either

$a_j^{(p)} = a_j^{(q)}$ or $a_j^{(p)} \neq a_j^{(q)}$. For $1 \leq i \leq n$, object $\mathbf{x}_i \in U$ can be represented as a vector $[x_{i1}, x_{i2}, \ldots, x_{im}]$, where $x_{ij} \in D_{a_j}$ is the value of object $\mathbf{x}_i$ in attribute a_j for $1 \leq j \leq m$. If each attribute in A is categorical, U is called a categorical data set.

In the fuzzy k-modes algorithm, the objective of clustering the objects in U into k clusters is to find W and Z that minimize [22]

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} d(\mathbf{z}_l, \mathbf{x}_i) \tag{1}$$

subject to

$$\begin{cases} w_{li} \in [0, 1], & 1 \leq l \leq k,\ 1 \leq i \leq n, \\ \sum\limits_{l=1}^{k} w_{li} = 1, & 1 \leq i \leq n, \\ 0 < \sum\limits_{i=1}^{n} w_{li} < n, & 1 \leq l \leq k, \end{cases} \tag{2}$$

where n is the number of objects in U; $k(\leq n)$ is a known number of clusters; $\alpha \in (1, +\infty)$ is the fuzzy index; $W = [w_{li}]$ is a $k \times n$ real matrix, w_{li} is the membership degree of $\mathbf{x}_i$ to the lth cluster; $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\} \subseteq R$, where $R = D_{a_1} \times D_{a_2} \times \cdots \times D_{a_m}$ and $\mathbf{z}_l = [z_{l1}, z_{l2}, \ldots, z_{lm}]$ is the lth cluster prototype with categorical attributes $a_1, a_2, \ldots, a_m$; $d(\mathbf{z}_l, \mathbf{x}_i)$ is the simple matching dissimilarity measure between the object $\mathbf{x}_i$ and the prototype $\mathbf{z}_l$ of the lth cluster which is defined as

$$d(\mathbf{z}_l, \mathbf{x}_i) = \sum_{j=1}^{m} \delta(z_{lj}, x_{ij}), \tag{3}$$

where

$$\delta(z_{lj}, x_{ij}) = \begin{cases} 1, & z_{lj} \neq x_{ij}, \\ 0, & z_{lj} = x_{ij}. \end{cases} \tag{4}$$

Minimization of F with the constraints in (2) forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method towards optimization of F is to use partial optimization for Z and W. In this method, we first fix Z and find necessary conditions on W to minimize F. Then, we fix W and minimize F with respect to Z. The above optimization problem can be solved by iteratively solving the following two minimization problems:

Problem P_1**.** Fix $Z = \hat{Z}$, solve the reduced problem $F(W, \hat{Z})$ with the constraints in (2);

Problem P_2**.** Fix $W = \hat{W}$, solve the reduced problem $F(\hat{W}, Z)$ with the constraints in (2).

Given $Z = \hat{Z}$, if we compute W as follows [23]:

$$\hat{w}_{li} = \begin{cases} 1 & \text{if } d(\hat{\mathbf{z}}_l, \mathbf{x}_i) = 0, \\ 0 & \text{if } d(\hat{\mathbf{z}}_h, \mathbf{x}_i) = 0,\ h \neq l, \\ 1 \Big/ \sum_{h=1}^{k} \left[\dfrac{d(\hat{\mathbf{z}}_l, \mathbf{x}_i)}{d(\hat{\mathbf{z}}_h, \mathbf{x}_i)} \right]^{1/(\alpha-1)} & \text{if } d(\hat{\mathbf{z}}_h, \mathbf{x}_i) \neq 0,\ 1 \leq h \leq k, \end{cases} \tag{5}$$

for $1 \leq i \leq n$, $1 \leq l \leq k$, problem P_1 is solved.

Given $W = \hat{W}$, if we compute Z as follows [23]:

$$\hat{z}_{lj} = a_j^{(r)} \in D_{a_j}, \tag{6}$$

where

$$\frac{\sum_{x_{ij} = a_j^{(r)}, \mathbf{x}_i \in U} w_{li}^{\alpha}}{\sum_{\mathbf{x}_i \in U} w_{li}^{\alpha}} = \max_{q=1}^{n_j} \frac{\sum_{x_{ij} = a_j^{(q)}, \mathbf{x}_i \in U} w_{li}^{\alpha}}{\sum_{\mathbf{x}_i \in U} w_{li}^{\alpha}}, \tag{7}$$

for $1 \leq j \leq m$, $1 \leq l \leq k$, problem P_2 is solved.

In (7), the term $f_{ljr} = \sum_{x_{ij}=a_j^{(r)}, \mathbf{x}_i \in U} w_{li}^{\alpha} / \sum_{\mathbf{x}_i \in U} w_{li}^{\alpha}$ can be seen as the frequency of the categorical value $a_j^{(r)}$ in the lth fuzzy cluster. Here, a fuzzy cluster is represented by "mode", which is composed of the attribute value that occurs most frequently in each attribute domain of the cluster.

This process is formalized in the fuzzy k-modes algorithms as follows [22]:

Step 1. Choose an initial point set $Z^{(1)} \subseteq R$. Determine $W^{(1)}$ such that $F(W, Z^{(1)})$ is minimized. Set $t = 1$.
Step 2. Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized. If $F(W^{(t)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t)})$, then stop; otherwise goto Step 3.
Step 3. Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized. If $F(W^{(t+1)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t+1)})$, then stop; otherwise set $t = t + 1$ and goto Step 2.

The fuzzy k-modes clustering algorithm performs iteratively the partition step and the cluster prototypes generation step until convergence. The computational complexity of the algorithm is $O(nmkt)$ where t is the number of iterations in the clustering process.

It is noted that the fuzzy k-modes clustering algorithm faces the local minimum problem. That is, the clustering results guarantee local minimum solutions only. Its performance heavily depends on the initial cluster prototypes. Furthermore, according to (5) and (7), we remark that the update formulas of W and Z are only based on the within-cluster information. However, good cluster criteria should have high within-cluster similarity and low between-cluster similarity. The fuzzy k-modes algorithm ignores the between-cluster information, which often results in weak separation between clusters.

Let us demonstrate the importance of the between-cluster information from the following two respects. The one is to compute W when Z is fixed. The other is to compute Z when W is fixed.

(1) According to Fig. 1, we see that $d(\mathbf{x}_i, \mathbf{z}_1)$ is equal to $d(\mathbf{x}_i, \mathbf{z}_2)$. If the dissimilarity between the object and the cluster prototypes is only taken into account to compute W, $\mathbf{z}_1$ has no more representability to $\mathbf{x}_i$ than $\mathbf{z}_2$. However, the separation between $\mathbf{z}_2$ and other cluster prototypes is weak, compared to that between $\mathbf{z}_1$ and other cluster prototypes. In order to obtain a clustering result with low between-cluster similarity, we should take advantage of the between-cluster information to enhance the representability of $\mathbf{z}_1$ and reduce that of $\mathbf{z}_2$, which makes the objects in the boundary area between Cluster 1 and Cluster 2 more prone to belong to $\mathbf{z}_1$ than $\mathbf{z}_2$.
(2) When giving W to compute Z, the representability of each categorical value in a cluster is evaluated only based on its frequency in the cluster. This will lead to high importance when the value occurs frequently in this cluster. However, the representability of the categorical value in this cluster is likely to be overestimated because other clusters also contain this value with high frequency. For example, Fig. 2 shows an attribute distribution in the three clusters. The categorical value C is the most frequent value in Cluster 1. However, the categorical value C also occurs frequently in other clusters. In contrast, although the categorical value D is less frequent than the categorical value C in Cluster 1, the categorical value D mostly occurs in Cluster 1. Therefore, the categorical value D should have more representability in Cluster 1 than the categorical value C. This means that when we evaluate the importance of a categorical value in a cluster, we should not only consider the within-cluster information, i.e., the frequency in the cluster, but also consider the between-cluster information, i.e., its distribution between clusters.

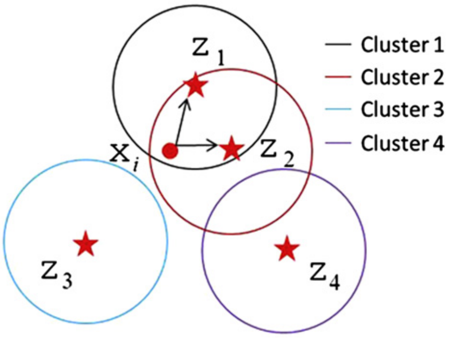

Fig. 1. The importance of the between-cluster information in computing W.

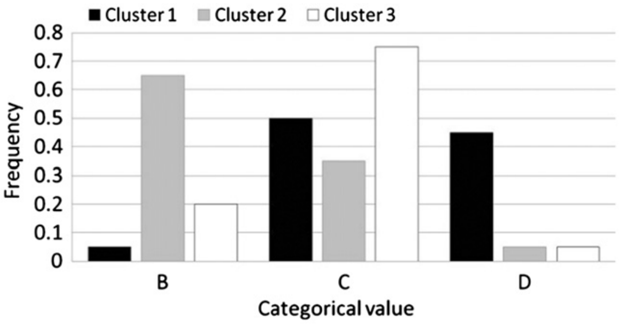

Fig. 2. The importance of the between-cluster information in computing Z.

According to the above analysis, we see that adding the between-cluster information to the iterative process can help us to obtain better W and Z. Therefore, in the next sections, we will first give the definition of the between-cluster information. Furthermore, a novel fuzzy clustering algorithm for categorical data will be proposed, where the within-cluster and between-cluster information is simultaneously considered.

3. The between-cluster information

We will present an objective function to evaluate the between-cluster separation which is defined as follows:

$$B(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha} S(\mathbf{z}_l) \tag{8}$$

where $S(\mathbf{z}_l)$ denotes the similarity between the lth cluster represented by $\mathbf{z}_l$ and other clusters, $\sum_{i=1}^{n} w_{li}^{\alpha}$ is a weight of $S(\mathbf{z}_l)$, which reflects the number of objects in the lth fuzzy cluster.

The similarity between the lth cluster and other clusters is generally measured by using the mean of the similarity between $\mathbf{z}_l$ and other cluster prototypes as

$$S(\mathbf{z}_l)=\frac{1}{k-1}\sum_{h=1,h\neq l}^{k} s(\mathbf{z}_l,\mathbf{z}_h). \tag{9}$$

Here, $s(\mathbf{z}_l,\mathbf{z}_h)$ is a similarity measure between $\mathbf{z}_l$ and $\mathbf{z}_h$ which is defined as

$$s(\mathbf{z}_l,\mathbf{z}_h)=\sum_{j=1}^{m}\phi(z_{lj},z_{hj}),$$

where

$$\phi(z_{lj},z_{hj})=\begin{cases}1, & z_{lj}=z_{hj},\\ 0, & z_{lj}\neq z_{hj}.\end{cases}$$

Similar to solving (1), the function B with the constraints in (2) is minimized by iteratively solving the following two minimization problems:

Problem P_3**.** Fix $Z=\hat{Z}$, obtain W to minimize $B(W,\hat{Z})$;

Problem P_4**.** Fix $W=\hat{W}$, obtain Z to minimize $B(\hat{W},Z)$.

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005

Given $Z = \hat{Z}$ fixed, we can obtain the minimum value of $B(W, \hat{Z})$ by the Lagrangian multiplier technique:

$$\tilde{\Phi}(W, \lambda) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} S(\hat{z}_l) + \sum_{i=1}^{n} \lambda_i \left(\sum_{l=1}^{k} w_{li} - 1 \right), \tag{10}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n]$ is the vector containing the Lagrangian multipliers. If $(\hat{W}, \hat{\lambda})$ is a minimizer of $\tilde{\Phi}(W, \lambda)$, the gradients in both sets of variables must vanish. Thus,

$$\frac{\partial \tilde{\Phi}(W, \lambda)}{\partial w_{li}} = \alpha w_{li}^{(\alpha-1)} S(\hat{z}_l) + \lambda_i = 0, \quad 1 \le l \le k, \ 1 \le i \le n, \tag{11}$$

and

$$\frac{\partial \tilde{\Phi}(W, \lambda)}{\partial \lambda_i} = \sum_{l=1}^{k} w_{li} - 1 = 0, \quad 1 \le i \le n. \tag{12}$$

From (11) and (12), we obtain

$$\hat{w}_{li} = \frac{1}{\sum_{h=1}^{k} \left[\frac{S(\hat{z}_l)}{S(\hat{z}_h)} \right]^{1/(\alpha-1)}} \tag{13}$$

for $1 \le l \le k$ and $1 \le i \le n$. This shows that (13) is the necessary condition for the optimization problem P_3 to reach its minimum when $\hat{Z}$ is fixed.

Given $W = \hat{W}$ fixed, let

$$\kappa_{l,j} = \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} \frac{1}{k-1} \sum_{h=1, h \ne l}^{k} \phi(z_{lj}, z_{hj}) \tag{14}$$

for $1 \le l \le k$ and $1 \le j \le m$. Then,

$$\begin{aligned} B(\hat{W}, Z) &= \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} \frac{1}{k-1} \sum_{h=1, h \ne k}^{k} s(\mathbf{z}_l, \mathbf{z}_h) \\ &= \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} \frac{1}{k-1} \sum_{h=1, h \ne k}^{k} \sum_{j=1}^{m} \phi(z_{lj}, z_{hj}) \\ &= \sum_{j=1}^{m} \sum_{l=1}^{k} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} \frac{1}{k-1} \sum_{h=1, h \ne k}^{k} \phi(z_{lj}, z_{hj}) \\ &= \sum_{j=1}^{m} \sum_{l=1}^{k} \kappa_{l,j}. \end{aligned}$$

According to (14), we see that if one wants to compute z_{lj} to minimize $\kappa_{l,j}$, z_{hj} must be given for $1 \le h \ne l \le k$. However, we do not know z_{hj}.

To compute z_{lj} independent of $z_{hj} (1 \le h \ne l \le k)$, we consider to employ the mean of the similarity between $\mathbf{z}_l$ and all the objects in the data set to evaluate the separation between clusters, instead of the mean of the similarity between $\mathbf{z}_l$ and other cluster prototypes. That is,

$$S(\mathbf{z}_l) = \frac{1}{n} \sum_{i=1}^{n} s(\mathbf{z}_l, \mathbf{x}_i). \tag{15}$$

Next, we will analyze whether $S(\mathbf{z}_l)$ can be used to measure the separation between clusters. Given a set U of objects, if there is a data point $\mathbf{v} \in R$ which can minimize the mean of the similarity between it and all the objects in

U, i.e., $\min_{\mathbf{v}}(1/|U|)\sum_{\mathbf{x}_i\in U} s(\mathbf{v},\mathbf{x}_i)$, the data point will be used as the representative point of U to reflect the global features of U. Therefore, the larger $S(\mathbf{z}_l)$ is, the closer $\mathbf{z}_l$ is to the representative point of U, and the more global features of U $\mathbf{z}_l$ reflects. This means that if $S(\mathbf{z}_l)$ is large, $\mathbf{z}_l$ reflects not only some features of the lth cluster but also some features of other clusters. In this case, z_l may be a boundary point among clusters, which has weak representability in the lth fuzzy cluster. If it is selected as a representative point of the lth fuzzy cluster, the separation between the lth fuzzy cluster and other clusters will be weak.

On the basis of the above idea, the between-cluster separation is evaluated by

$$B(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n} s(\mathbf{z}_l,\mathbf{x}_p).$$

Given $W=\hat{W}$ fixed, let

$$\kappa'_{l,j}=\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n}\phi(z_{lj},x_{pj}) \tag{16}$$

for $1\le l\le k$ and $1\le j\le m$. Then,

$$\begin{aligned}B(\hat{W},Z)&=\sum_{l=1}^{k}\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n} s(\mathbf{z}_l,\mathbf{x}_p)\\&=\sum_{l=1}^{k}\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n}\sum_{j=1}^{m}\phi(z_{lj},x_{pj})\\&=\sum_{j=1}^{m}\sum_{l=1}^{k}\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n}\phi(z_{lj},x_{pj})\\&=\sum_{j=1}^{m}\sum_{l=1}^{k}\kappa'_{l,j}.\end{aligned}$$

Since each $\kappa'_{l,j}$ is nonnegative and independent of each other. Thus, minimizing $B(\hat{W},Z)$ is equivalent to minimizing $\kappa'_{l,j}$. Note that

$$\begin{aligned}\kappa'_{l,j}&=\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1}^{n}\phi(z_{lj},x_{pj})\\&=\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\left[\sum_{p=1,x_{pj}=z_{lj}}^{n}\phi(z_{lj},x_{pj})+\sum_{p=1,x_{pj}\neq z_{lj}}^{n}\phi(z_{lj},x_{pj})\right]\\&=\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}\sum_{p=1,x_{pj}=z_{lj}}^{n}\phi(z_{lj},x_{pj})\\&=\sum_{i=1}^{n}\hat{w}_{li}^{\alpha}\frac{1}{n}|\{\mathbf{x}_i|x_{ij}=z_{lj},\mathbf{x}_i\in U\}|.\end{aligned}$$

It is clear that $\kappa'_{l,j}$ is minimized iff $z_{lj}=a_j^{(r)}$ which satisfies

$$|\{\mathbf{x}_i|x_{ij}=a_j^{(r)},\mathbf{x}_i\in U\}|=\min_{q=1}^{n_j}|\{\mathbf{x}_i|x_{ij}=a_j^{(q)},\mathbf{x}_i\in U\}| \tag{17}$$

for $1\le j\le m$ and $1\le l\le k$.

4. A novel fuzzy *k*-modes algorithm

In this section, we will present a new fuzzy *k*-modes algorithm for categorical data. In the new algorithm, we modify the objective function (1) by adding the between-cluster information to it so that one can simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation.

The new objective function is written as follows:

$$F_n(W, Z, \gamma) = F(W, Z) + \gamma B(W, Z)$$
$$= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha} d(\mathbf{z}_l, \mathbf{x}_i) + \gamma \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha} \frac{1}{n}\sum_{p=1}^{n} s(\mathbf{z}_l, \mathbf{x}_p) \tag{18}$$

subject to the same conditions as in (2), where the parameter γ is used to maintain a balance between the effect of the within-cluster information and that of the between-cluster information on the minimization process of (18). It has the following features in control of the clustering process:

- When $\gamma > 0$, the between-cluster similarity term $B(W, Z)$ will play an important role in the minimization of (18). The clustering process will attempt to assign each object to a cluster farther from the representative point of U to make the between-cluster similarity term smaller. When the locations of objects are fixed, in order to minimize the term, the clustering process will move the cluster prototypes to some locations which is farther from the representative point of U. However, the value of γ should not be too large. The reason is that when γ is very large so that the between-cluster similarity term dominates the clustering process, the cluster prototypes are moved to the locations of outliers in U.
- When $\gamma = 0$, the between-cluster similarity term will not play any role in the clustering process. The new objective function (18) will become the original objective function (1). The clustering process turns to minimize the within-cluster dispersion.
- When $\gamma < 0$, the clustering process will try to move the cluster prototypes to the location of the representative point of U. This is contradictory to the original idea of clustering. Therefore, γ cannot be smaller than zero.

The above properties tell us that an appropriate γ can enhance the performance of the fuzzy *k*-modes algorithm in clustering categorical data. However, the appropriate setting of γ depends on the domain knowledge of the data sets, it is difficult to directly choose a suitable value. Therefore, in the new clustering algorithm, we will not select a fixed γ value but a sequence Γ which includes several γ values. In clustering process, a larger γ value will be first used to obtain a clustering result (W, Z). Furthermore, we will gradually reduce the γ value and weaken the effect of the between-cluster information in clustering the given data set until the γ value is equal to 0 which makes minimizing the new objective function (18) is equivalent to minimizing the original objective function (1). This means that in the proposed clustering algorithm, instead of directly minimizing the objective function (1) with the constraints in (2), we consider a scheme of obtaining a solution of the problem at the limit of $\gamma \downarrow 0$ of

$$\min F_n(W, Z, \gamma) \text{ subject to (2).}$$

The basic description of the scheme is as follows:

Step 1. Let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_o\}$ be a sequence such that $\gamma_1 > \gamma_2 > \cdots > \gamma_o = 0$. Choose an initial point set $Z_1 \subseteq R$ and set $e = 1$.

Step 2. Use Z_e as an initial set of cluster prototypes to compute $(\hat{W}_e, \hat{Z}_e)$ which is a local optimal solution of

$$\min F_n(W, Z, \gamma_e) \text{ subject to (2).}$$

Step 3. Set $Z_{e+1} = \hat{Z}_e$. If $e \geq o$, then output $(\hat{W}_e, \hat{Z}_e)$ and stop; otherwise set $e = e + 1$ and goto Step 2.

According to the above description, we know that since

$$F_n(W, Z_e, \gamma_e) \geq F_n(\hat{W}_e, \hat{Z}_e, \gamma_e) \geq F_n(W, \hat{Z}_e, \gamma_{e+1}) \geq F_n(\hat{W}_{e+1}, \hat{Z}_{e+1}, \gamma_{e+1}),$$

the sequence $F_n(\cdot, \cdot, \cdot)$ produced by the above procedure is decreasing. This indicates that we can search a good clustering result by gradually updating W, Z and γ. When the initial set Z_e of cluster prototypes and γ_e are given, for $1 \leq e \leq o$, a key issue is how to derive rigorously the updating formulas of W and Z and guarantee that a local minimal solution of $F_n(W, Z, \gamma_e)$ can be obtained in a finite number of iterations. The matrices W and Z are calculated according to the following two theorems:

Theorem 1. *Let $\hat{Z}$ and γ_e be fixed and consider the problem*:

$$\min_{W} F_n(W, \hat{Z}, \gamma_e) \text{ subject to } (2).$$

The minimizer $\hat{W}$ is given by

$$\hat{w}_{li} = \begin{cases} 1, & d(\hat{\mathbf{z}}_l, \mathbf{x}_i) + \gamma_e \dfrac{1}{n} \displaystyle\sum_{p=1}^{n} s(\hat{\mathbf{z}}_l, \mathbf{x}_p) = 0, \\ 0, & d(\hat{\mathbf{z}}_h, \mathbf{x}_i) + \gamma_e \dfrac{1}{n} \displaystyle\sum_{p=1}^{n} s(\hat{\hat{\mathbf{z}}}_h, \mathbf{x}_p) = 0, \;\; h \neq l, \\ \dfrac{1}{\sum_{h=1}^{k} \left[\dfrac{d(\hat{\mathbf{z}}_l, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_l, \mathbf{x}_p)}{d(\hat{\mathbf{z}}_h, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_h, \mathbf{x}_p)} \right]^{1/(\alpha-1)}} & \textit{otherwise}. \end{cases}$$

Proof. Given that $\hat{Z}$ fixed, the Lagrangian multiplier technique is used to obtain the following unconstrained minimization problem:

$$\tilde{P}(W, \lambda) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} \left[d(\hat{\mathbf{z}}_l, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_l, \mathbf{x}_p) \right] + \sum_{i=1}^{n} \lambda_i \left(\sum_{l=1}^{k} w_{li} - 1 \right), \tag{19}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n]$ is the vector containing the Lagrangian multipliers. If $(\hat{W}, \hat{\lambda})$ is a minimizer of $\tilde{P}(W, \lambda)$, the gradients in both sets of variables must vanish. Thus,

$$\frac{\partial \tilde{P}(W, \lambda)}{\partial w_{li}} = \alpha w_{li}^{(\alpha-1)} \left[d(\hat{\mathbf{z}}_l, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_l, \mathbf{x}_p) \right] + \lambda_i = 0, \quad 1 \leq l \leq k, \;\; 1 \leq i \leq n, \tag{20}$$

and

$$\frac{\partial \tilde{P}(W, \lambda)}{\partial \lambda_i} = \sum_{l=1}^{k} w_{li} - 1 = 0, \quad 1 \leq i \leq n. \tag{21}$$

From (20) and (21), we obtain

$$\hat{w}_{li} = \frac{1}{\sum_{h=1}^{k} \left[\dfrac{d(\hat{\mathbf{z}}_l, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_l, \mathbf{x}_p)}{d(\hat{\mathbf{z}}_h, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\hat{\mathbf{z}}_h, \mathbf{x}_p)} \right]^{1/(\alpha-1)}} \tag{22}$$

for $1 \leq i \leq n$ and $1 \leq l \leq k$.

This shows that (22) is the necessary condition for the optimization objective function (18) to reach its minimum when $\hat{Z}$ and γ_e are fixed and $d(\hat{\mathbf{z}}_h, \mathbf{x}_i) + \gamma_e(1/n) \sum_{p=1}^{n} s(\hat{\mathbf{z}}_h, \mathbf{x}_p) \neq 0$ for $1 \leq h \leq k$. □

Theorem 2. *Let $\hat{W}$ and γ_e be fixed and consider the problem*:

$$\min_{Z} F_n(\hat{W}, Z, \gamma_e) \text{ subject to } (2).$$

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005

The minimizer $\hat{z}$ *is given by*

$$\hat{z}_{lj} = a_j^{(r)} \in D_{a_j},$$

where

$$\sum_{i=1, x_{ij}=a_j^{(r)}}^{n} \hat{w}_{li}^{\alpha} - \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(r)}, \mathbf{x}_p \in U\}|$$

$$\geq \sum_{i=1, x_{ij}=a_j^{(q)}}^{n} \hat{w}_{li}^{\alpha} - \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(q)}, \mathbf{x}_p \in U\}|, \; 1 \leq q \leq n_j,$$

for $1 \leq j \leq m$ *and* $1 \leq l \leq k$.

Proof. Given that $\hat{W}$ fixed, all the inner sums of the quantity

$$\sum_{l=1}^{k} \sum_{i=1}^{n} \hat{w}_{li} \left[d(\mathbf{z}_l, \mathbf{x}_i) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} s(\mathbf{z}_h, \mathbf{x}_p) \right]$$

$$= \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{w}_{li} \left[\delta(z_{lj}, x_{ij}) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} \phi(z_{lj}, x_{pj}) \right],$$

are nonnegative and independent. Minimizing the quantity is equivalent to minimizing each inner sum. We write the l, jth inner sum ($1 \leq l \leq k$ and $1 \leq j \leq m$) as

$$\psi_{l,j} = \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} \left[\delta(z_{lj}, x_{ij}) + \gamma_e \frac{1}{n} \sum_{p=1}^{n} \phi(z_{lj}, x_{pj}) \right].$$

When $f(z_l, a_j) = a_j^{(q)}$, we have

$$\psi_{l,j} = \sum_{i=1, x_{ij} \neq a_j^{(h)}}^{n} w_{li}^{\alpha} + \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(r)}, \mathbf{x}_p \in U\}|$$

$$= \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} - \sum_{i=1, x_{ij}=a_j^{(q)}}^{n} w_{li}^{\alpha} + \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(r)}, \mathbf{x}_p \in U\}|$$

$$= \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} - \left(\sum_{i=1, x_{ij}=a_j^{(q)}}^{n} \hat{w}_{li}^{\alpha} - \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(r)}, \mathbf{x}_p \in U\}| \right).$$

When $\hat{W}$ is given, $\sum_{i=1}^{n} \hat{w}_{li}^{\alpha}$ is fixed. It is clear that $\psi_{l,j}$ is minimized iff

$$\sum_{i=1, x_{ij}=a_j^{(q)}}^{n} \hat{w}_{li}^{\alpha} - \gamma_e \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{li}^{\alpha} |\{\mathbf{x}_p | x_{pj} = a_j^{(r)}, \mathbf{x}_p \in U\}|$$

is maximal for $1 \leq q \leq n_j$. The result follows. □

Combining Theorems 1 and 2 forms an iterative optimization method to minimize the objective function (18) in which the partition matrix W is computed according to Theorem 1 and the set Z of cluster prototypes is updated according to Theorem 2 in each iteration.

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005

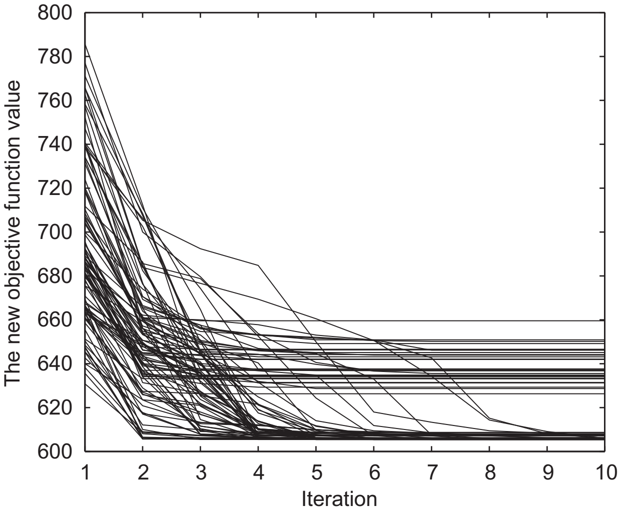

Fig. 3. The objective function values against the iterations with different initial guesses.

Theorem 3. *For any given $\gamma_e(\geq 0)$, the optimization method converges to a local minimal solution of F_n in a finite number of iterations.*

Proof. We first note that there are only a finite number ($N = \prod_{j=1}^{m} n_j$) of possible cluster prototypes (modes). We then show that each possible prototype appears at most once in the iterative process. Assume that $Z^{(t_1)} = Z^{(t_2)}$, where $t_1 \neq t_2$. We can compute the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ for $Z^{(t_1)}$ and $Z^{(t_2)}$, respectively. Therefore, we have

$$F_n(W^{(t_1)}, Z^{(t_1)}, \gamma_e) = F_n(W^{(t_1)}, Z^{(t_2)}, \gamma_e) = F_n(W^{(t_2)}, Z^{(t_2)}, \gamma_e).$$

However, the sequence $F_n(\cdot, \cdot, \gamma_e)$ generated by the iterative method is strictly decreasing. Hence, the result follows. □

In Fig. 3, we show the 100 curves, where each curve refers to the objective function F_n values against the iterations of the above method with different initial cluster prototypes on the soybean data set from UCI. It is clear from the figure that the objective function values are decreasing in each curve, which is consistent with our results in Theorems 1 and 2. We also see in Fig. 3 that the iterative process stops after a finite number of iterations, i.e., the objective function values do not decrease any more. This is exactly the result we showed in Theorem 3. Therefore, the iterative process guarantees to converge to a local minimal solution.

The proposed clustering algorithm is scalable to the number of objects, attributes or clusters. This is because the proposed algorithm only adds a new computational cost to the fuzzy *k*-modes clustering process to calculate the between-cluster separation. The runtime complexity can be analyzed as follows. We only consider the three major computational steps:

- *Computing the between-cluster information. We have*

$$\begin{aligned} S(\mathbf{z}_l) &= \frac{1}{n}\sum_{p=1}^{n} s(\mathbf{z}_l, \mathbf{x}_p) \\ &= \frac{1}{n}\sum_{j=1}^{m}\sum_{p=1}^{n} \phi(z_{lj}, x_{pj}) \\ &= \frac{1}{n}\sum_{j=1}^{m}\sum_{p=1}^{n} |\{\mathbf{x}_p | x_{pj} = z_{lj}, \mathbf{x}_p \in U\}|. \end{aligned}$$

Therefore, before implementing the proposed algorithm, we calculate and save the frequency of each categorical value of each attribute in U, which will be used to update W and Z. The step takes $O(n\sum_{j=1}^{m} n_j)$ operations.

- *Updating the membership matrix.* Given the cluster prototypes Z, compute the memberships in Theorem 1 for each object in all k clusters. Thus, the computational complexity for this step is $O(mnk)$ operations.
- *Updating the cluster prototypes.* Given the membership matrix W, updating cluster prototypes is finding the modes of the objects in the same cluster. Thus, for k clusters, the computational complexity for this step is $O(mnk)$ operations.

If one needs t_e iterations to obtain a local minimal solution of F_n for each $\gamma_e (e = 1, 2, \ldots, o)$, the total computational complexity of the proposed algorithm is $O(n\sum_{j=1}^{m} n_j + mnk\sum_{e=1}^{o} t_e)$. This shows that the computational complexity increases linearly with the number of objects, attributes or clusters.

5. Experimental analysis

The main aim of this section is to demonstrate the performance of the new fuzzy k-modes algorithm (NFKM) by a thorough experimental study on the real categorical data sets. In Section 5.1, the test environment and the data sets used are described. In Section 5.2, we introduce several validity indices which are used to evaluate the effectiveness of the clustering results. In Section 5.3, we present the comparisons of the proposed algorithms with the hard k-modes algorithm (HKM) [21], the weighting k-modes algorithm (WKM) [24] and the fuzzy k-modes algorithm (FKM) [23] on the given data sets.

5.1. Test environment and data sets

To ensure that the comparisons are in a uniform environmental condition, we give the parameters setting for these clustering algorithms as follows:

(1) We set the number of clusters is equal to the number of classes for each of the given data sets.
(2) Due to the fact that the performance of these algorithms depends on initial cluster prototypes, we randomly select 100 initial sets of cluster prototypes and carry out 100 runs of each algorithm on each data set, respectively. In each run, the same initial cluster prototypes are used in different algorithms.
(3) For the fuzzy clustering algorithms, it is necessary to set the fuzzy index α appropriately. Theoretical and empirical results on the study of setting the fuzzy index have been obtained for numerical data [7,9,20,34,35]. Among them, Pal and Bezdek [7] gave heuristic guidelines regarding the best choice for α, suggesting that it is probably in the interval (1.5, 2.5). However, the fuzzy clustering algorithms for categorical data have bad performance in the interval. This indicates that the interval is not appropriate for clustering categorical data. Unfortunately, the corresponding theoretical study for categorical data has not been reported. In [23], Huang and Ng suggested to specify $\alpha = 1.1$ since they tried several values of α and found that $\alpha = 1.1$ provides the least value of the objective function (1). In the following experimental analysis, we will test the performance of the proposed algorithm with $\alpha \in (1, 3)$.

Table 1
The nine UCI data sets.

Data set	The number of objects	The number of categorical attributes	The number of clusters
Lung cancer	32	56	3
Small soybean	47	35	4
Teaching	151	5	3
Heart disease	303	8	2
Dermatology	366	33	6
Credit approval	690	8	2
Breast cancer	699	9	2
Letters (E,F)	1543	16	2
Mushroom	8124	22	2

Table 2
Means of AC, PR, RE for 100 runs of four algorithms on the nine data sets.

Data set	Index	HKM	WKM	FKM	NFKM
Lung cancer	*AC*	0.5313	0.5497	0.5306	**0.6012**
	PR	0.5880	0.5965	0.5885	**0.6688**
	RE	0.5374	0.5626	0.5306	**0.5954**
Small soybean	*AC*	0.8553	0.8613	0.8336	**0.9264**
	PR	0.9020	0.8948	0.8840	**0.9426**
	RE	0.8407	0.8471	0.8176	**0.9216**
Teaching	*AC*	0.4137	0.4124	0.4166	**0.4528**
	PR	0.5005	0.5028	0.4818	**0.5122**
	RE	0.4153	0.4126	0.4344	**0.4624**
Heart disease	*AC*	0.7462	0.7472	0.7487	**0.7882**
	PR	0.7573	0.7566	0.7556	**0.7890**
	RE	0.7446	0.7455	0.7464	**0.7872**
Dermatology	*AC*	0.6869	0.6854	0.6698	**0.7423**
	PR	0.7633	0.7692	0.7286	**0.8443**
	RE	0.5750	0.5765	0.5566	**0.6100**
Credit approval	*AC*	0.7517	0.7513	0.7491	**0.7701**
	PR	0.7629	0.7513	0.7638	**0.7701**
	RE	0.7638	0.7715	0.7629	**0.7729**
Breast cancer	*AC*	0.8482	0.8530	0.8343	**0.9446**
	PR	0.8731	0.8733	0.8613	**0.9456**
	RE	0.7893	0.7968	0.7665	**0.9312**
Letters	*AC*	0.6910	0.6930	0.6790	**0.7458**
	PR	0.7016	0.7050	0.6843	**0.7512**
	RE	0.6911	0.6932	0.6790	**0.7461**
Mushroom	*AC*	0.7176	0.7106	0.7001	**0.8298**
	PR	0.7453	0.7414	0.7166	**0.8469**
	RE	0.7132	0.7056	0.6947	**0.8257**

(4) Before implementing the proposed algorithm, we need to provide a sequence $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_o\}$. We set $\gamma_1 = 1$, $\gamma_o = 0$ and $\gamma_e = \gamma_{e+1} - 0.1$, $1 \leq e < o$.

On the basis of the above parameters setting, we use the nine standard data sets (shown in Table 1) obtained from the UCI Machine Learning Repository [14] to test the performance of the proposed algorithm. These data sets are introduced as follows:

Lung cancer data. The data set was used by Hong and Young to illustrate the power of the optimal discriminant plane even in ill-posed settings. This data set has 32 instances described by 56 categorical attributes. It contains three class.

Small soybean data. The data set has 47 records, each of which is described by 35 attributes. Each record is labeled as one of the four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for Phytophthora Rot which has 17 records, all other diseases have ten records each.

Teaching data. The data set consists of evaluations of teaching performance over three regular semesters and two summer semesters of 151 teaching assistant (TA) assignments at the Statistics Department of the University of Wisconsin-Madison. The scores were divided into three roughly equal-sized categories ("low", "medium", and "high") to form the class variable.

Table 3
Standard deviations of AC, PR, RE for 100 runs of four algorithms on the nine data sets.

Data set	Index	HKM	WKM	FKM	NFKM
Lung cancer	*AC*	0.0485	0.0526	0.0487	**0.0375**
	PR	0.0744	0.0767	0.0762	**0.0477**
	RE	0.0582	0.0652	0.0558	**0.0421**
Small soybean	*AC*	0.1101	0.1072	0.1135	**0.0889**
	PR	0.0793	0.0815	0.0836	**0.0656**
	RE	0.1263	0.1265	0.1289	**0.1036**
Teaching	*AC*	0.0274	0.0277	0.0322	**0.0274**
	PR	0.0398	0.0471	0.0394	**0.0394**
	RE	0.0295	0.0300	0.0320	**0.0348**
Heart disease	*AC*	0.0819	0.0825	0.0830	**0.0266**
	PR	0.0788	0.0806	0.0792	**0.0276**
	RE	0.0869	0.0871	0.0896	**0.0125**
Dermatology	*AC*	0.0735	0.0683	0.0746	**0.0481**
	PR	0.0777	0.0731	0.0899	**0.0437**
	RE	0.0800	0.0770	0.0774	**0.0735**
Credit approval	*AC*	0.0936	0.0952	0.0919	**0.0487**
	PR	0.0850	0.0952	0.0948	**0.0524**
	RE	0.1005	0.1100	0.1067	**0.0465**
Breast cancer	*AC*	0.0735	0.0843	0.1028	**0.0157**
	PR	0.0606	0.0808	0.0814	**0.0084**
	RE	0.1058	0.1194	0.1522	**0.0254**
Letters	*AC*	0.0665	0.0678	0.0779	**0.0505**
	PR	0.0698	0.0717	0.0794	**0.0532**
	RE	0.0667	0.0680	0.0780	**0.0507**
Mushroom	*AC*	0.1237	0.1204	0.1260	**0.1113**
	PR	0.1336	0.1321	0.1361	**0.1184**
	RE	0.1239	0.1208	0.1267	**0.1113**

Heart disease data. The data set generated at the Cleveland Clinic has 303 instances with eight categorical and five numerical features. It contains two classes: normal (164 data objects) and heart patient (139 data objects). In the test, all numerical attributes are removed from the data set.

Dermatology data. The data set describes clinical features and histopathological features of erythemato-squamous diseases in dermatology. It contains 366 elements and 33 categorical attributes. It has six classes: psoriasis (112 data objects), seboreic dermatitis (61 data objects), lichen planus (72 data objects), pityriasis rosea (49 data objects), cronic dermatitis (52 data objects) and pityriasis rubra pilaris (20 data objects).

Credit approval data. The data set contains data from credit card organization, where customers are divided into two classes. It is a mixed data set with eight categorical and six numeric features. It contains 690 data objects belonging to two classes: negative (383 data objects) and positive (307 data objects). In the test, we only consider the categorical attributes on the data set.

Breast cancer data. The data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. It consists of 699 data objects and nine categorical attributes. It has two classes: Benign (458 data objects) and Malignant (241 data objects).

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005

Table 4
Means of PC for 100 runs with respect to different α values on the nine data sets.

α	Algorithm	Lung cancer	Small soybean	Teaching	Heart Disease	Dermatology	Credit approval	Breast cancer	Letters (E, F)	Mushroom
1.1	FKM	0.8106	0.9148	0.6737	0.9145	0.7035	0.8904	0.8743	0.7543	0.8744
	NFKM	**0.8296**	**0.9439**	**0.7409**	**0.9293**	**0.8181**	**0.8847**	**0.9237**	**0.8017**	**0.9206**
1.3	FKM	0.4632	0.6828	0.4626	0.7576	0.3075	0.7026	0.7366	0.5512	0.6789
	NFKM	**0.5002**	**0.6995**	**0.5349**	**0.7905**	**0.3749**	**0.7405**	**0.8222**	**0.6012**	**0.7254**
1.5	FKM	0.3781	0.5061	0.3828	0.6604	0.2000	0.6166	0.6360	0.5157	0.5726
	NFKM	**0.3863**	**0.5151**	**0.4441**	**0.7063**	**0.2172**	**0.6468**	**0.7453**	**0.5379**	**0.6346**
1.7	FKM	0.3526	0.3941	0.3602	0.6047	0.1725	0.5702	0.5744	0.5060	0.5366
	NFKM	**0.3538**	**0.4136**	**0.3801**	**0.6473**	**0.1800**	**0.5945**	**0.6447**	**0.5165**	**0.5730**
1.9	FKM	0.3425	0.3473	0.3527	0.5750	0.1681	0.5464	0.5423	0.5034	0.5212
	NFKM	**0.3446**	**0.3667**	**0.3669**	**0.6026**	**0.1681**	**0.5677**	**0.5429**	**0.5061**	**0.5490**
2.1	FKM	0.3391	0.3129	0.3496	0.5563	0.1675	0.5347	0.5362	0.5022	0.5134
	NFKM	**0.3393**	**0.3406**	**0.3587**	**0.5798**	**0.1676**	**0.5515**	**0.5372**	**0.5035**	**0.5234**
2.3	FKM	0.3368	0.2974	0.3491	0.5442	0.1672	0.5267	0.5325	0.5015	0.5080
	NFKM	**0.3371**	**0.3313**	**0.3541**	**0.5648**	**0.1673**	**0.5421**	**0.5336**	**0.5023**	**0.5164**
2.5	FKM	0.3359	0.2740	0.3488	0.5363	0.1671	0.5230	0.5301	0.5011	0.5055
	NFKM	**0.3368**	**0.3285**	**0.3517**	**0.5544**	**0.1672**	**0.5349**	**0.5313**	**0.5016**	**0.5124**
2.7	FKM	0.3349	0.2708	0.3488	0.5319	0.1670	0.5207	0.5284	0.5008	0.5039
	NFKM	**0.3505**	**0.3274**	**0.3528**	**0.5470**	**0.1678**	**0.5300**	**0.5302**	**0.5013**	**0.5098**
2.9	FKM	0.3349	0.2648	0.3463	0.5284	0.1670	0.5189	0.5272	0.5007	0.5027
	NFKM	**0.3817**	**0.3250**	**0.3508**	**0.5415**	**0.1728**	**0.5264**	**0.5289**	**0.5010**	**0.5079**

Letters data. The data set contains character image features of 26 capital letters in the English alphabet. We take data objects with similar looking alphabets, *E* and *F* alphabets from this data set. There are 1543 data objects (768 *E* and 775 *F*) described by 16 attributes which are integer valued and seen as categorical attributes in the experiments.

Mushroom data. The data set includes descriptions of hypothetical samples corresponding to 22 species of gilled mushrooms in the Agaricus and Lepiota Family. It consists of 8124 data objects and 22 categorical attributes. Each object belongs to one of two classes, edible (4208 objects) and poisonous (3916 objects).

5.2. Validity indices

To evaluate the performance of clustering algorithms in the experiments, we consider the five validity indices [26,30,33]: accuracy (AC), precision (PR), recall (RE), partition coefficient (PC) and partition entropy (PE).

The first three indices are external criteria which take advantage of the true class labels to evaluate the clustering result on each of these given data sets. If the cluster result is close to the true class distribution, then the values of these evaluation measures are high. The three validity indices are defined as follows [33]:

$$AC = \frac{\sum_{l=1}^{k} a_l}{n}, \quad PR = \frac{\sum_{l=1}^{k}\left(\dfrac{a_l}{a_l + b_l}\right)}{k}, \quad RE = \frac{\sum_{l=1}^{k}\left(\dfrac{a_l}{a_l + c_l}\right)}{k},$$

Table 5
Standard deviations of PC for 100 runs with respect to different α values on the nine data sets.

α	Algorithm	Lung cancer	Small soybean	Teaching	Heart Disease	Dermatology	Credit approval	Breast cancer	Letters (E, F)	Mushroom
1.1	FKM	0.0407	0.0525	0.0841	0.0370	0.0651	0.0539	0.0883	0.0338	0.0449
	NFKM	**0.0218**	**0.0431**	**0.0856**	**0.0083**	**0.0217**	**0.0287**	**0.0001**	**0.0212**	**0.0094**
1.3	FKM	0.0246	0.0650	0.0471	0.0424	0.0466	0.0534	0.0991	0.0164	0.0474
	NFKM	**0.0214**	**0.0577**	**0.0137**	**0.0032**	**0.0182**	**0.0152**	**0.0006**	**0.0143**	**0.0164**
1.5	FKM	0.0128	0.0473	0.0253	0.0381	0.0186	0.0408	0.0680	0.0054	0.0461
	NFKM	**0.0107**	**0.0446**	**0.0139**	**0.0079**	**0.0221**	**0.0119**	**0.0012**	**0.0050**	**0.0108**
1.7	FKM	0.0067	0.0302	0.0120	0.0310	0.0066	0.0301	0.0344	0.0017	0.0267
	NFKM	**0.0015**	**0.0274**	**0.0102**	**0.0070**	**0.0048**	**0.0160**	**0.0249**	**0.0024**	**0.0038**
1.9	FKM	0.0028	0.0193	0.0103	0.0246	0.0011	0.0230	0.0025	0.0009	0.0172
	NFKM	**0.0015**	**0.0189**	**0.0086**	**0.0063**	**0.0001**	**0.0124**	**0.0014**	**0.0016**	**0.0050**
2.1	FKM	0.0015	0.0188	0.0092	0.0201	0.0001	0.0183	0.0031	0.0007	0.0114
	NFKM	**0.0006**	**0.0181**	**0.0086**	**0.0060**	**0.0001**	**0.0100**	**0.0016**	**0.0008**	**0.0020**
2.3	FKM	0.0011	0.0134	0.0094	0.0155	0.0001	0.0146	0.0037	0.0005	0.0070
	NFKM	**0.0002**	**0.0133**	**0.0083**	**0.0058**	**0.0001**	**0.0060**	**0.0017**	**0.0003**	**0.0022**
2.5	FKM	0.0043	0.0103	0.0088	0.0132	0.0001	0.0121	0.0039	0.0004	0.0052
	NFKM	**0.0002**	**0.0093**	**0.0084**	**0.0056**	**0.0001**	**0.0050**	**0.0018**	**0.0002**	**0.0007**
2.7	FKM	0.0159	0.0084	0.0092	0.0121	0.0020	0.0101	0.0043	0.0003	0.0038
	NFKM	**0.0002**	**0.0049**	**0.0088**	**0.0055**	**0.0000**	**0.0043**	**0.0019**	**0.0002**	**0.0012**
2.9	FKM	0.0131	0.0075	0.0088	0.0112	0.0039	0.0091	0.0044	0.0002	0.0030
	NFKM	**0.0001**	**0.0036**	**0.0076**	**0.0054**	**0.0000**	**0.0038**	**0.0019**	**0.0001**	**0.0010**

where a_l is the number of objects that are correctly assigned to the lth class ($1 \leq l \leq k$), b_l is the number of objects that are incorrectly assigned to the lth class, c_l is the number of objects that should be in, but are not correctly assigned to the lth class.

The last two indices are internal criteria which measure the fuzziness of the clustering results obtained by fuzzy clustering algorithms. The lower the fuzziness of a clustering result is, the less uncertainty the clustering result has [27]. The two fuzzy cluster indices are defined as follows [26,30]:

$$PC = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li}^2, \quad PE = -\frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} \log_2 w_{li}.$$

The lower the fuzziness of a partition is, the larger the PC value (or the smaller the PE value) is.

5.3. *Performance analysis*

On each of the above data sets, the performance analysis consists of the following two parts:

Part 1. We use the three external indices, i.e., AC, PR and RE, to compare the clustering results of the proposed algorithm with those of the hard k-modes algorithm , the weighting k-modes algorithm and the fuzzy k-modes algorithm. Unlike the hard clustering algorithms, the fuzzy clustering algorithms produce a fuzzy partition matrix W. To calculate the values of these evaluation measures, we need to obtain the hard partition matrix from W as follows. The object

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005

Table 6
Means of PE for 100 runs with respect to different α values on the nine data sets.

α	Algorithm	Lung cancer	Small soybean	Teaching	Heart Disease	Dermatology	Credit approval	Breast cancer	Letters (E, F)	Mushroom
1.1	FKM	0.4813	0.2246	0.7656	0.1999	0.7940	0.2582	0.2756	0.5490	0.2921
	NFKM	**0.4340**	**0.1438**	**0.5963**	**0.1687**	**0.4840**	**0.2572**	**0.1770**	**0.4527**	**0.1946**
1.3	FKM	1.3071	0.8793	1.2987	0.5449	2.0530	0.6527	0.5677	0.9217	0.6955
	NFKM	**1.2247**	**0.8331**	**1.1184**	**0.4803**	**1.8043**	**0.5796**	**0.4012**	**0.8386**	**0.6078**
1.5	FKM	1.4904	1.3471	1.4794	0.7319	2.4504	0.8099	0.7668	0.9769	0.8851
	NFKM	**1.4718**	**1.3205**	**1.3410**	**0.6485**	**2.3825**	**0.7576**	**0.5690**	**0.9429**	**0.7816**
1.7	FKM	1.5439	1.6211	1.5276	0.8300	2.5603	0.8876	0.8745	0.9912	0.9444
	NFKM	**1.5411**	**1.5813**	**1.4856**	**0.7579**	**2.5261**	**0.8483**	**0.7595**	**0.9757**	**0.8871**
1.9	FKM	1.5653	1.7466	1.5427	0.8791	2.5787	0.9254	0.9258	0.9951	0.9684
	NFKM	**1.5607**	**1.6982**	**1.5124**	**0.8334**	**2.5790**	**0.8921**	**0.9241**	**0.9912**	**0.9257**
2.1	FKM	1.5726	1.8331	1.5488	0.9091	2.5814	0.9435	0.9350	0.9968	0.9803
	NFKM	**1.5721**	**1.7623**	**1.5292**	**0.8703**	**2.5810**	**0.9177**	**0.9325**	**0.9949**	**0.9654**
2.3	FKM	1.5774	1.8730	1.5492	0.9279	2.5825	0.9556	0.9405	0.9979	0.9882
	NFKM	**1.5768**	**1.7847**	**1.5387**	**0.8941**	**2.5821**	**0.9318**	**0.9379**	**0.9967**	**0.9760**
2.5	FKM	1.5794	1.9346	1.5494	0.9400	2.5831	0.9612	0.9440	0.9984	0.9920
	NFKM	**1.5773**	**1.7906**	**1.5435**	**0.9102**	**2.5828**	**0.9427**	**0.9413**	**0.9976**	**0.9818**
2.7	FKM	1.5807	1.9433	1.5497	0.9462	2.5833	0.9644	0.9465	0.9988	0.9943
	NFKM	**1.5449**	**1.7932**	**1.5444**	**0.9215**	**2.5804**	**0.9500**	**0.9427**	**0.9981**	**0.9856**
2.9	FKM	1.5815	1.9598	1.5596	0.9513	2.5837	0.9671	0.9482	0.9990	0.9961
	NFKM	**1.4713**	**1.7994**	**1.5552**	**0.9297**	**2.5625**	**0.9554**	**0.9447**	**0.9985**	**0.9885**

$\mathbf{x}_i$ is assigned to the lth cluster if $w_{li} = \max_{1 \le h \le k} w_{hi}$. If the maximum is not unique, then $\mathbf{x}_i$ is assigned to the cluster of first achieving the maximum. In this part, we set $\alpha = 1.1$ which was suggested in [23]. Tables 2 and 3 show the means and standard deviations of AC, PR, RE for the 100 runs of each algorithm on these given data sets.

Part 2. We employ the two fuzzy cluster indices PC and PE to measure the fuzziness of the clustering results obtained by fuzzy clustering algorithms. We apply the fuzzy k-modes algorithm and the proposed algorithm to cluster these data sets with different α values, respectively. For each α, we will compute the means and standard deviations of the fuzzy cluster indices for the 100 runs of each algorithm on each data set. Tables 4–7 show the comparison results of the two algorithms for PC and PE with different α values (the value of α is from 1.1 to 2.9 with a step length of 0.2).

Performance results. First, we see from Table 2 that the proposed algorithm has higher accuracy in clustering these given data sets than other clustering algorithms. Furthermore, Table 3 illustrates that the performances of the proposed algorithm are relatively consistent in the 100 clustering results of each given data set, compared to those of other clustering algorithms. According to Tables 4 and 6, we see that the cluster results produced by the proposed algorithms have less uncertainty than those produced by the fuzzy k-modes algorithm. Tables 5 and 7 illustrate that the proposed algorithms have relatively robust results in clustering these given data sets, compared to the fuzzy k-modes algorithm. In addition, we see from the analysis of the fuzzy index α that as the fuzzy index α value increases, the PC value decreases (or the PE value increases), which was explained in [26]. When the value of α is larger than 1.5, the fuzzy k-modes algorithm and the proposed algorithm have bad performance. Therefore, we suggest that the value of α should be on the interval (1, 1.5] when clustering categorical data sets.

Table 7
Standard deviations of PE for 100 runs with respect to different α values on the nine data sets.

α	Algorithm	Lung cancer	Small soybean	Teaching	Heart disease	Dermatology	Credit approval	Breast cancer	Letters (E, F)	Mushroom
1.1	FKM	0.0944	0.1353	0.1947	0.0717	0.1913	0.0944	0.1763	0.0682	0.0915
	NFKM	**0.0459**	**0.1125**	**0.1529**	**0.0142**	**0.0536**	**0.0488**	**0.0002**	**0.0428**	**0.0223**
1.3	FKM	0.0517	0.1582	0.1103	0.0826	0.1693	0.0954	0.1921	0.0264	0.0827
	NFKM	**0.0448**	**0.1337**	**0.0300**	**0.0067**	**0.0578**	**0.0304**	**0.0014**	**0.0242**	**0.0331**
1.5	FKM	0.0265	0.1070	0.0558	0.0680	0.0730	0.0692	0.1188	0.0082	0.0743
	NFKM	**0.0217**	**0.1034**	**0.0319**	**0.0152**	**0.0866**	**0.0218**	**0.0023**	**0.0077**	**0.0187**
1.7	FKM	0.0138	0.0733	0.0263	0.0527	0.0280	0.0489	0.0548	0.0024	0.0413
	NFKM	**0.0031**	**0.0605**	**0.0223**	**0.0125**	**0.0215**	**0.0271**	**0.0438**	**0.0036**	**0.0065**
1.9	FKM	0.0059	0.0481	0.0235	0.0408	0.0052	0.0371	0.0060	0.0013	0.0259
	NFKM	**0.0032**	**0.0472**	**0.0194**	**0.0121**	**0.0006**	**0.0208**	**0.0032**	**0.0023**	**0.0079**
2.1	FKM	0.0032	0.0487	0.0213	0.0329	0.0004	0.0297	0.0071	0.0010	0.0170
	NFKM	**0.0013**	**0.0469**	**0.0198**	**0.0115**	**0.0004**	**0.0168**	**0.0035**	**0.0012**	**0.0030**
2.3	FKM	0.0024	0.0346	0.0219	0.0255	0.0004	0.0240	0.0079	0.0007	0.0103
	NFKM	**0.0004**	**0.0353**	**0.0192**	**0.0112**	**0.0003**	**0.0097**	**0.0037**	**0.0005**	**0.0033**
2.5	FKM	0.0101	0.0245	0.0206	0.0221	0.0003	0.0202	0.0083	0.0006	0.0076
	NFKM	**0.0005**	**0.0274**	**0.0196**	**0.0110**	**0.0002**	**0.0082**	**0.0038**	**0.0003**	**0.0010**
2.7	FKM	0.0375	0.0225	0.0216	0.0205	0.0076	0.0172	0.0089	0.0004	0.0055
	NFKM	**0.0003**	**0.0131**	**0.0206**	**0.0108**	**0.0002**	**0.0072**	**0.0038**	**0.0003**	**0.0018**
2.9	FKM	0.0315	0.0200	0.0207	0.0193	0.0139	0.0156	0.0090	0.0004	0.0044
	NFKM	**0.0002**	**0.0096**	**0.0179**	**0.0107**	**0.0002**	**0.0066**	**0.0039**	**0.0002**	**0.0015**

6. Conclusions

In this paper, we have presented a novel fuzzy clustering algorithm for categorical data which is an extension of the fuzzy *k*-modes algorithm. In this algorithm, we have integrated the within-cluster and between-cluster information. Furthermore, we have rigorously derived the updating formulas of the membership matrix and the set of cluster prototypes in the clustering process and proved the convergence of the proposed algorithm under the optimization framework. The time complexity of the algorithm has been analyzed which is linear with respect to the number of data objects, attributes or clusters. We have tested the algorithm using several real data sets from UCI. Experimental results have shown that the proposed algorithm is effective in clustering categorical data sets.

Acknowledgments

The authors are very grateful to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Nos. 71031006, 70971080), GRF: CityU 112809 of Hong Kong SAR Government, the Foundation of Doctoral Program Research of Ministry of Education of China (No. 20101401110002), the National Key Basic Research and Development Program of China (973) (No. 2011CB311805), the Key Problems in Science and Technology Project of Shanxi (No. 20110321027-01).

References

[1] C.C. Aggarwal, C. Magdalena, P.S. Yu, Finding localized associations in market basket data, IEEE Trans. Knowl. Data Eng. 14 (1) (2002) 51–62.

[2] A. Baxevanis, F. Ouellette, et al., Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd ed., Wiley, NY, 2001.

[3] D. Barbara, S. Jajodia, et al., Applications of Data Mining in Computer Security, Kluwer, Dordrecht, 2002.

[4] L. Bai, J.Y. Liang, C.Y. Dang, F.Y. Cao, A novel attribute weighting algorithm for clustering high-dimensional categorical data, Pattern Recogn. 44 (12) (2011) 2843–2861.

[5] L. Bai, J.Y. Liang, C.Y. Dang, An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data, Knowl-Based Syst. 24 (6) (2011) 785–795.

[6] D. Barbara, Y. Li, J. Couto, Coolcat: an entropy-based algorithm for categorical clustering, in: Information and Knowledge Management, 2002, pp. 582–589.

[7] J.C. Bezdek, A physical interpretation of fuzzy ISODATA, IEEE Trans. Syst. Man Cybern. 6 (5) (1976) 387–390.

[8] F.Y. Cao, J.Y. Liang, L. Bai, et al., A framework for clustering categorical time-evolving data, IEEE Trans. Fuzzy Syst. 18 (5) (2010) 872–882.

[9] R.L. Cannon, J.V. Dave, J.C. Bezdek, Efficient implementation of the fuzzy c-means clustering algorithms, IEEE Trans. Pattern Anal. Mach. Int. 8 (2) (1986) 248–255.

[10] E. Cesario, G. Manco, R. Ortale, Top-down parameter-free clustering of high-dimensional categorical data, IEEE Trans. Knowl. Data Eng. 19 (12) (2007) 1607–1624.

[11] A. Chaturvedi, P. Green, J. Carroll, K-modes clustering, J. Classif. 18 (1) (2001) 35–55.

[12] H.L. Chen, K.T. Chuang, M.S. Chen, On data labeling for clustering categorical data, IEEE Trans. Knowl. Data Eng. 20 (11) (2008) 1458–1472.

[13] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Mach. Learn. 2 (2) (1987) 139–172.

[14] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.

[15] V. Ganti, J.E. Gekhre, R. Ramakrishnan, CACTUS-clustering categorical data using summaries, in: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 73–83.

[16] K.C. Gowda, E. Diday, Symbolic clustering using a new dissimilarity measure, Pattern Recogn. 24 (6) (1991) 567–578.

[17] D. Graves, W. Pedrycz, Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study, Fuzzy Set Syst. 161 (4) (2010) 522–543.

[18] S. Guha, R. Rastogi, S. Kyuseok, ROCK: a robust clustering algorithm for categorical attributes, in: Proceedings of 15th International Conference on Data Engineering, Sydney, Australia, vol. 23–26, 1999, pp. 512–521.

[19] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Cluster validity methods: Part I and II, SIGMOD Rec. 31 (2) (2002) 40–45.

[20] L.O. Hall, A.M. Bensaid, L.P. Clarke, A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain, IEEE Trans. Neural Networ. 3 (5) (1992) 672–682.

[21] Z.X. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, in: Proceedings of the SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery, 1997, pp. 1–8.

[22] Z.X. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Disc. 2 (3) (1998) 283–304.

[23] Z.X. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999) 446–452.

[24] Z.X. Huang, M.K. Ng, H. Rong, Z. Li, Automated variable weighting in k-means type clustering, IEEE Trans. Pattern Anal. Mach. Int. 27 (5) (2005) 657–668.

[25] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[26] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Syst. 3 (3) (1995) 370–379.

[27] Y.H. Qian, J.Y. Liang, C.Y. Dang, Consistency measure, inclusion degree and fuzzy measure in decision tables, Fuzzy Set Syst. 159 (18) (2008) 2353–2377.

[28] M. Lee, W. Pedrycz, The fuzzy c-means algorithm with fuzzy p-mode prototypes for clustering objects having mixed features, Fuzzy Set Syst. 160 (24) (2009) 3590–3600.

[29] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, vol. 1, 1967, pp. 281–297.

[30] W.N. Wang, Y.J. Zhang, On fuzzy cluster validity indices, Fuzzy Set Syst. 158 (19) (2007) 2095–2117.

[31] N. Wrigley, Categorical Data Analysis for Geographers and Environmental Scientists, Longman, London, 1985.

[32] K.L. Wu, J. Yu, M.S. Yang, A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality test, Pattern Recogn. Lett. 26 (5) (2005) 639–652.

[33] Y.M. Yang, An evaluation of statistical approaches to text categorization, J. Inform. Retrieval 1 (1–2) (1999) 67–88.

[34] J. Yu, M.S. Yang, Optimality test for generalized FCM and its application to parameter selection, IEEE Trans. Fuzzy Syst. 13 (1) (2005) 164–176.

[35] J. Yu, Q.S. Cheng, H.K. Huang, Analysis of the weighting exponent in the FCM, IEEE Trans. Syst. Man Cybern. B Cybern. 34 (1) (2004) 164–176.

[36] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (3) (1965) 338–353.

Please cite this article as: L. Bai, et al., A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems (2012), http://dx.doi.org/10.1016/j.fss.2012.06.005