Neurocomputing 133 (2014) 111–121

Contents lists available at ScienceDirect

Neurocomputing

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

The *k*-modes type clustering plus between-cluster information for categorical data

Liang Bai, Jiye Liang*

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

ARTICLE INFO

Article history:
Received 24 May 2013
Received in revised form
7 September 2013
Accepted 12 November 2013
Communicated by Zhi Yong Liu
Available online 11 January 2014

Keywords:
Cluster analysis
Categorical data
The *k*-modes type algorithms
Optimization objective function
The between-cluster information

ABSTRACT

The *k*-modes algorithm and its modified versions are widely used to cluster categorical data. However, in the iterative process of these algorithms, the updating formulae, such as the partition matrix, cluster centers and attribute weights, are computed based on within-cluster information only. The between-cluster information is not considered, which maybe result in the clustering results with weak separation among different clusters. Therefore, in this paper, we propose a new term which is used to reflect the separation. Furthermore, the new optimization objective functions are developed by adding the proposed term to the objective functions of several existing *k*-modes algorithms. Under the optimization framework, the corresponding updating formulae and convergence of the iterative process is strictly derived. The above improvements are used to enhance the effectiveness of these existing *k*-modes algorithms whilst keeping them simple. The experimental studies on real data sets from the UCI (University of California Irvine) Machine Learning Repository illustrate that these improved algorithms outperform their original counterparts in clustering categorical data sets and are also scalable to large data sets for their linear time complexity with respect to either the number of data objects, attributes or clusters.

1. Introduction

Clustering is an unsupervised classification technique that is used to partition a set of unlabeled objects and ensure the objects which have high similarity into the same clusters. To tackle this problem, various types of clustering algorithms have been developed in the literature (e.g., [1–8] and references therein), which has extensive applications in various domains, including information retrieval, image processing and biological engineering. Since there are a number of categorical data produced in our real lives, recently increasing attention has been paid to clustering categorical data [9–13].

However, the lack of intuitive geometric properties for categorical data imposes several difficulties on clustering them [14,15]. For example, since the domains of categorical attributes are unordered, the distance functions for numerical values fail to capture resemblance between categorical values. Furthermore, for numerical data, the representative of a cluster is often defined as the mean of objects in the cluster. However, it is infeasible to compute the mean for categorical values. These imply that the techniques used in clustering numerical data are not applicable to categorical data. Therefore, it is widely recognized that designing clustering techniques to directly tackle this kind of data is very important for many applications. So far, several clustering algorithms have been reported [15–24] to solve the problem. Among them, the *k*-modes algorithm [19,20] and its several modified versions [25–28] are well known for their efficiency.

The *k*-modes type clustering algorithms often begin with an initial set of cluster centers and use the alternating minimization method to solve a nonconvex optimization problem in finding cluster solutions [1]. However, in the alternative process, the update formulae of partition matrix and cluster centers 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 may result in the clustering results with weak between-cluster separation. In [29], we proposed a fuzzy clustering algorithm with the between-cluster information. The numerical experimental studies illustrated that when handling data sets with fuzzy boundaries between clusters, the between-cluster information can effectively help users to find out good clustering results. Therefore, we will use the between-cluster information to improve the effectiveness of the *k*-modes algorithm and its modified versions in this paper. The major contributions are as follows:

- Unlike most existing *k*-modes type clustering algorithms, both the within-cluster compactness and between-cluster separation are employed at the same time to develop the new optimization objective functions, which are used to derive the improved clustering algorithms.

* Corresponding author.
E-mail addresses: sxbailiang@hotmail.com (L. Bai), ljy@sxu.edu.cn (J. Liang).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.024

- The updating formulae of the proposed clustering algorithms are derived, and the convergence of the proposed algorithms under the optimization framework is proved.
- The performance and scalability of the proposed clustering algorithms is investigated by using real data sets from UCI.

The rest of this paper is organized as follows. A detailed review of the k-modes type algorithms is presented in Section 2. In Section 3, the new k-modes type algorithms are presented and analyzed. Section 4 illustrates the performance and scalability of the proposed algorithms. Finally, a concluding remark is given in Section 5.

2. The k-modes algorithm and its modified versions

2.1. Categorical data

As we know, the structural data are stored in a table, where each row (tuple) represents facts about an object. A data table is also called an information system in rough set theory [30–32]. Data in the real world usually contain categorical attributes [12]. More formally, a categorical data table is defined as a quadruple $IS=(U,A,V,f)$, where:

- $U=\{x_1,x_2,\ldots,x_n\}$ is a nonempty set of n data points, called a universe;
- $A=\{a_1,a_2,\ldots,a_m\}$ is a nonempty set of m categorical attributes;
- V is the union of attribute domains, i.e., $V=\bigcup_{j=1}^{m}V_{a_j}$, where $V_{a_j}=\{a_j^{(1)},a_j^{(2)},\ldots,a_j^{(n_j)}\}$ is the value domain of categorical attribute a_j and is finite and unordered, e.g., for any $1\le p\le q\le n_j$, either $a_j^{(p)}=a_j^{(q)}$ or $a_j^{(p)}\ne a_j^{(q)}$. Here, n_j is the number of categories of attribute a_j for $1\le j\le m$;
- $f: R\times A\to V$ is an information function such that $f(x_i,a_j)\in V_{a_j}$ for $1\le i\le n$ and $1\le j\le m$, where $R=V_{a_1}\times V_{a_2}\times\cdots\times V_{a_m}$ and $U\subseteq R$.

2.2. The k-modes clustering algorithm

The k-modes clustering algorithm is an extension of the k-means algorithm [2] by using a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. These extensions have removed the numeric-only limitation of the k-means algorithm and enable the k-means clustering process to be used to efficiently cluster large categorical data sets from real world databases.

The objective function of the k-modes algorithm is defined as follows [20]:

$$F_0(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}d_0(z_l,x_i) \tag{1}$$

subject to

$$\begin{cases} w_{li}\in\{0,1\}, & 1\le l\le k,1\le i\le n,\\ \sum_{l=1}^{k}w_{li}=1, & 1\le i\le n,\\ 0<\sum_{i=1}^{n}w_{li}<n, & 1\le l\le k, \end{cases} \tag{2}$$

where $k(\le n)$ is a known number of clusters; $W=[w_{li}]$ is a k-by-n $\{0,1\}$ matrix, w_{li} is a binary variable, and indicates whether object x_i belongs to the lth cluster, $w_{li}=1$ if x_i belongs to the lth cluster and 0 otherwise; $Z=[z_1,z_2,\ldots,z_k]$ and $z_l=[f(z_l,a_1),f(z_l,a_2),\ldots,f(z_l,a_m)]$ is the lth cluster center with categorical attributes a_1, a_2, ..., a_m; $d_0(z_l,x_i)$ is the simple matching dissimilarity measure between object x_i and the center z_l of the lth cluster which is described as

$$d_0(z_l,x_i)=\sum_{j=1}^{m}\delta_0^{a_j}(z_l,x_i), \tag{3}$$

where

$$\delta_0^{a_j}(z_l,x_i)=\begin{cases}1, & f(z_l,a_j)\ne f(x_i,a_j),\\ 0, & f(z_l,a_j)=f(x_i,a_j).\end{cases} \tag{4}$$

Similar to the k-means algorithm, the k-modes algorithm uses the alternating method to minimize the function F_0 with the constraints in (2). In each iteration, W and Z are updated by the following formulae: When Z is given, W is updated by

$$\hat{w}_{li}=\begin{cases}1 & \text{if } d_0(\hat{z}_l,x_i)\le d_0(\hat{z}_h,x_i),1\le h\le k,\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

for $1\le i\le n$, $1\le l\le k$. When W is given, Z is updated by

$$f(z_l,a_j)=a_j^{(r)}\in V_{a_j} \tag{6}$$

where

$$\begin{aligned}&|\{x_i|f(x_i,a_j)=a_j^{(r)},w_{li}=1,x_i\in U\}|\\ &\ge|\{x_i|f(x_i,a_j)=a_j^{(t)},w_{li}=1,x_i\in U\}|,\quad 1\le t\le n_j,\end{aligned} \tag{7}$$

for $1\le j\le m$. Here, $V_{a_j}=\{a_j^{(1)},a_j^{(2)},\ldots,a_j^{(n_j)}\}$, n_j is the number of categories of attribute a_j for $1\le j\le m$.

2.3. Ng's improved k-modes algorithm

Ng and He et al. [25,26] introduced a new dissimilarity measure based on the relative attribute frequencies of the cluster modes. Using the new dissimilarity measure to the k-modes algorithm can improve the accuracy of the clustering results and strengthen the within-cluster similarity. The new dissimilarity measure $d_1(z_l,x_i)$ is defined as follows:

$$d_1(z_l,x_i)=\sum_{j=1}^{m}\delta_1^{a_j}(z_l,x_i), \tag{8}$$

where

$$\delta_1^{a_j}(z_l,x_i)=\begin{cases}1, & f(z_l,a_j)\ne f(x_i,a_j),\\ 1-\dfrac{|c_{ljr}|}{|c_l|}, & f(z_l,a_j)=f(x_i,a_j)\end{cases} \tag{9}$$

where $|c_l|=|\{x_i|w_{li}=1,x_i\in U\}|$ and $|c_{ljr}|=|\{x_i|f(x_i,a_j)=a_j^{(r)},w_{li}=1,x_i\in U\}|$. According to the definition of $\delta_1^{a_j}$, the dominant level of the mode category is considered in the calculation of the dissimilarity measure. This modification allows the algorithm to recognize a cluster with weak intra-similarity and, therefore, assign less similar objects to such a cluster so that the generated clusters have strong intra-similarities. Correspondingly, the k-modes objective function with the new dissimilarity measure is written as

$$F_1(W,Z)=\sum_{i=1}^{n}\sum_{l=1}^{k}w_{li}d_1(z_l,x_i) \tag{10}$$

subject to the same conditions as in (2). The updating formulae of the modified k-modes algorithm [26] are the same as the k-modes algorithm.

2.4. The weighted k-modes algorithm

To effectively handle high-dimensional categorical data sets, Huang [28] proposed the weighted k-modes clustering algorithm. The algorithm can automatically compute variable weights in the k-modes clustering process. It extends the standard k-modes algorithm with one additional step to compute variable weights at each iteration of the clustering process. The variable weight

is inversely proportional to the sum of the within-cluster variances of the variable. As such, noise variables can be identified and their effects on the clustering result are significantly reduced. The weighted k-modes objective function [28] is written as

$$F_2(W,Z,\Lambda)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{li}d_2(z_l,x_i) \tag{11}$$

subject to the conditions in (2) and

$$\lambda_j\in[0,1],\quad \sum_{j=1}^{m}\lambda_j=1,\ \ 1\le j\le m, \tag{12}$$

where

$$d_2(z_l,x_i)=\sum_{j=1}^{m}\delta_2^{a_j}(z_l,x_i)=\sum_{j=1}^{m}\lambda_j^{\beta}\delta_0^{a_j}(z_l,x_i) \tag{13}$$

is the weighted dissimilarity measure, $\Lambda=[\lambda_j]$ is a vector, λ_j is the weight for the jth attribute, which is used to identify the importance of the attribute in clustering, and $\beta\in(1,+\infty)$ is a parameter for controlling attribute weight λ_j.

Similar to solving (1), the objective function (11) can be locally minimized by iteratively updating W, Z and V. Here, the updating formulae of W and Z are the same as the k-modes algorithm. When W and Z are given, V is updated by [28]

$$\hat{\lambda}_j=\begin{cases}0 & \text{if } D_j=0,\\ \dfrac{1}{\sum_{h=1,D_h\neq 0}^{m}\left[\dfrac{D_j}{D_h}\right]^{1/(\beta-1)}} & \text{if } D_j\neq 0\end{cases} \tag{14}$$

for $1\le j\le m$, where

$$D_j=\sum_{i=1}^{n}\sum_{l=1}^{k}\hat{w}_{li}\delta_0^{a_j}(\hat{z}_l,x_i). \tag{15}$$

2.5. The lack of between-cluster information

Although the above modified algorithms can effectively improve the accuracy of the clustering results of the k-modes algorithm, it is noted that the k-modes algorithm and its modified versions face the local minimum problem. That is, the clustering results guarantee local minimum solutions only. Its performance heavily depends on the initial cluster centers. Furthermore, according to (5), (6), (7) and (13), we remark that the update formulae 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. These algorithms ignore 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.

According to Fig. 1, we see that $d_g(x_i,z_1)$ is equal to $d_g(x_i,z_1^*)$ $(g=0,1,2)$. If the dissimilarity between the object and the cluster centers is only taken into account to compute W, z_1 has no more representability to x_i than z_1^*. However, the separation between z_1^* and other cluster centers is weak, compared to that between z_1 and other cluster centers. 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 z_1 and reduce that of z_1^*, which makes more objects prone to belong to z_1 than z_1^*.

When computing 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.

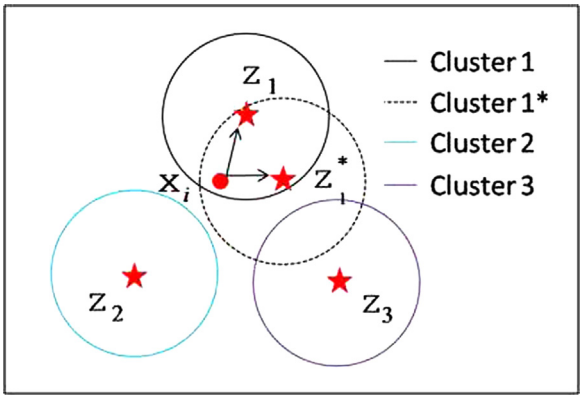


Fig. 1. An illustrative example about the effect of the between-cluster information on W

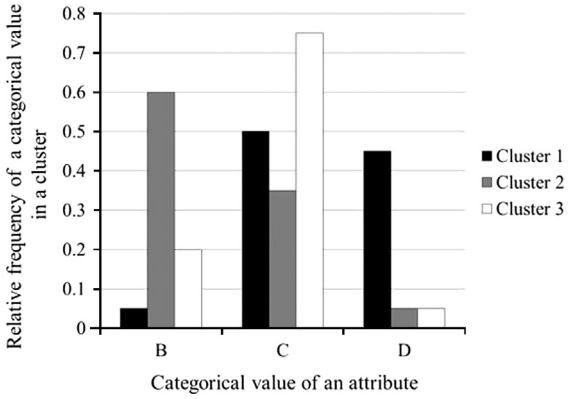


Fig. 2. An illustrative example about the effect of the between-cluster information on Z.

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.

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 section, we will propose a novel clustering technique for clustering categorical data, where the within-cluster and between-cluster information will be simultaneously employed to derive updating formulae.

3. Plus the between-cluster information

In this section, we will first give a definition of the between-cluster similarity term which is used to evaluate the between-cluster separation. Furthermore, we will respectively modify the k-modes objective functions (1), (10) and (11) by adding the between-cluster similarity term to them. Finally, we will develop the new k-modes type clustering algorithms based on these modified objective functions to obtain updating formulae.

3.1. The between-cluster similarity term

The between-cluster similarity term is defined as

$$B_g(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}S_g(z_l) \tag{16}$$

where $S_g(z_l)$ denotes the similarity between the lth cluster represented by z_l and other clusters, $\sum_{i=1}^{n} w_{li}$ is a weight of $S_g(z_l)$, which is the number of objects in the lth cluster, $g\in\{0,1,2\}$ is used to denote the different k-modes algorithms.

To compute z_l independent of z_h $(1\le h\ne l\le k)$, we employ the mean of the similarity between z_l and all the objects in the data set to evaluate the separation between clusters, instead of the mean of the similarity between z_l and other cluster centers. That is,

$$S_g(z_l)=\frac{1}{n}\sum_{i=1}^{n} s_g(z_l,x_i). \tag{17}$$

where $s_g(z_l,x_i)$ is a similarity measure between z_l and x_i. In the following subsections, we will provide the different definitions of s_g according to the different k-modes algorithms.

Next, we will explain whether the mean of the similarity between z_l and all the objects in the data set can be used to evaluate the separation between clusters. Given a set U of objects, if there is a data point $v\in R$ which can minimize the mean of the similarity between it and all the objects in U, i.e., $\min_v(1/|U|)\sum_{x_i\in U}s_g(v,x_i)$, the data point will be used as the representative point of U to reflect the global features of U. The larger the $S_g(z_l)$ is, the closer the z_l is to the representative point of U, and the more global features of U z_l reflects. However, a cluster tends to exist in local space. When $S_g(z_l)$ is very large, 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 cluster. If it is selected as a representative point of the lth cluster, the separation between the lth cluster and other clusters will be weak. Therefore, we can use (17) to evaluate the between-cluster separation.

3.2. Huang's k-modes plus the between-cluster similarity term

We modify the objective function (1) by adding the between-cluster similarity term to it so that we can simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation. The new objective function is written as follows:

$$F_{n_0}(W,Z,\gamma)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}d_0(z_l,x_i)+\gamma\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}S_0(z_l), \tag{18}$$

subject to the same conditions as in (2), where the parameter γ is 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). Here, the similarity measure s_0 is defined as

$$s_0(z_l,x_h)=\sum_{j=1}^{m}\phi_0^{a_j}(z_l,x_h), \tag{19}$$

where

$$\phi_0^{a_j}(z_l,x_h)=1-\delta_0^{a_j}(z_l,x_h). \tag{20}$$

When the initial set Z of cluster centers and γ are given, a key issue is how to derive rigorously the updating formulae of W and Z and guarantee that a local minimal solution of $F_{n_0}(W,Z,\gamma)$ 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 be fixed and consider the problem:*

$\min_W F_{n_0}(W,\hat{Z},\gamma)$ *subject to* (2).

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

$$\hat{w}_{li}=\begin{cases}1 & \text{if } d_0(\hat{z}_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p)\\ & \quad\le d_0(\hat{z}_h,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_h,x_p),\\ & \quad 1\le h\le k,\\ 0 & \text{otherwise.}\end{cases}$$

Proof. For a given Z, all the inner sums of the quantity

$$\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}d_0(z_l,x_i)+\gamma\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p)$$
$$=\sum_{i=1}^{n}\sum_{l=1}^{k} w_{li}\left[d_0(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p)\right],$$

are independent. Minimizing the quantity is equivalent to minimizing each inner sum. We write the ith inner sum $(1\le i\le n)$ as

$$\varphi_i=\sum_{l=1}^{k} w_{li}\left[d_0(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p)\right].$$

When $w_{hi}=1$, we have $w_{ti}=0, 1\le t\le k, t\ne h$ and

$$\varphi_i=d_0(z_h,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p).$$

It is clear that φ_i is minimized iff $d_0(z_h,x_i)+\gamma(1/n)\sum_{p=1}^{n}s_0(z_h,x_p)$ is minimal for $1\le h\le k$. The result follows. □

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

$\min_Z F_{n_0}(\hat{W},Z,\gamma)$ *subject to* (2).

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

$$f(\hat{z}_l,a_j)=a_j^{(r)}\in V_{a_j}$$

which satisfies

$$\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}\ge\frac{|c_{ljq}|}{|c_l|}-\gamma\frac{|c_{jq}|}{n},\quad 1\le q\le n_j,$$

for $1\le j\le m$, where $|c_{ljr}|=|\{x_i|f(x_i,a_j)=a_j^{(r)},\hat{w}_{li}=1\}|$, $|c_l|=\sum_{i=1}^{n}\hat{w}_{li}$ and $|c_{jr}|=|\{x_p|f(x_p,a_j)=a_j^{(r)},x_p\in U\}|$.

Proof. For a given W, all the inner sums of the quantity

$$\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\left[d_0(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_0(z_l,x_p)\right]$$
$$=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{li}\left[\delta_0^{a_j}(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}\phi_0^{a_j}(z_l,x_p)\right],$$

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

$$\psi_{l,j}=\sum_{i=1}^{n} w_{li}\left[\delta_0^{a_j}(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}\phi_0^{a_j}(z_l,x_p)\right].$$

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

$$\psi_{l,j}=\sum_{i=1,f(x_i,a_j)\ne a_j^{(q)}}^{n} w_{li}+\gamma\sum_{i=1}^{n} w_{li}\frac{1}{n}\sum_{p=1}^{n}\phi_0^{a_j}(z_l,x_p)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} w_{li} - \sum_{i=1, f(x_i,a_j)=a_j^{(q)}}^{n} w_{li} \\
&\quad + \gamma \sum_{i=1}^{n} w_{li} \frac{1}{n} \sum_{p=1}^{n} \phi_0^{a_j}(z_l, x_p) \\
&= \sum_{i=1}^{n} w_{li} - \left(\sum_{i=1, f(x_i,a_j)=a_j^{(q)}}^{n} w_{li} - \gamma \sum_{i=1}^{n} w_{li} \frac{1}{n} \sum_{p=1}^{n} \phi_0^{a_j}(z_l, x_p) \right) \\
&= \sum_{i=1}^{n} w_{li} - (|\{x_i | f(x_i, a_j) = a_j^{(q)}, w_{li} = 1\}| \\
&\quad - \gamma \frac{1}{n} \sum_{i=1}^{n} w_{li} |\{x_p | f(x_p, a_j) = a_j^{(q)}, x_p \in U\}|) \\
&= |c_l| - |c_l| \left(\frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n} \right).
\end{aligned}
$$

When W is given, $|c_l|$ is fixed. It is clear that $\psi_{l,j}$ is minimized iff $\frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n}$ is maximal for $1 \le q \le n_j$. 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 centers is updated according to Theorem 2 in each iteration. The convergence of the proposed algorithm can be obtained as in Theorem 3 below.

Theorem 3. *For any given $\gamma (\ge 0)$, the k-modes algorithm with the between-cluster similarity term converges to a local minimal solution 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 centers (modes). We then show that each possible center appears at most once in the iterative process. Assume that $Z^{(t_1)} = Z^{(t_2)}$, where $t_1 \ne t_2$. We can compute the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ for $Z^{(t_1)}$ and $Z^{(t_2)}$, respectively. Therefore, we have

$$F_{n_0}(W^{(t_1)}, Z^{(t_1)}, \gamma) = F_{n_0}(W^{(t_1)}, Z^{(t_2)}, \gamma) = F_{n_0}(W^{(t_2)}, Z^{(t_2)}, \gamma).$$

However, the sequence $F_{n_0}(\cdot, \cdot, \gamma)$ generated by the iterative method is strictly decreasing. Hence, the result follows. □

3.3. Ng's improved k-modes plus the between-cluster similarity term

The objective function (10) is modified as follows:

$$
\begin{aligned}
F_{n_1}(W, Z, \gamma) &= \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d_{n_1}(z_l, x_i) \\
&\quad + \gamma \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} S_1(z_l),
\end{aligned} \tag{21}
$$

subject to the same conditions as in (2). The dissimilarity measure d_{n_1} is defined as

$$d_{n_1}(z_l, x_i) = \sum_{j=1}^{m} \delta_{n_1}^{a_j}(z_l, x_i), \tag{22}$$

where

$$
\delta_{n_1}^{a_j}(z_l, x_i) = \begin{cases} 1, & f(z_l, a_j) \ne f(x_i, a_j), \\ 1 - \dfrac{\frac{|c_{ljr}|}{|c_l|} - \gamma \frac{|c_{jr}|}{n}}{\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n} \right)}, & f(z_l, a_j) = f(x_i, a_j). \end{cases} \tag{23}
$$

Here, $|c_{ljr}|/|c_l| - \gamma |c_{jr}|/n$ is used to reflect the dominant level of the mode category. When $|c_{ljr}|/|c_l| - \gamma |c_{jr}|/n = \sum_{q=1}^{n_j} (|c_{ljq}|/|c_l| - \gamma |c_{jq}|/n)$, the mode category is 100% dominant and d_{n_1} is the same as d_0 in the original k-modes algorithm. When $\gamma = 0$, d_{n_1} becomes d_1 in Ng's improved k-modes algorithm. The similarity measure s_1 is defined as

$$s_1(z_l, x_h) = \sum_{j=1}^{m} \phi_1^{a_j}(z_l, x_h), \tag{24}$$

where

$$\phi_1^{a_j}(z_l, x_h) = 1 - \delta_{n_1}^{a_j}(z_l, x_h). \tag{25}$$

To obtain the local minimal value of the modified objective function, we will provide Theorems 4 and 5 to iteratively update W and Z.

Theorem 4. *Let $\hat{Z}$ and γ be fixed and consider the problem:*

$\min_W F_{n_1}(W, \hat{Z}, \gamma)$ *subject to* (2).

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

$$
\hat{w}_{li} = \begin{cases} 1 & \text{if } d_{n_1}(\hat{z}_l, x_i) + \gamma \frac{1}{n} \sum_{p=1}^{n} s_1(z_l, x_p) \\ & \quad \le d_{n_1}(\hat{z}_h, x_i) + \gamma \frac{1}{n} \sum_{p=1}^{n} s_1(z_h, x_p), \\ & \quad 1 \le h \le k, \\ 0 & \text{otherwise}. \end{cases}
$$

Proof. Similar to Theorem 1. □

Theorem 5. *Let $\hat{W}$ and γ be fixed and consider the problem:*

$\min_Z F_{n_1}(\hat{W}, Z, \gamma)$ *subject to* (2).

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

$$f(\hat{z}_l, a_j) = a_j^{(r)} \in V_{a_j}$$

which satisfies

$$\frac{|c_{ljr}|}{|c_l|} - \gamma \frac{|c_{jr}|}{n} \ge \frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n}, \quad 1 \le q \le n_j,$$

for $1 \le j \le m$.

Proof. For a given W, all the inner sums of the quantity

$$
\begin{aligned}
&\sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} \left[d_{n_1}(z_l, x_i) + \gamma \frac{1}{n} \sum_{p=1}^{n} s_1(z_l, x_p) \right] \\
&\quad = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{li} \left[\delta_{n_1}^{a_j}(z_l, x_i) + \gamma \frac{1}{n} \sum_{p=1}^{n} \phi_1^{a_j}(z_l, x_p) \right],
\end{aligned}
$$

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

$$\psi_{l,j} = \sum_{i=1}^{n} w_{li} \left[\delta_{n_1}^{a_j}(z_l, x_i) + \gamma \frac{1}{n} \sum_{p=1}^{n} \phi_1^{a_j}(z_l, x_p) \right].$$

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

$$
\begin{aligned}
\psi_{l,j} &= |c_l| - |c_{ljr}| + |c_{ljr}| \left(1 - \frac{\frac{|c_{ljr}|}{|c_l|} - \gamma \frac{|c_{jr}|}{n}}{\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n} \right)} \right) \\
&\quad + \gamma \sum_{i=1}^{n} w_{li} \frac{1}{n} \sum_{p=1}^{n} \phi_{a_j}(z_l, x_p) \\
&= |c_l| - |c_{ljr}| + |c_{ljr}| \left(1 - \frac{\frac{|c_{ljr}|}{|c_l|} - \gamma \frac{|c_{jr}|}{n}}{\sum_{q=1}^{n_j} \left(\frac{|c_{ljq}|}{|c_l|} - \gamma \frac{|c_{jq}|}{n} \right)} \right)
\end{aligned}
$$

$$+\gamma|c_l|\frac{|c_{jr}|}{n}\frac{\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}}{\sum_{q=1}^{n_j}\left(\frac{|c_{ljq}|}{|c_l|}-\gamma\frac{|c_{jq}|}{n}\right)}$$

$$=|c_l|-|c_l|\frac{\left(\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}\right)^2}{\sum_{q=1}^{n_j}\left(\frac{|c_{ljq}|}{|c_l|}-\gamma\frac{|c_{jq}|}{n}\right)}$$

$$=|c_l|-|c_l|\frac{\left(\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}\right)^2}{1-\gamma}.$$

When W is given, $|c_l|$ is fixed. It is clear that $\psi_{l,j}$ is minimized iff

$$\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}$$

is maximal for $1\le r\le n_j$. The result follows. □

By comparing the results in Theorems 2 and 5, the cluster centers Z are updated in the same manner even when we use different dissimilarity measures in (3) and (22), respectively. Combining Theorems 4 and 5 forms an iterative optimization method to minimize the objective function (21) in which the partition matrix W is computed according to Theorem 4 and the set Z of cluster centers is updated according to Theorem 5 in each iteration. The convergence of the proposed algorithm can be obtained as in Theorem 6 below.

Theorem 6. *For any given $\gamma(\ge 0)$, Ng's improved k-modes algorithm with the between-cluster similarity term converges to a local minimal solution 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 centers (modes). We then show that each possible center appears at most once in the iterative process. Assume that $Z^{(t_1)}=Z^{(t_2)}$, where $t_1\ne t_2$. We can compute the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ for $Z^{(t_1)}$ and $Z^{(t_2)}$, respectively. Therefore, we have

$$F_{n_1}(W^{(t_1)},Z^{(t_1)},\gamma)=F_{n_1}(W^{(t_1)},Z^{(t_2)},\gamma)=F_{n_1}(W^{(t_2)},Z^{(t_2)},\gamma).$$

However, the sequence $F_{n_1}(\cdot,\cdot,\gamma)$ generated by the iterative method is strictly decreasing. Hence, the result follows. □

3.4. The weighted k-modes plus the between-cluster similarity term

We modify the objective function (11) by adding the between-cluster similarity term to it. The new objective function is written as follows:

$$F_{n_2}(W,Z,\Lambda,\gamma)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}d_2(z_l,x_i)+\gamma\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}S_2(z_l),\tag{26}$$

subject to the same conditions as in (2) and (12). The similarity measure s_2 is defined as

$$s_2(z_l,x_h)=\sum_{j=1}^{m}\phi_2^{a_j}(z_l,x_h),\tag{27}$$

where

$$\phi_0^{a_j}(z_l,z_h)=\lambda_j^{\beta}-\delta_2^{a_j}(z_l,x_h).\tag{28}$$

When γ is given, a key issue is how to derive rigorously the updating formulae of W, Z and Λ and guarantee that a local minimal solution of $F_{n_2}(W,Z,\Lambda,\gamma)$ can be obtained in a finite number of iterations. The matrices W, Z and Λ are calculated according to the following three theorems:

Theorem 7. *Let $\hat{Z}$, $\hat{\Lambda}$ and γ be fixed and consider the problem:*

$$\min_{W}F_{n_2}(W,\hat{Z},\hat{\Lambda},\gamma)\quad \textit{subject to}\ (2).$$

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

$$\hat{w}_{li}=\begin{cases}1 & \text{if } d_2(\hat{z}_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_2(z_l,x_p)\\ & \quad\le d_2(\hat{z}_h,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_2(z_h,x_p),\\ & \quad 1\le h\le k,\\ 0 & \text{otherwise.}\end{cases}$$

Proof. Similar to Theorem 1. □

Theorem 8. *Let $\hat{W}$, $\hat{\Lambda}$ and γ be fixed and consider the problem:*

$$\min_{Z}F_{n_2}(\hat{W},Z,\hat{\Lambda},\gamma)\quad \textit{subject to}\ (2).$$

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

$$f(\hat{z}_l,a_j)=a_j^{(r)}\in V_{a_j}$$

which satisfies

$$\frac{|c_{ljr}|}{|c_l|}-\gamma\frac{|c_{jr}|}{n}\ge\frac{|c_{ljq}|}{|c_l|}-\gamma\frac{|c_{jq}|}{n},\quad 1\le q\le n_j,$$

for $1\le j\le m$.

Proof. For a given W and Λ, all the inner sums of the quantity

$$\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}\left[d_2(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}s_2(z_l,x_p)\right]$$

$$=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{li}\left[\delta_2^{a_j}(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}\phi_2^{a_j}(z_l,x_p)\right],$$

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

$$\psi_{l,j}=\sum_{i=1}^{n}w_{li}\left[\delta_2^{a_j}(z_l,x_i)+\gamma\frac{1}{n}\sum_{p=1}^{n}\phi_2^{a_j}(z_l,x_p)\right].$$

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

$$\psi_{l,j}=\sum_{i=1,f(x_i,a_j)\ne a_j^{(t)}}^{n}w_{li}\lambda_j^{\beta}+\gamma\sum_{i=1}^{n}w_{li}\frac{1}{n}\sum_{p=1}^{n}\phi_2^{a_j}(z_l,x_p)$$

$$=(|c_l|-|c_{ljt}|)\lambda_j^{\beta}+\gamma|c_l|\frac{|c_{ljt}|}{n}\lambda_j^{\beta}$$

$$=|c_l|\lambda_j^{\beta}-|c_l|\left(\frac{|c_{ljt}|}{|c_l|}-\gamma\frac{|c_{jt}|}{n}\right)\lambda_j^{\beta}.$$

When W and Λ are given, $|c_l|$ and λ_j are fixed. It is clear that $\psi_{l,j}$ is minimized iff

$$\frac{|c_{ljt}|}{|c_l|}-\gamma\frac{|c_{jt}|}{n}$$

is maximal for $1\le t\le n_j$. The result follows. □

Theorem 8 tells us that the cluster centers Z are updated in the same manner as the original k-modes algorithm even when we use the new weighted dissimilarity measure. It implies that computing the minimizer $\hat{Z}$ is independent of $\hat{\Lambda}$.

Theorem 9. *Let $\hat{W}$, $\hat{Z}$ and γ be fixed and $\beta>1$, $F_{n_2}(\hat{W},\hat{Z},\Lambda,\gamma)$ reaches a local minimum only if Λ satisfies the following conditions:*

$$\hat{\lambda}_j=\begin{cases}0 & \text{if } D'_j=0,\\ \dfrac{1}{\sum_{h=1,D'_h\neq 0}^{m}\left[\dfrac{D'_j}{D'_h}\right]^{1/(\beta-1)}} & \text{if } D'_j\neq 0\end{cases} \tag{29}$$

for $1\le j\le m$, where

$$D'_j=\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right), \tag{30}$$

$|c_{ljr_l}|=|\{x_i|f(x_i,a_j)=f(z_l,a_j),w_{li}=1,x_i\in U\}|$,
$|c_{jr_l}|=|\{x_i|f(x_i,a_j)=f(z_l,a_j),x_i\in U\}|$ *for* $1\le l\le k$.

Proof. We rewrite F_{n_1} as

$$\begin{aligned}
&F_{n_2}(\hat{W},\hat{Z},\Lambda,\gamma)\\
&=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}\sum_{j=1}^{m}\lambda_j^{\beta}\delta_0^{a_j}(z_l,x_i)\\
&\quad+\gamma\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}\sum_{j=1}^{m}\lambda_j^{\beta}\frac{1}{n}\sum_{p=1}^{n}\phi_0^{a_j}(z_l,x_p)\\
&=\sum_{j=1}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}\delta_0^{a_j}(z_l,x_i)\\
&\quad+\gamma\sum_{j=1}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\sum_{i=1}^{n}w_{li}\frac{1}{n}\sum_{p=1}^{n}\phi_0^{a_j}(z_l,x_p)\\
&=\sum_{j=1}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|\right)\\
&\quad+\gamma\sum_{j=1}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}|c_l|\frac{|c_{jr_l}|}{n}\\
&=\sum_{j=1}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right)\\
&=\sum_{j=1,D'_j\neq 0}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right)
\end{aligned}$$

We know that $|c_l|$, $|c_{ljr_l}|$ and $|c_{jr_l}|$ are constants for fixed $\hat{W}$ and $\hat{Z}$. The Lagrangian multiplier technique is used to obtain the following unconstrained minimization problem:

$$\begin{aligned}
\tilde{P}(\Lambda,\alpha)=&\sum_{j=1,D'_j\neq 0}^{m}\lambda_j^{\beta}\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right)\\
&-\alpha\left(\sum_{j=1,D'_j\neq 0}^{m}\lambda_j-1\right),
\end{aligned} \tag{31}$$

where α is the Lagrangian multiplier. If $(\hat{\Lambda},\hat{\alpha})$ is a minimizer of $\tilde{P}(\Lambda,\alpha)$, the gradients in both sets of variables must vanish. Thus,

$$\frac{\partial\tilde{P}(\Lambda,\alpha)}{\partial\lambda_j}=\beta\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right)\lambda_j^{\beta-1}-\alpha=0,\quad 1\le j\le m, \tag{32}$$

and

$$\frac{\partial\tilde{P}(\Lambda,\alpha)}{\partial\alpha}=\sum_{j=1,D'_j\neq 0}^{m}\lambda_j-1=0. \tag{33}$$

From (32) and (33), we obtain

$$\hat{\lambda}_j=\frac{1}{\sum_{h=1,D'_h\neq 0}^{m}\left[\dfrac{\sum_{l=1}^{k}\left(|c_l|-|c_{ljr_l}|+\gamma|c_l|\frac{|c_{jr_l}|}{n}\right)}{\sum_{l=1}^{k}\left(|c_l|-|c_{lhr_l}|+\gamma|c_l|\frac{|c_{hr_l}|}{n}\right)}\right]^{1/(\beta-1)}} \tag{34}$$

where $D'_j\neq 0$. This shows that (34) is the necessary conditions to reach the minimum value of the objective function F_{n_2} when W and Z are fixed. □

Theorem 10. *For any given $\gamma(\ge 0)$, the weighted k-modes algorithm with the between-cluster similarity term converges to a local minimal solution in a finite number of iterations.*

Proof. We first note that there are only a finite number of possible partitions W. We then show that each possible partition W appears at most once by the algorithm. Assume that $W^{(t_1)}=W^{(t_2)}$, where $t_1\neq t_2$. We note that, given $W^{(t)}$, we can compute the minimizer $Z^{(t)}$ which is independent of $\Lambda^{(t)}$. For $W^{(t_1)}$ and $W^{(t_2)}$, we have the minimizers $Z^{(t_1)}$ and $Z^{(t_2)}$, respectively. Using $W^{(t_1)}$ and $Z^{(t_1)}$, and $W^{(t_2)}$ and $Z^{(t_2)}$, we can compute the minimizers $\Lambda^{(t_1)}$ and $\Lambda^{(t_2)}$, respectively, according to Theorem 9. Although $Z^{(t_1)}$ may not be equal to $Z^{(t_2)}$, $|c_{ljr}^{(t_1)}|/|c_l^{(t_1)}|-\gamma|c_{jr}^{(t_1)}|/n=|c_{ljr}^{(t_2)}|/|c_l^{(t_2)}|-\gamma|c_{jr}^{(t_2)}|/n$ and $|c_l^{(t_1)}|=|c_l^{(t_2)}|$ for $1\le j\le m$, $1\le l\le k$. It is clear that $\Lambda^{(t_1)}=\Lambda^{(t_2)}$. Therefore, we obtain

$$F_{n_2}(W^{(t_1)},Z^{(t_1)},\Lambda^{(t_1)},\gamma)=F_{n_2}(W^{(t_2)},Z^{(t_2)},\Lambda^{(t_2)},\gamma).$$

However, the sequence $F_{n_2}(\cdot,\cdot,\cdot,\gamma)$ generated by the algorithm is strictly decreasing. Hence, the result follows. □

3.5. The effect of the parameter γ

The parameter γ is used to maintain a balance between the effect of the within-cluster information and that of the between-cluster information. It has the following features in control of the clustering process:

- When $\gamma>0$, the between-cluster similarity term $B_g(W,Z)$ will play an important role in the minimization of F_{n_g}, $g\in\{0,1,2\}$. 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 centers to some locations which are 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 centers are moved to the locations of outliers in U. Therefore, we suggest $\gamma<1$.
- When $\gamma=0$, the between-cluster similarity term will not play any role in the clustering process. F_{n_g} will become the original objective functions F_g. The clustering process turns to minimize the within-cluster dispersion.
- When $\gamma<0$, the clustering process will try to move the cluster centers 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 k-modes type algorithms 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 proposed clustering algorithms, we will not select a fixed γ value but a sequence Γ which includes several γ values. In clustering process, a larger γ value is first used to obtain a clustering result (W,Z) or (W,Z,Λ). Furthermore, we 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 functions F_{n_g} is equivalent to minimizing the

original objective functions F_g. This means that in the proposed clustering algorithms, instead of directly minimizing the objective functions F_g with the constraints in (2), we consider a scheme of obtaining a solution of the problem at the limit of $\gamma \downarrow 0$ of

$$\min F_{n_g}(W,Z,\gamma) \text{ or } F_{n_g}(W,Z,\Lambda,\gamma) \quad \text{subject to (2).}$$

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 $1>\gamma_1>\gamma_2>\cdots>\gamma_o=0$. Initialize Z_1 and Λ_1 and set $e=1$.

Step 2: If $g\in\{0,1\}$, use Z_e as the initial values to compute $(\hat{W}_e,\hat{Z}_e)$ which is a local optimal solution of

$$\min F_{n_g}(W,Z,\gamma_e) \quad \text{subject to (2),}$$

and set $\mathbb{F}(e)=F_{n_g}(\hat{W}_e,\hat{Z}_e,\gamma_e)$; otherwise use Z_e and Λ_e as the initial values to compute $(\hat{W}_e,\hat{Z}_e,\hat{\Lambda}_e)$ which is a local optimal solution of

$$\min F_{n_g}(W,Z,\Lambda,\gamma_e) \quad \text{subject to (2),}$$

and set $\mathbb{F}(e)=F_{n_g}(\hat{W}_e,\hat{Z}_e,\hat{\Lambda}_e,\gamma_e)$.

Step 3: If $e>1$ and $\mathbb{F}(e)=\mathbb{F}(e-1)$ or $e\geq o$, then output $(\hat{W}_e,\hat{Z}_e)$ or $(\hat{W}_e,\hat{Z}_e,\hat{\Lambda}_e)$ and stop; otherwise set $Z_{e+1}=\hat{Z}_e$, $e=e+1$ and goto Step 2.

3.6. The computational complexity

The proposed clustering algorithms are scalable to the number of objects, attributes or clusters. This is because the proposed algorithms only add a new computational cost to the k-modes clustering process to calculate the between-cluster term. The runtime complexity can be analyzed as follows. We only consider the four major computational steps:

- *Computing the between-cluster similarity term*: Before implementing the proposed algorithms, we calculate and save the frequency of each categorical value of each attribute in U, which will be used to compute B_g. The step takes $O(n\sum_{j=1}^{m}n_j)$ operations.
- *Partitioning the objects*: Given Z or Z and Λ, each object is assigned to a cluster. This process simply computes 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 centers*: Given W or W and Λ, updating cluster centers 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.
- *Calculating attribute weights*: If $g=2$, we need to calculate Λ based on the given W and Z. In this step, we only go through the whole data set once to update the attribute weights. The computational complexity of this step is also $O(mnk)$.

If one needs t_e iterations to obtain a local minimal solution of F_{n_g} for each γ_e $(e=1,2,\ldots,o)$, the total computational complexity of the proposed algorithms 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.

4. Experimental analysis

The main aim of this section is to evaluate the clustering performance and scalability of the proposed algorithms. We have selected the ten categorical data sets from the UCI Machine Learning Repository [33] which are widely used in other published papers to test the clustering algorithms. In the data sets, if the attribute value of an object is missing, then we denote the attribute value by $*$.

4.1. Performance analysis

To evaluate the effectiveness of clustering algorithms, we will first introduce three evaluation indices, i.e., accuracy (*AC*), precision (*PE*), and recall (*RE*) [34], which are defined as

$$AC=\frac{\sum_{i=1}^{k}a_i}{n},\quad PE=\frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+b_i}\right)}{k},\quad RE=\frac{\sum_{i=1}^{k}\left(\frac{a_i}{a_i+c_i}\right)}{k},$$

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.

Furthermore, we will use the proposed algorithms to improve the k-modes algorithm [20], the weighted k-modes algorithm [28] and Ng's k-modes algorithm [26]. These improved algorithms will be compared with those original algorithms. Due to the fact that the performance of the k-modes type algorithms depends on initial cluster centers, we randomly select 100 initial cluster centers and carry out 100 runs of each algorithm on these data sets. In each run, the same initial cluster centers are used in these algorithms. Before implementing these improved algorithms, we need to provide a sequence $\Gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_o\}$. We set $\gamma_1=0.9$, $\gamma_o=0$ and $\gamma_{e+1}=\gamma_e-0.1$, $1\leq e<o$.

We present the comparative results of clustering on the following data sets.

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 has 32 instances described by 56 categorical attributes. It contains three classes.

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 10 records each.

Zoo data: Zoo data set contains 101 elements described by 17 Boolean-valued attributes and 1 type attribute. Data set with 101 elements belongs to seven classes.

Heart disease data: The data set generated at the Cleveland Clinic has 303 instances with eight categorical and five numeric 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 clusters: psoriasis (112 data objects), seborrheic dermatitis (61 data objects), lichen planus (72 data objects), pityriasis rosea (49 data objects), chronic 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 9 categorical attributes. It has two clusters: Benign (458 data objects) and Malignant (241 data objects).

Letter recognition 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 experiment.

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 the two classes, edible (4208 objects) and poisonous (3916 objects).

Performance results: According to Tables 1–9, we see that the performances of these improved algorithms on most data sets are evidently better than the corresponding original algorithms for *AC*, *PE*, and *RE*. Since the between-cluster terms are added, the clustering accuracies of the original algorithms are enhanced by around 4–5%. On the breast cancer data set, these experimental results shown us that the between-cluster terms can help the original algorithms to find the better clustering results and weaken the effect of initial cluster centers. However, we also notice that some of the original algorithms provide superior results to the improved ones for *PE* in Tables 6 and 7. That could happen. The main reason is that as the accuracy *AC* of a clustering is enhanced, the average of the cluster purity *PE* does not necessarily increase. In some cases, while the number of objects correctly classified (i.e., $\sum_{i=1}^{k} a_i$) of a clustering increases the number of objects incorrectly classified (b_i) in some of the clusters may be reduced. In these cases, it is not enough to only consider *PE* to evaluate the clustering result. We do not deny *PE*. Conversely, we believe that the larger the value of *PE*, the better the clustering solution. We mean that other measures should be simultaneously considered.

4.2. Scalability analysis

In the scalability analysis, we test the scalability of the original *k*-modes algorithm plus the between-cluster term on the connect-4 and census data sets. The computational results are performed by

Table 1
Means of AC, PE, RE for 100 runs of algorithms on the lung cancer data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.5322	0.5803	0.5344	0.5631	0.5516	0.6003
PE	0.5886	0.6196	0.5967	0.5972	0.6181	0.6480
RE	0.5293	0.5800	0.5352	0.5646	0.5427	0.6069

Table 2
Means of AC, PE, RE for 100 runs of algorithms on the soybean data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.8553	0.9234	0.8613	0.9068	0.9396	0.9979
PE	0.9020	0.9462	0.8948	0.9291	0.9598	0.9983
RE	0.8407	0.9121	0.8471	0.8924	0.9291	0.9975

Table 3
Means of AC, PE, RE for 100 runs of algorithms on the zoo data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.8324	0.8509	0.8283	0.8552	0.8528	0.8900
PE	0.8433	0.8572	0.6528	0.8534	0.8374	0.8525
RE	0.6576	0.6646	0.8345	0.6947	0.7058	0.7525

Table 4
Means of AC, PE, RE for 100 runs of algorithms on the dermatology data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.6869	0.7718	0.6854	0.8195	0.7642	0.8211
PE	0.7633	0.8660	0.7692	0.8545	0.7756	0.8905
RE	0.5750	0.6709	0.5765	0.7704	0.6607	0.7668

Table 5
Means of AC, PE, RE for 100 runs of algorithms on the heart disease data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.7462	0.7882	0.7472	0.7728	0.7836	0.8053
PE	0.7573	0.7886	0.7566	0.7746	0.7839	0.8015
RE	0.7446	0.7869	0.7455	0.7767	0.7788	0.8056

Table 6
Means of AC, PE, RE for 100 runs of algorithms on the credit approval data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.7367	0.7647	0.7442	0.7578	0.7612	0.7942
PE	0.7617	0.7602	0.7403	0.7585	0.7574	0.7923
RE	0.7358	0.7656	0.7455	0.7517	0.7608	0.7952

Table 7
Means of AC, PE, RE for 100 runs of algorithms on the breast cancer data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.8482	0.9270	0.8530	0.8991	0.8645	0.8770
PE	0.8731	0.9343	0.8733	0.9119	0.9062	0.8770
RE	0.7893	0.9050	0.7968	0.8959	0.8066	0.8245

Table 8
Means of AC, PE, RE for 100 runs of algorithms on the letters data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.6910	0.7350	0.6836	0.7229	0.7299	0.7523
PE	0.7016	0.7496	0.6943	0.7339	0.7450	0.7684
RE	0.6911	0.7354	0.6838	0.7232	0.7304	0.7529

Table 9
Means of AC, PE, RE for 100 runs of algorithms on the mushroom data set.

Index	Huang's *k*-modes		Weighted *k*-modes		Ng's *k*-modes	
	Original	Improved	Original	Improved	Original	Improved
AC	0.7176	0.8190	0.7106	0.8006	0.7969	0.8366
PE	0.7453	0.8360	0.7414	0.8239	0.8079	0.8494
RE	0.7132	0.8149	0.7056	0.7956	0.7933	0.8330

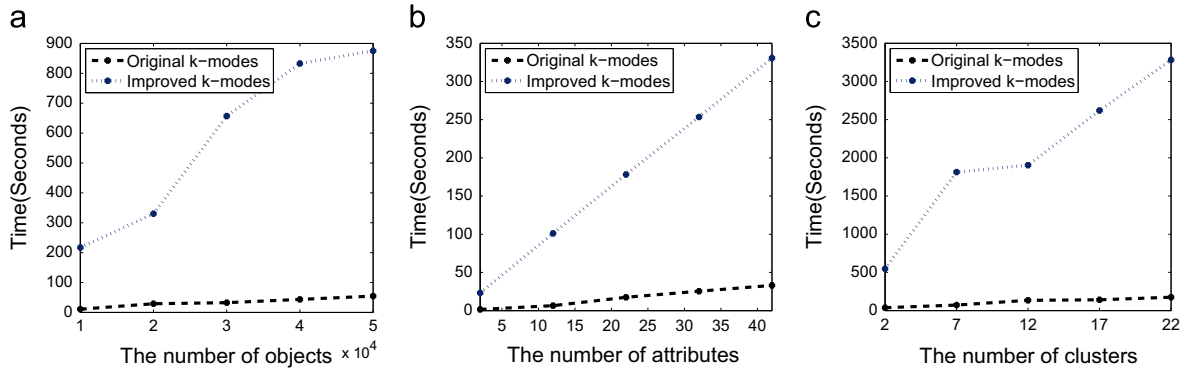


Fig. 3. Test on the connect-4 data: (a) Computational times for different numbers of objects. (b) Computational times for different numbers of attributes. (c) Computational times for different numbers of clusters.

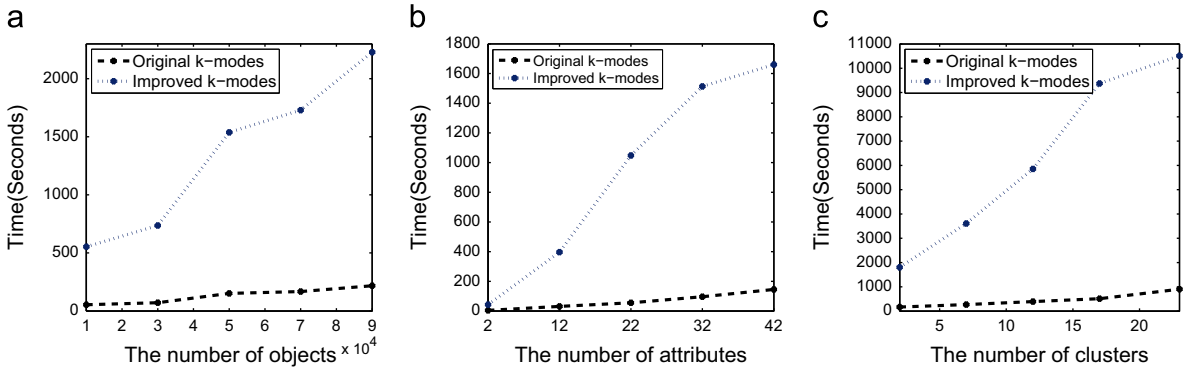


Fig. 4. Test on the census data: (a) Computational times for different numbers of objects. (b) Computational times for different numbers of attributes. (c) Computational times for different numbers of clusters.

using a machine with an Intel Q9400 and 2 G RAM. The computational times of the proposed algorithm is plotted with respect to the number of objects, attributes and clusters, while the other corresponding parameters are fixed.

Connect-4 data: The data set contains all legal 8-ply positions in the game of connect-4 in which neither player has won yet, and in which the next move is not forced. This data set contains 67,557 instances and 42 categorical attributes. It has three class: win (44,473), loss (16,635) and draw (6449). We take 50,000 records from this data set to test the scalability of the algorithms. Fig. 3 (a) shows the computational times against the number of objects, while the number of attributes is 42 and the number of clusters is 3. Fig. 3(b) shows the computational times against the number of attributes, while the number of clusters is 3 and the number of objects is 30,000. Fig. 3(c) shows the computational times against the number of clusters, while the number of attributes is 42 and the number of objects is 30,000.

Census data: The census data has 2,458,284 records with 68 categorical attributes, about 352 Mbytes in total. It was derived from the USCensus1990raw data set which was obtained from the (U.S. Department of Commerce) Census Bureau website using the Data Extraction System. We take 100,000 records from this data set to test the scalability of the algorithms. Fig. 4(a) shows the computational times against the number of objects, while the numbers of attributes are 68 and the number of clusters is 3. Fig. 4(b) shows the computational times against the number of attributes, while the numbers of clusters are 3 and the number of objects is 50,000. Fig. 4(c) shows the computational times against the number of clusters, while the numbers of attributes are 68 and the number of objects is 50,000.

Scalability results: According to Figs. 3 and 4, the improved *k*-modes algorithm requires more computational times than the original *k*-modes algorithm. It is an expected outcome since it requires the more additional arithmetic operations of the between-cluster information and the numbers of iterations than the original one. Since the between-cluster term is added, the solution capability of the original algorithm is boosted. Therefore, its numbers of iterations in the solution process increase. From these figures, we see that its computational times are around tenfold those of the original *k*-modes algorithm. However, we also see that the proposed algorithm is scalable, i.e., the computational times increase linearly with respect to either the number of objects, attributes or clusters. Therefore, it can cluster large categorical data efficiently.

5. Conclusions

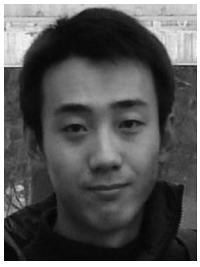
In this paper, we have presented a novel clustering technique for categorical data, which simultaneously minimizes the within-cluster dispersion and enhances the between-cluster separation in the clustering process. This technique is used to improve the performance of the existing *k*-modes algorithms. Furthermore, we rigorously derive the updating formulae and the convergence of the improved algorithms under the optimization framework. The time complexity of the proposed algorithms has been analyzed which is linear with respect to either the number of data objects, attributes or clusters. We have tested the proposed algorithms using several real data sets from UCI. Experimental results have shown that the improved algorithms are effective and scalable in clustering categorical data sets.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 71031006, 61305073), the National Key Basic Research and Development Program of China (973) (No. 2013CB329404), the Foundation of Doctoral Program Research of Ministry of Education of China (No. 20131401120001).

References

[1] A. Jain, R. Dubes, Algorithms for Clustering Data, Prentice Hall, New Jersey, 1988.
[2] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1967, pp. 281–297.
[3] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. 39 (1) (1977) 1–38.
[4] A. Likas, M. Vlassis, J. Verbeek, The global *k*-means clustering algorithm, Pattern Recognit. 35 (2) (2003) 451–461.
[5] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: SIGMOD Conference, Berkeley, 1996, pp. 103–114.
[6] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
[7] Y. Andrew, M. Ng, Y. Jordan, On spectral clustering: analysis and an algorithm, Adv. Neural Inf. Process. Syst. 14 (2001) 849–856.
[8] J. Brendan, D. Delbert, Clustering by passing messages between data points, IEEE Trans. Fuzzy Syst. 315 (16) (2007) 972–976.
[9] C. Aggarwal, C. Magdalena, P. Yu, Finding localized associations in market basket data, IEEE Trans. Knowl. Data Eng. 14 (1) (2002) 51–62.
[10] D. Barbara, S. Jajodia, Applications of Data Mining in Computer Security, Kluwer, Dordrecht, 2002.
[11] A. Baxevanis, F. Ouellette, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd edn., Wiley, NY, 2001.
[12] K. Gowda, E. Diday, Symbolic clustering using a new dissimilarity measure, Pattern Recognit. 24 (6) (1991) 567–578.
[13] N. Wrigley, Categorical Data Analysis for Geographers and Environmental Scientists, Longman, London, 1985.
[14] 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.
[15] H. Chen, K. Chuang, M. Chen, On data labeling for clustering categorical data, IEEE Trans. Knowl. Data Eng. 20 (11) (2008) 1458–1472.
[16] D. Fisher, Knowledge acquisition via incremental conceptual clustering, Mach. Learn. 2 (2) (1987) 139–172.
[17] V. Ganti, J. 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.
[18] S. Guha, R. Rastogi, S. Kyuseok, Rock: a robust clustering algorithm for categorical attributes, in: Proceedings of 15th International Conference on Data Engineering, no. 23–26, Sydney, Australia, 1999, pp. 512–521.
[19] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, in: Proceedings of SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery, 1997, pp. 1–8.
[20] Z. Huang, Extensions to the *k*-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (3) (1998) 283–304.
[21] L. Bai, J. Liang, C. 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.
[22] L. Bai, J. Liang, C. Dang, F. Cao, A novel attribute weighting algorithm for clustering high-dimensional categorical data, Pattern Recognit. 44 (12) (2011) 2843–2861.
[23] D. Barbara, Y. Li, J. Couto, Coolcat: an entropy-based algorithm for categorical clustering, in: Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, pp. 582–589.
[24] K. Chen, L. Liu, He-tree: a framework for detecting changes in clustering structure for categorical data streams, VLDB J. 18 (5) (2009) 1241–1260.
[25] Z. He, S. Deng, X. Xu, Improving *k*-modes algorithm considering frequencies of attribute values in mode, in: Proceedings of Computational Intelligence and Security, 2005, pp. 157–162.
[26] M. Ng, M.J. Li, Z.X. Huang, Z. He, On the impact of dissimilarity measure in *k*-modes clustering algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 503–507.
[27] O. San, V. Huynh, Y. Nakamori, An alternative extension of the *k*-means algorithm for clustering categorical data, Pattern Recognit. 14 (2) (2004) 241–247.
[28] Z. Huang, M. Ng, H. Rong, Z. Li, Automated variable weighting in *k*-means type clustering, IEEE Trans. Fuzzy Syst. 27 (5) (2005) 657–668.
[29] L. Bai, J. Liang, C. Dang, F. Cao, A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets Syst. 215 (2013) 55–73.
[30] J. Liang, J. Wang, Y. Qian, A new measure of uncertainty based on knowledge granulation for rough sets, Inf. Sci. 179 (4) (2009) 458–470.
[31] Z. Pawlak, Rough Sets—Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, London, Dordrecht, Boston, 1991.
[32] Y. Qian, J. Liang, W. Pedrycz, C. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artif. Intell. 174 (5–6) (2010) 597–618.
[33] Uci Machine Learning Repository, 2011 ⟨http://www.ics.uci.edu/mlearn/MLRepository.html⟩.
[34] Y. Yang, An evaluation of statistical approaches to text categorization, J. Inf. Retr. 1 (1–2) (2004) 67–88.



Liang Bai received the PhD degree of computer science in 2012, from the School of Computer and Information Technology at Shanxi University, China, where he is currently a lecturer. His research interests are in the areas of data mining and machine learning.

Jiye Liang received the MS and PhD degrees from Xi′an Jiaotong University, Xi′an, China, in 1990 and 2001, respectively. He is currently a professor with the School of Computer and Information Technology and the Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education, Shanxi University, Taiyuan, China. He has authored or coauthored more than 150 journal papers in his research fields. His current research interests include computational intelligence, granular computing, data mining, and machine learning.