Space Structure and Clustering of Categorical Data

Yuhua Qian, *Member, IEEE*, Feijiang Li, *Student Member, IEEE*, Jiye Liang, Bing Liu, *Fellow, IEEE*, and Chuangyin Dang, *Senior Member, IEEE*

***Abstract*— Learning from categorical data plays a fundamental role in such areas as pattern recognition, machine learning, data mining, and knowledge discovery. To effectively discover the group structure inherent in a set of categorical objects, many categorical clustering algorithms have been developed in the literature, among which k-modes-type algorithms are very representative because of their good performance. Nevertheless, there is still much room for improving their clustering performance in comparison with the clustering algorithms for the numeric data. This may arise from the fact that the categorical data lack a clear space structure as that of the numeric data. To address this issue, we propose, in this paper, a novel data-representation scheme for the categorical data, which maps a set of categorical objects into a Euclidean space. Based on the data-representation scheme, a general framework for space structure based categorical clustering algorithms (SBC) is designed. This framework together with the applications of two kinds of dissimilarities leads two versions of the SBC-type algorithms. To verify the performance of the SBC-type algorithms, we employ as references four representative algorithms of the k-modes-type algorithms. Experiments show that the proposed SBC-type algorithms significantly outperform the k-modes-type algorithms.**

***Index Terms*— Categorical data, clustering, dissimilarity, k-modes-type algorithms, space structure.**

I. Introduction

IN DATA mining and knowledge discovery, there are many types of data, including the numeric data, the categorical data, the text data, the image data, the audio data, and so on. Transforming the unstructured data into the structured data is one of the common research methods. The existing data processing methods often map these data types to one of the numeric data or the categorical data [14], [15], [21], [32], [62]. Specially, data mining from the categorical data is an important research topic, in which several types of methods have been developed, such as decision trees [44], rough sets [38], [42], [45], [47], [48], concept lattice [52], [54], granular computing [43], [46], [49], [59], [61], and others.

Clustering is an important tool in data mining and knowledge discovery, which is used to discover the grouping structure inherent in a set of objects. It has many applications in areas, such as text mining, granular computing, information retrieval, bioinformatics, Web data mining, customer analysis, and scientific data exploration [1], [6]–[8], [19], [41], [53]. The objective of a clustering algorithm is to group a set of unlabeled objects into several meaningful clusters so that the objects within the same clusters are close to each other and those from different clusters are very dissimilar with objects within other clusters. To address this problem, many types of clustering algorithms have been proposed for various clustering tasks in [3], [6], [9], [12], [16], [17], [20], [25], [26], [56], and [57].

In a clustering algorithm, a predefined similarity/dissimilarity measurement among objects is one of its key concepts, which significantly affect the clustering results of the objects and the performance of the algorithm.

For the numeric data, many successful algorithms have been developed according to various data distributions. The k-means-type algorithms are very representative [28], [30], [35], [60], which can effectively and efficiently organize the objects into several clusters [29], [37]. In k-means-type algorithms, all objects are described by several features with numeric domains. Therefore, these objects can be considered in a Euclidean space, and the dissimilarity/similarity between two objects can be measured by a Euclidean distance, a cosine distance, and so on [29]. When the dimensionality of a given data set is small, the Euclidean distance is often used, and when the dimensionality of a given data set is big, the cosine distance is usually employed.

Recently, increasing attention has been paid to the clustering categorical data, in which the objects are made up of nonnumerical data [13], [25], [55]. In fact, how to learn dissimilarity among the categorical data is a difficult problem all the time. For a supervised setting, Xie *et al.* [55] repeatedly updated the assignment of categorical symbols to real numbers to minimize the classification error based on the nearest neighbor technique. For an unsupervised setting, Alamuri *et al.* [2] gave a survey of distance/similarity measures for the categorical data. More generally, the 0–1 distance or its extensions are widely used for analyzing the categorical data.

Manuscript received August 31, 2014; revised December 23, 2014, April 1, 2015, and May 21, 2015; accepted June 19, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61322211, Grant 61432011, Grant 71031006, Grant 61202018, and Grant 61303008, in part by the Program for New Century Excellent Talents in University under Grant NCET-12-1031, in part by the National Key Basic Research and Development Program of China (973) under Grant 2013CB329404, in part by the Research Fund for the Doctoral Program of Higher Education under Grant 20121401110013, and in part by the Program for the Innovative Talents of Higher Learning Institutions of Shanxi, China, under Grant 20120301.

Y. Qian, F. Li, and J. Liang are with the School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China (e-mail: jinchengqyh@sxu.edu.cn; lfjfly@163.com; ljy@sxu.edu.cn).

B. Liu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: liub@cs.uic.edu).

C. Dang is with the Department of Manufacture Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: mecdang@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2451151

To discover the grouping structures of categorical data, several algorithms have also been reported [2], [3], [6], [12], [17], [20], [25], [26], [56]. Among these algorithms, the k-modes-type clustering algorithms are very representative and popular techniques, which can be used to the quickly clustering categorical data. This type of algorithms can overcome the limitation of the k-means-type algorithms only working on the numeric data. The k-modes-type algorithm was originally proposed in [25]–[27], which adopts the same paradigm as the k-means-type algorithms. Therefore, the k-modes algorithm can effectively cluster large categorical data sets from real-world databases. In recent years, several extended versions of the k-modes algorithm have been published [4], [5], [11], [22], [31], [34], [39], [40], [51]. In this paper, we call the k-modes algorithm and their extensions the k-modes-type algorithms.[1]

There are two key issues in the k-modes-type algorithms. One is how to measure the dissimilarity between two objects and the other is how to update each of cluster modes in each iteration. The former is addressed by introducing a 0–1 distance or its extensions, and the latter is solved by combining the frequencies of feature values in each feature domain in a cluster to represent the cluster in each iteration. Although the k-modes-type algorithms have good performance, one may consider the following problems.

1) Whether the performance of the clustering algorithms for the categorical data can be further enhanced or not.
2) Whether the 0–1 distance and its extensions are effective in revealing the grouping structure of objects or not.
3) Whether the method of updating cluster prototypes has the same excellent performance as that of computing the mean value of objects in a cluster with the k-means-type algorithms or not. These three problems induce the most important problem.
4) Whether the numeric data and the categorical data can intrinsically be unified in a clustering algorithm or not.

The above four problems are the main motivations of this paper. Unlike numeric data, the categorical data have no clear linear space, which brings a challenge for effectively discovering the grouping structure of objects. In this paper, we will not be concerned with how to improve the k-modes-type algorithms themselves. Instead, we will focus on how to construct a space structure for the categorical data and propose a new type of algorithms based on this space structure for organizing the categorical data.

In this paper, we first present a new data-representation scheme for mapping the objects of a categorical data set into a Euclidean space, in which each categorical data object represents a single coordinate. Based on the space structure, we then give a new type of algorithm (just SBC) for the clustering categorical data, which combines the construction of the space structure of the categorical data and the k-means algorithm paradigm together. The framework of SBC algorithms does not change the convergence of k-means-type algorithms, and can simultaneously minimize the within cluster dispersion. Finally, we employ a Euclidean distance and a cosine distance for verifying the validity of the SBC algorithm. By incorporating these two distances into the SBC algorithm, we obtain two versions of the algorithm. Numerical experiments on nine real data sets from the UCI repository show that each of these two versions always converges, and statistically possesses much better performance than the k-modes-type algorithms for clustering categorical data. It is worth noting that the clustering performance of the SBC algorithm with the cosine distance is getting better as the size of the categorical data set increasing.

The rest of this paper is organized as follows. The k-modes-type algorithms are briefly reviewed in Section II. In Section III, through analyzing the challenge of the clustering categorical data, we establish a space structure for representing a categorical data set, which maps all the objects into a Euclidean space. In Section IV, a general framework of algorithms based on the space structure is introduced to cluster a categorical data set, just named SBC. In Section V, we give a series of experimental analyses that include the rationality, the validity, and the efficiency of the SBC algorithm. Finally, Section VI concludes this paper with some remarks.

II. k-Modes-Type Algorithms

In this section, we will review the theoretical framework of the k-modes-type algorithms with some remarks.

We assume the set of objects $U = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ to be clustered is stored in a database table T defined by a set of features, $A = \{a_1, a_2, \ldots, a_m\}$. Each feature a_j describes a domain of values, denoted by $V(a_j)$, associated with a defined semantic and a data type. A domain $V(a_j)$ is defined as categorical if it is finite and unordered, i.e., $V(a_j) = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)}\}$, where n_j is the number of categories of feature a_j, $1 \le j \le m$. In this paper, we only consider a data set with a single data type. If each feature in A is categorical, then U is called a categorical data set.

The k-modes-type algorithms use the k-means-type paradigm to organize the categorial data. The objective function of clustering n categorical objects into k clusters is to find W and Z that minimize [25], [26]

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

subject to

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

where n is the number of objects in U, and k is the known number of clusters. $W = [w_{li}]$ is a k-by-n matrix, w_{li} shows whether $\mathbf{x}_i$ belongs to the lth cluster, and $w_{li} = 1$ if x_i belongs to the lth cluster and 0 otherwise. $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\} \subseteq R$, where $R = V(a_1) \times V(a_2) \times \cdots \times V(a_m)$ and $\mathbf{z_l} = [z_{l1}, z_{l2}, \ldots, z_{lm}]$ are the lth cluster prototype with categorical features $a_1, a_2, \ldots, a_m$. $d(\mathbf{z}_l, \mathbf{x}_i)$ is the

[1] In fact, there is another k-modes algorithm (originated in [10]) that works with continuous data, which uses a local bandwidth at each point rather than a global one. Because we only pay attention to how to organize the categorical data, in this paper, the discussed k-modes algorithm especially refers to the method proposed in [25]–[27].

simple matching dissimilarity measure between $\mathbf{x}_i$ 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

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

The above optimization problem can be solved by iteratively solving the following process of the k-modes algorithm [25]–[27].

Step 1: Choose k initial cluster prototypes $Z^{(1)} \subseteq R$. Determine W^1 such that $F(W^1, Z^{(1)})$ 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 go to 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 go to Step 2.

However, the clustering result obtained by the above k-modes algorithm may not be a globe optimization, but a local optimization.

As we know, Z of the k-modes algorithm is determined based on the frequencies of feature values in the cluster. The most frequent feature value in each feature domain in a cluster is selected to represent the cluster in each iteration, which can minimize the within-cluster dissimilarity. As Bai *et al.* [4] pointed out, however, this updated method of cluster prototypes ignores the representability of other feature values, which may effect the clustering performance of algorithms for categorical data sets. In order to overcome this limitation, several improved algorithms were proposed in [4], [5], [11], [22], [31], [34], [39], [40], and [51], where a prototype in a cluster is a list of several categorical values in the feature with their frequencies in the cluster as the weights in each iteration of a k-modes algorithm. This idea implies that the higher the weight of a categorical value in the cluster is, the more representability the categorical value has in the cluster. These clustering methods can further improve the performance of the original k-modes algorithm, which all can be induced to a framework of the k-modes-type algorithms.

III. Space Structure of the Categorical Data

In this section, we first analyze dissimilarity measures in the k-modes-type algorithms and its shortcomings. Based on the analysis, we propose a new data-representation scheme of categorical data with a space structure. Finally, we discuss several advantages of the data-representation scheme.

From the working mechanism of the k-modes-type algorithms, it can be seen that its validity depends on two core concepts: 1) the dissimilarity measure and 2) the updating method of cluster prototypes.

Given a database table with the categorical data $T = (U, A)$, where $U = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ be a set of objects, and $A = \{a_1, a_2, \ldots, a_m\}$ is a set of features. The dissimilarity measure between two objects is defined by (3). For determining the cluster label of each object, it is solved by the following method:

$$\forall \mathbf{x_i}, \quad \text{if} \quad d(\mathbf{x_i}, \mathbf{z_l}) \leq d(\mathbf{x_i}, \mathbf{z_h}), \quad 1 \leq h \leq k$$

then $\mathbf{x_i}$ belongs to the cluster w_l.

For updating cluster prototypes, it is solved by

$$z_{lj} = a_j^{(r)} \in V(a_j)$$

where $\sum_{x_{ij}=a_j^{(r)}, \mathbf{x_i} \in U} w_{li} = \max_{q=1}^{n_j} \{\sum_{x_{ij}=a_j^{(q)}, \mathbf{x_i} \in U} w_{li}\}$ for $1 \leq j \leq m$. Here, n_j is the number of categories of feature a_j for $1 \leq j \leq m$. In essence, the method of updating cluster prototypes can also be equivalently characterized by the following form:

$$d\left(x_{ij}, z_{lj}^{(r)}\right) = \min_{q=1}^{n_j} \left\{ d\left(x_{ij}, z_{lj}^{(q)}\right) \right\} \tag{4}$$

for $1 \leq j \leq m$, where $z_{lj}^{(q)}$ is the object described by the feature a_j in the cluster w_l.

Although in several modified versions [4], [5], [11], [22], [31], [34], [39], [40], [51], a prototype is often updated by combining several categorical values in the feature with their frequencies in a cluster as the weights, they have the same working mechanism as the classical k-modes algorithm.

From the above analysis, one can see that the dissimilarity measure d is the most important concept in the k-modes-type algorithms, which significantly affects the clustering performance of a categorical data set. However, the 0–1 distance may not be a finer-grained metric for measuring the difference between two categorical objects, and the method of updating cluster prototypes determined by the 0–1 distance may not be a much better solution. Specially, the limitation brings a pity, i.e., we cannot use some of existing proven methodologies for the numeric data to deal with the categorical data. Hence, it is also difficult for the numeric data and the categorical data to be clustered together in a semantically unified framework.[2]

To address the above issue, in what follows, we will try to give a new data-representation scheme for the categorical data, such that a categorical data set can possess a space structure like the numeric data.

To facilitate discussion, we give a categorical data set in the table form, which is shown in Table I.

TABLE I
Categorical Data Set

	a_1	a_2	$\cdots$	a_m
x_1	$a_1(x_1)$	$a_2(x_1)$	$\cdots$	$a_m(x_1)$
x_2	$a_1(x_2)$	$a_2(x_2)$	$\cdots$	$a_m(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_n	$a_1(x_n)$	$a_2(x_n)$	$\cdots$	$a_m(x_n)$

[2]It is a very important issue to cluster a mixed data set with the numeric data and the categorical data. Several researchers have tried to solve it in [20], [23], [24], [26], [34], and [36]. In these studies, a usual method is to fuse these two types of features with a variable weight. Lee and Yun [33] proposed a new procedure using multidimensional scaling for the clustering categorical and the numerical data. In existing approaches, in essence, the numeric data and the categorical data are still dealt with separately, which is not really unified semantically.

TABLE II
SPACE STRUCTURE MATRIX OF THE CATEGORICAL DATA

	$x_1(c_1)$	$x_2(c_2)$	$\cdots$	$x_n(c_n)$
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_n	p_{n1}	p_{n2}	$\cdots$	p_{nn}

In this table, $U = \{x_1, x_2, \ldots, x_n\}$ are n categorical objects and $A = \{a_1, a_2, \ldots, a_m\}$ are m categorical features with weights $W = \{w_1, w_2, \ldots, w_m\}$, where $a_i(x_j)$ is the categorical value of the object x_j under the feature a_i. Now, we give a new data-representation scheme of a categorical object. First, we compute the probability that any two categorical objects are equal to each other, which is formally defined as

$$p_{ij} = \sum_{k=1}^{m} w_k \theta_k(\mathbf{x_i}, \mathbf{x_j}) \Big/ \sum_{k=1}^{m} w_k \tag{5}$$

where

$$\theta_k(\mathbf{x_i}, \mathbf{x_j}) = \begin{cases} 1, & a_k(x_i) = a_k(x_j) \\ 0, & a_k(x_i) \neq a_k(x_j). \end{cases}$$

Through rerepresenting these objects of the categorical data set, we can obtain a space structure (it is a Euclidean space) described by the probabilities of similarities among objects. With loss of generality, we assume that all attributes have the same weight in this paper. In this method, one describes an object with a set of new dimensions $\{c_i = x_i,\ 1 \leq i \leq n\}$, which is shown in Table II.

Its working mechanism is illustrated by the following example.

Example 1: Given a categorical data set about three hexahedrons with six features, which is shown in Table III. Without loss of generality, we assume that the features have the same weight $w_1 = w_2 = \cdots = w_6 = 1/6$.

Using classical set theory, one can obtain set relationships among these three objects for each of six features, as shown in Fig. 1(a). Indeed, it can describe the grouping structure of the data set in each of the features. However, this method cannot well discover the grouping structure inherent in the data set in the entire feature space.

In what follows, we establish its space structure. First, from Table III, one can compute the probability of $\mathbf{x_i} = \mathbf{x_j}$ for any two objects in the categorical data set. Through computing, their results are shown in Table IV.

Through the space structure matrix, the original categorical data set is transformed into a Euclidean space with three new features $c_1 = x_1, c_2 = x_2$, and $c_3 = x_3$. The geometric structure of these three objects can clearly be given, in which objects x_1 and x_2 seem to be much closer.

Let (U, C) be a Euclidean space mapped by a categorical data set (U, A). In order to differentiate objects in (U, C) and those in (U, A), we denote a vector in (U, C) by $\mathbf{x}_C$ and that in (U, A) by $\mathbf{x}_A$, respectively.

Property 1: Let (U, C) be a Euclidean space mapped by a categorical data set (U, A). The following properties hold.

1) $\forall \mathbf{x}_C$ lies in the first quadrant of (U, C).
2) $\mathbf{x}_C = \mathbf{y}_C$ if $\mathbf{x}_A = \mathbf{y}_A$.
3) $p_{ij} = 1 - (1/m)d(\mathbf{x_i}_A, \mathbf{x_j}_A)$.
4) $0 \leq \langle \mathbf{x}_C, \mathbf{y}_C \rangle \leq 90°$.

In our opinion, though the 0–1 distance had widely been used for the clustering categorical data, it may not be a much finer-grained metric for measuring the difference between two categorical objects. After we transform an original categorical data set into its corresponding Euclidean space, the metric D on the Euclidean space would have much finer-grained performance than the 0–1 distance. It can be revealed by the following theorem. Easily, we assume that $D = |\mathbf{x_C} - \mathbf{y_C}|^2$ is a metric on (U, C).

Theorem 1: Let (U, A) be a categorical data set, $x_i, x_j, x_k \in U$, and (U, C) be the corresponding Euclidean space mapped by (U, A). If $d(\mathbf{x_i}_A, \mathbf{x_j}_A) = d(\mathbf{x_i}_A, \mathbf{x_k}_A)$, then $D(\mathbf{x_i}_C, \mathbf{x_j}_C) = D(\mathbf{x_i}_C, \mathbf{x_k}_C)$ may not hold.

Proof: From $d(\mathbf{x_i}_A, \mathbf{x_j}_A) = d(\mathbf{x_i}_A, \mathbf{x_k}_A)$ and (5), one has that $p_{ij} = p_{ik}$. Then

$$\begin{aligned} &D(\mathbf{x_i}_C, \mathbf{x_j}_C) - D(\mathbf{x_i}_C, \mathbf{x_k}_C) \\ &\quad = |\mathbf{x_i}_C - \mathbf{x_j}_C|^2 - |\mathbf{x_i}_C, \mathbf{x_k}_C|^2 \\ &\quad = \Bigg((p_{ii} - p_{ji})^2 + (p_{ij} - p_{jj})^2 + (p_{ik} - p_{jk})^2 \\ &\qquad + \sum_{t=1, t\neq i,j,k}^{n} (p_{it} - p_{jt})^2 \Bigg) \\ &\qquad - \Bigg((p_{ii} - p_{ki})^2 + (p_{ij} - p_{kj})^2 + (p_{ik} - p_{kk})^2 \\ &\qquad + \sum_{t=1, t\neq i,j,k}^{n} (p_{it} - p_{kt})^2 \Bigg) \\ &\quad = \sum_{t=1, t\neq i,j,k}^{n} (p_{it} - p_{jt})^2 - \sum_{t=1, t\neq i,j,k}^{n} (p_{it} - p_{kt})^2. \end{aligned}$$

In this equation, p_{jt} and p_{kt} may not be equal. Hence, $D(\mathbf{x_i}_C, \mathbf{x_j}_C)$, and $D(\mathbf{x_i}_C, \mathbf{x_k}_C)$ may also not be equal. ■

If we adopt one of the other metrics on an Euclidean space, the above theorem may also hold. We do not prove them one-by-one because of paper's compactness.

Theorem 1 implies such a phenomena that when objects y and z with respect to x on the original feature set A cannot be distinguished, they could be still differentiated on the Euclidean space induced by the new data-representation scheme. Hence, the new data-representation scheme can provide much finer characterization for relationships among categorical objects than the original one.

We can observe several advantages of the data-representation scheme as follows.

1) The object values under all features and their weights can be fused together into the data-representation scheme without loss of any information.
2) The objects can be described by a normalized form, in which each object value lies in the interval [0, 1].

TABLE III
DESCRIPTION OF SOME HEXAHEDRONS

Objects	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
x_1	Red	Green	Blue	Orange	Purple	Yellow
x_2	Red	Blue	Blue	Orange	Purple	Yellow
x_3	Green	Blue	Blue	Purple	Orange	Yellow

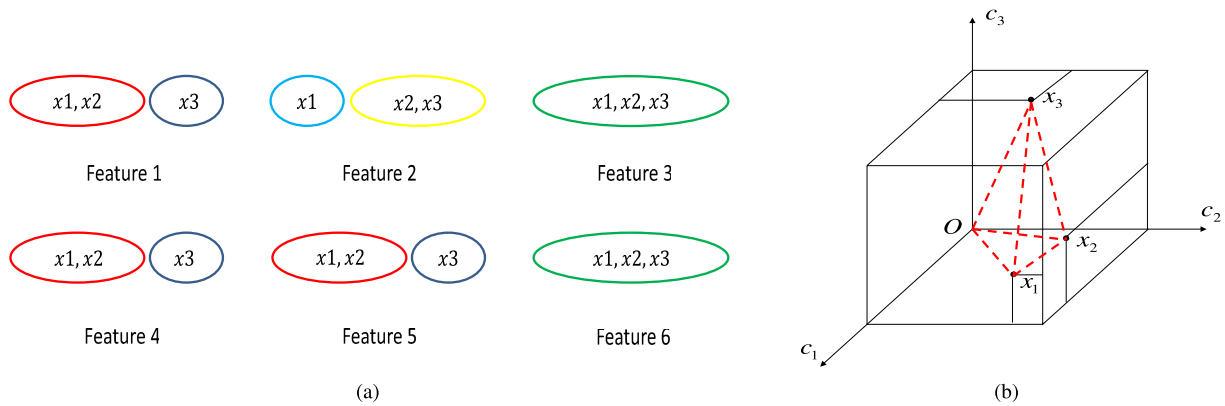


Fig. 1. Set relationship of the categorical data and its space structure. (a) Set relationship of three objects. (b) Space structure of three objects.

TABLE IV
SPACE STRUCTURE MATRIX OF THE DATA IN EXAMPLE 1

	x_1	x_2	x_3
x_1	1	5/6	2/6
x_2	5/6	1	3/6
x_3	2/6	3/6	1

3) The space structure is a Euclidean space, which eliminates the limitation of the classical data-representation scheme without a clear structure among categorical objects.
4) Through transforming the categorical data to a Euclidean space, we can use many existing excellent theories and methods for data mining and knowledge discovery from a categorical data set.

Due to these four advantages above, the new data-representation scheme will be very helpful for discovering the grouping structure inherent in a set of categorical objects.

IV. STRUCTURE-BASED CLUSTERING ALGORITHM

Based on the data-representation scheme in Section III, we can map a categorical data set into s Euclidean space with new dimensions. This allows us to take the advantage of the existing k-means-type algorithms for the clustering categorical data. In this section, we discuss the dissimilarity measure among objects in the space structure and propose a structure-based clustering algorithm.

In real applications, the scale of a categorical data set may be very large. When it is mapped into a Euclidean space, the dimensionality of the space will be much higher. Hence, we first introduce the following dissimilarity measure:

$$d_1(\mathbf{x}, \mathbf{y}) = \sqrt{2(1 - \cos\langle \mathbf{x}, \mathbf{y}\rangle)} \tag{6}$$

where $\cos\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}/\|\mathbf{x}\|\|\mathbf{y}\|$.

In general, the size of the data set for discovering its grouping structure is often of large scale. When the original categorical data was mapped to its corresponding Euclidean space, the updated data set will possess much higher dimensionality. In this paper, therefore, we will use this distance to quantify the difference between the objects in the Euclidean space. Based on the distance d_1, we can construct a corresponding objective function for a clustering algorithm for the categorical data.

The objective of clustering a set of n categorical objects into k clusters is to find Z that minimizes

$$F_1(Z) = \sum_{j=1}^{k} \sum_{\mathbf{x}\in\omega_j} d_1(\mathbf{x}, \mathbf{z_j}) \tag{7}$$

where $k(\leq n)$ is a known number of clusters, $Z = \{\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_k}\}$, $\mathbf{z_j}$ is the jth cluster center, and $d_1(\mathbf{x}, \mathbf{z_j}) = (2(1 - \cos\langle \mathbf{x}, \mathbf{z_j}\rangle))^{(1/2)}$.

In what follows, we discuss the dissimilarity measure in a data set with a much smaller scale.

Given a database table with the numeric data $T = (U, A)$, where $U = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ be a set of objects, and $A = \{a_1, a_2, \ldots, a_m\}$ is a set of features. Unlike the 0–1 distance in (3), in the classical k-means-type algorithms, a Euclidean distance is often employed for measuring the difference between two objects $\mathbf{x}$ and $\mathbf{y}$, which is formally defined as follows:

$$d'(\mathbf{x}, \mathbf{y}) = ||\mathbf{x}, \mathbf{y}||_p = ((\mathbf{x} - \mathbf{y})^p)^{\frac{1}{p}} \tag{8}$$

where $||\cdot||_p$ is a p-norm in a finite linear space. Moreover, it can be represented as

$$d'(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{m} (a_k(x) - a_k(y))^p \right)^{\frac{1}{p}}$$

where m is the number of features of the numeric data set.

If we transform objects in a categorical data set $T = (U, A)$ into a Euclidean space (U, C) with n new features $C = \{c_1, c_2, \ldots, c_n\}$, then the distance in (5) can be rewritten as follows:

$$d_2(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} (c_k(x) - c_k(y))^p \right)^{\frac{1}{p}} \tag{9}$$

where n is the number of new features in the Euclidean space (U, C), and $c_k(x)$ is computed by (5).

Based on the Euclidean distance, the objective of clustering a set of n categorical objects into k clusters is to find Z that minimizes

$$F_2(Z) = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \omega_j} d_2(\mathbf{x}, \mathbf{z_j}) \tag{10}$$

where $k(\leq n)$ is a known number of clusters, $Z = \{\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_k}\}$, $\mathbf{z_j}$ is the jth cluster center, and $d_2(\mathbf{x}, \mathbf{z_j}) = \|\mathbf{x}, \mathbf{z_j}\|_p$.

For these two measures above, the Euclidean distance measures the relative distance between two objects, whereas the cosine distance measures the angular distance between two objects. Many existing distance measures including them all can be used to measure the distance between two objects in the corresponding Euclidean space. The main motivation of this paper is to propose one type of space structure-based algorithm for clustering a categorical data set, in which every distance measure can generate its corresponding version.

In what follows, we propose one type of space structure-based algorithm for clustering a categorical data set, in which the space structure construction and the mechanism of the k-means algorithm are combined together.

In Algorithm 1 (its code can be downloaded [64]), the time complexity of mapping n objects into a space structure is $O(n^2)$. In addition, the time complexity of determining class labels of n objects at each iteration is $O(kn)$. Let the iterative times be p when the algorithm stops, one easily knows that the time complexity of the SBC algorithm is determined by n^2 and pkn. For clustering on many data sets, due to $pk < n$, its time complexity is often written as $O(n^2)$. Although the time complexity of the SBC is nonlinear, it may not be disappointing because of its better clustering performance. This can be seen in the experiment analyses in Section V.

Because of the convergence of the k-means-type algorithms, the minimization of F in (7) and that in (10) can also be finished by the SBC algorithm when the distance $\mathcal{D} = d_1$ and d_2, respectively.

Algorithm 1 Space Structure-Based Algorithm for Clustering a Categorical Data Set (SBC)

Input: A set of categorical objects $U = \{\mathbf{x_1}_A, \mathbf{x_2}_A, \ldots, \mathbf{x_n}_A\}$ and features $A = \{a_1, a_2, \ldots, a_m\}$;

Output: k clusters.

(1) Mapping all objects into a space structure $(U, C) = \{p_{ij}, 1 \leq i, j \leq n\}$ by Equation (5), and generating corresponding n objects $\{\mathbf{x_i}_C : i \leq n\}$, where $\mathbf{x_i}_C = (p_{i1}, p_{i2}, \ldots, p_{in})$.

(2) In the space structure, randomly choose k objects as initial cluster centers: $\mathbf{z_1}^0, \mathbf{z_2}^0, \ldots, \mathbf{z_k}^0$. Let $l = 0$.

(3) If $\mathcal{D}(\mathbf{x_i}_C, \mathbf{z_c}^{(l)}) = \min_j \{\mathcal{D}(\mathbf{x_i}_C, \mathbf{z_j}^{(l)})\}$, $i = 1, 2, \ldots, n$, then put $\mathbf{x_i}_C$ into the cluster $\omega_c^{(l+1)}$, which generates new clusters $\omega_j^{(l+1)}$ $(j = 1, 2, \ldots, k)$.

(4) Computing the center of each new cluster by

$$\mathbf{z_j}^{(l+1)} = \frac{1}{n_j^{(l+1)}} \sum_{\mathbf{x_i}_C \in \omega_j^{(l+1)}} \mathbf{x_i}_C, \quad j = 1, 2, \ldots, k$$

where $n_j^{(l+1)}$ is the number of objects in the cluster $\omega_j^{(l+1)}$.

(5) If $\mathbf{z_j}^{(l+1)} = \mathbf{z_j}^{(l)}$ $(j = 1, 2, \ldots, k)$, then the algorithm stops; otherwise, $l = l + 1$, go to (3).

V. Experimental Analysis

The primary objective of clustering algorithms is to discover the grouping structures inherent in the data. As the assumption is that a certain structure may exist in a given data set, a clustering algorithm is used to verify the assumption and recover the inherent structure. If an algorithm can discover the structure, then it may be a good solution.

In this section, we aim to verify the rationality, the clustering performance, and the computational efficiency of the SBC algorithm for pattern recognition from a categorical data set.

The nine categorical data sets used in the experiments are outlined in Table V, which were all downloaded from UCI repository of machine learning databases [63]. In this table, the class distribution shows the real partition status of each of these nine categorical data sets.

A. Correlation Analysis

To demonstrate how the proposed data representation can reflects the distribution information in an original categorical data set, this section first test the correlation between its original categorical data space and its corresponding Euclidean space, where the former refers to the original structure of the data induced by the 0–1 distance measure.

To conveniently address this issue, we construct the variable $X = \{X_1, X_2, \ldots, X_{C_n^2}\}$ from an original categorical data set and the variable $Y = \{Y_1, Y_2, \ldots, Y_{C_n^2}\}$ from its corresponding Euclidean space, where X_i means the value of normalized 0–1 distance between the ith pair of objects in an original categorical data set, and Y_j means the value of normalized cosine distance (or normalized Euclidean distance) between the jth pair of objects in its corresponding Euclidean space. In this experiment, through employing these two variables, we analyze the correlation between the new represented data and the original categorical data, which is quantified by the Pearson correlation coefficient [50]. Given two variables X and Y, the Pearson correlation coefficient $\rho_{X,Y}$ between X and Y

TABLE V
DESCRIPTION OF THE NINE PUBLIC DATA SETS FROM UCI

	Data sets	Number of samples	Number of features	Classes	Class distribution
1	Fitting contact lenses	24	4	3	$\{4, 5, 15\}$
2	Balloon	20	4	2	$\{8, 12\}$
3	Space shuttle autolanding	15	6	2	$\{6, 9\}$
4	Soybean-small	47	35	4	$\{10, 10, 10, 17\}$
5	Hayes-Roth-Hayes-Roth	132	4	3	$\{51, 51, 30\}$
6	Lymphography domain	142	18	2	$\{81, 61\}$
7	Vote	435	16	2	$\{168, 267\}$
8	Breast cancer	699	9	2	$\{458, 241\}$
9	Promoters	106	57	2	$\{53, 53\}$

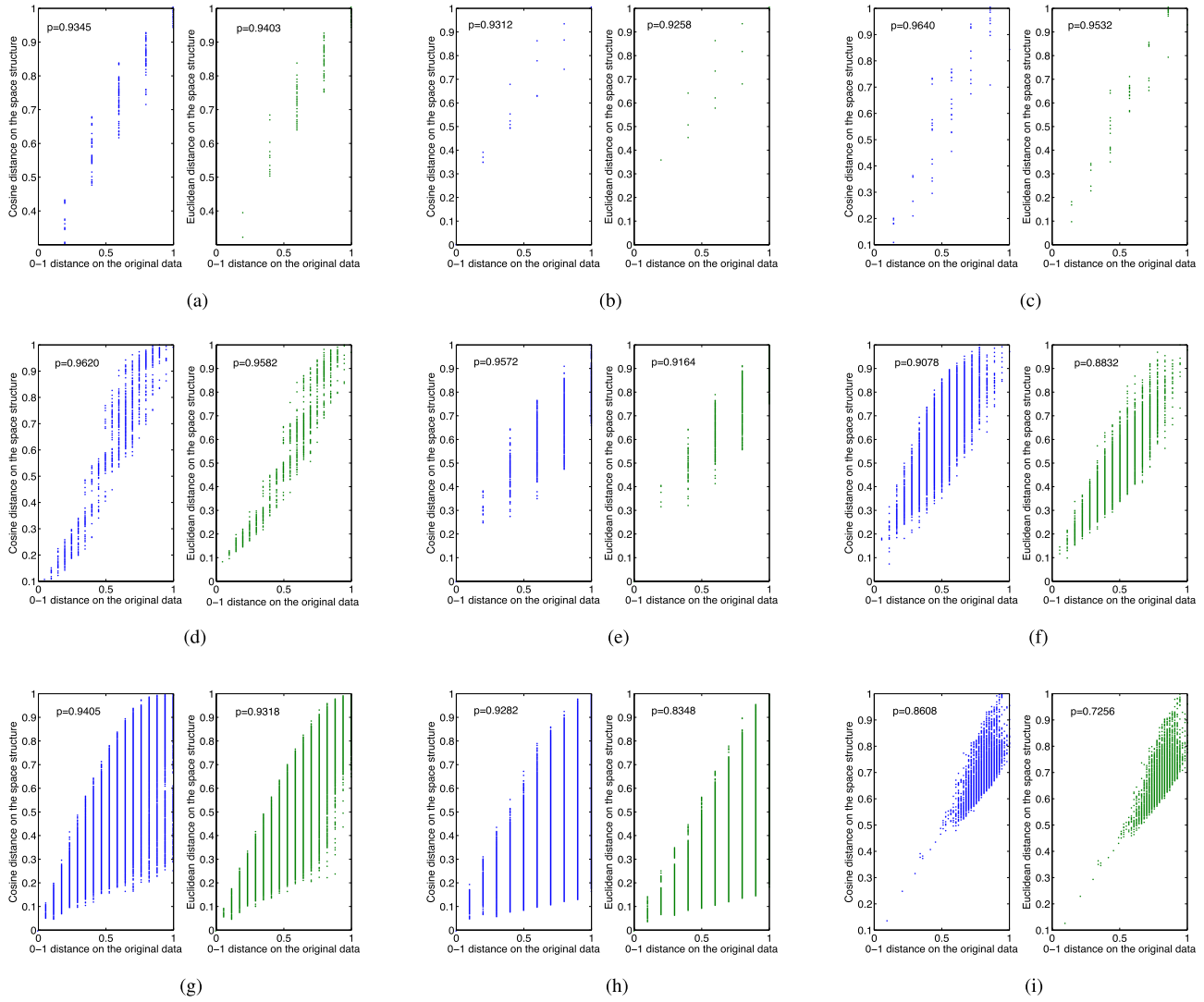


Fig. 2. Correlation analysis between the original categorical data space and the corresponding Euclidean space. (a) Fitting contact lenses. (b) Balloon. (c) Space shuttle autolanding. (d) Soybean-small. (e) Hayes-Roth–Hayes-Roth. (f) Lymphography Domain. (g) Vote. (h) Breast cancer. (i) Promoters.

is defined by

$$p_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{11}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively, and μ_X and μ_Y are the means of X and Y, respectively. The correlation analysis results and their Pearson correlation coefficients are shown in Fig. 2.

From Fig. 2, an obvious positive correlation between two variables X and Y can be observed. For example, the values of Pearson correlation coefficients between X in an original space and Y (induced by the cosine distance) in its corresponding

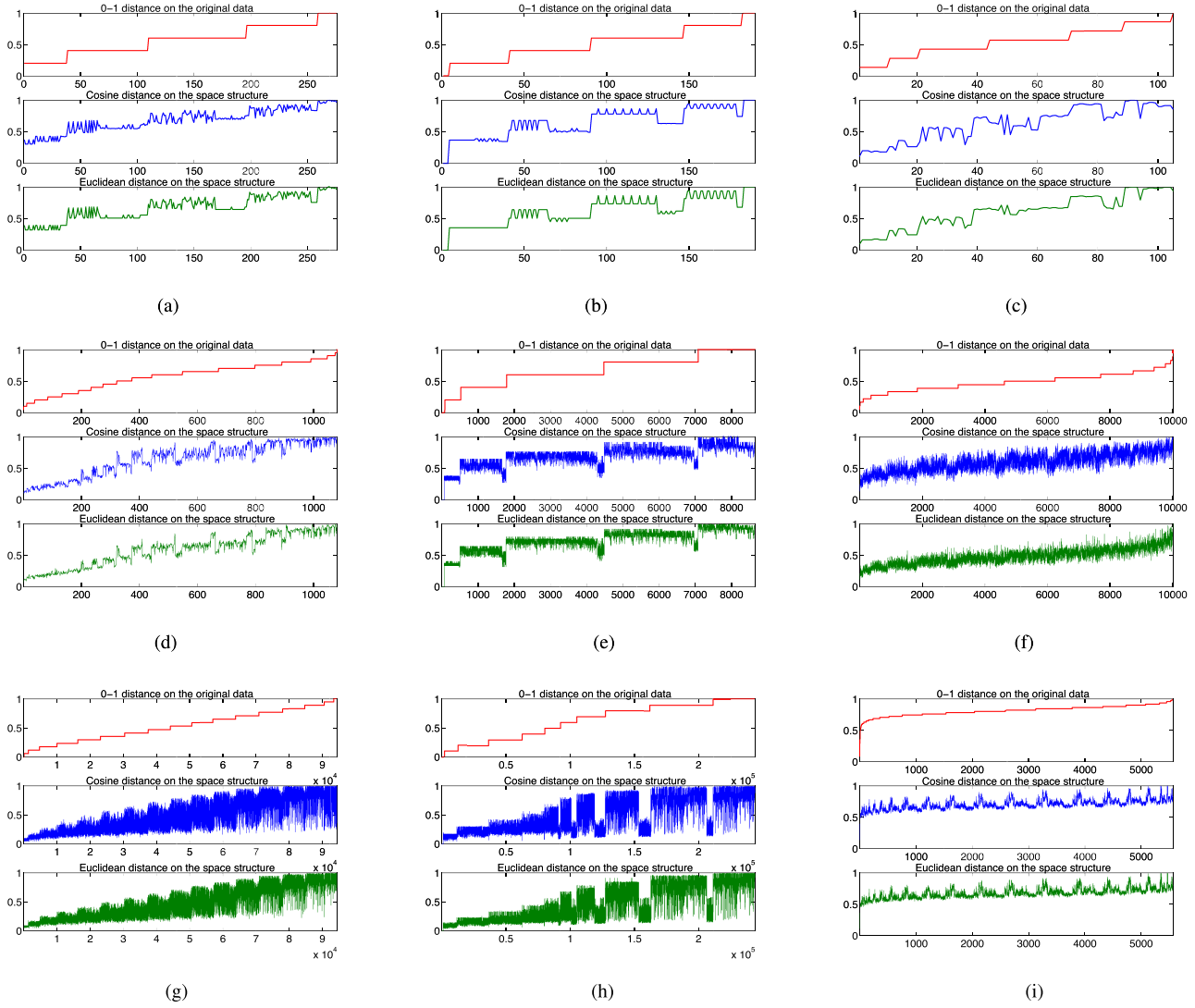


Fig. 3. Difference analysis between the original categorical data space and the corresponding Euclidean space. (a) Fitting contact lenses. (b) Balloon. (c) Space shuttle autolanding. (d) Soybean-small. (e) Hayes-Roth–Hayes-Roth. (f) Lymphography Domain. (g) Vote. (h) Breast cancer. (i) Promoters.

Euclidean space are beyond 0.9 on eight data sets. The strong positive correlation between X and Y means that a larger X is likely to result in a larger value of Y. This implies that the mapped Euclidean space almost reflect the original structure information of a categorical data set. Besides this advantage, it can be seen from Fig. 3 that the mapped Euclidean space can provide more distinct information. When two pairs of categorical objects cannot be distinguished with the 0–1 distance measure, they may be differentiated in its corresponding Euclidean space. This mechanism may improve the clustering performance of a categorical data set. Hence, we can effectively organize a categorical data set on its corresponding Euclidean space.

B. Clustering Performance Analysis

Many k-modes-type algorithms have been developed for the categorical data [4], [5], [11], [22], [31], [34], [39], [40], [51]. In order to verify the clustering performance of the proposed SBC algorithm, we employ four representative k-modes-type algorithms for organizing the categorical data, which are the classic k-mode algorithm [25], Chan's algorithm [11], Mkm-nof algorithm [4], and Mkm-ndm algorithm [4]. The objective of the following experiments is to show the performance of the proposed SBC algorithm for clustering a categorical data set.

We first consider a widely used index [accuracy (AC)] for evaluating the performance of a clustering algorithm, which is suitable for both the numeric data and the categorical data. This type of methods can be regarded as set matching. The main idea is to measure the shared set cardinality between two clustering results on a given data set. The set matching technique is an external criterion, in which external information-class labels need be used. It computes the best matches between clusters from each of the two clustering results and returns a value equaling to the total number of points shared between pairs of matched clusters. In this type of techniques, the set matching AC [58] is the simplest form, which is defined as

$$\mathrm{AC} = \sum_{i=1}^{k} \frac{\max\{n_{ij} : j \leq k'\}}{n} \tag{12}$$

TABLE VI
NOTATIONS ON TWO CLUSTERING RESULTS FROM THE SAME DATA SET

Clustering results	p_1	p_2	$\cdots$	$p_{k'}$	Sums
c_1	n_{11}	n_{12}	$\cdots$	$n_{1k'}$	b_1
c_2	n_{21}	n_{22}	$\cdots$	$n_{2k'}$	b_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k	n_{k1}	n_{k2}	$\cdots$	$n_{kk'}$	b_k
Sums	d_1	d_2	$\cdots$	$d_{k'}$	

where k and k' are the values from Table VI, and $n_{ij} = |c_i \cap p_j|$ is the number of common objects of cluster c_i and cluster p_j. Table VI is a contingency table including two clustering results $\{c_1, c_2, \ldots, c_k\}$ and $\{p_1, p_2, \ldots, p_{k'}\}$ on a data set with n objects, in which $\bigcup_{i=1}^{k} c_i = \bigcup_{j=1}^{k'} p_j = n$ and $c_i \cap c_{i'} = p_j \cap p_{j'} = \emptyset$, $1 \le i \ne i' $ and $1 \le j \ne j''$. The AC index will be equal to its maximum value 1 when the clustering result and its real partition are equivalent. If the clustering result is much closer to the true class distribution, the value of the AC will also be much higher.

Through using the above AC, we compare the proposed SBC algorithm with those four representative algorithms for the clustering categorical data, which are denoted by k-modes, Chan, Mkm-nf, and Mkm-ndm, respectively. In order to impartially compare, it is necessary to put each of the algorithms into a uniform environmental condition. The uniform environmental condition includes two sides. One is to set the number of clusters the same as the true number of classes of each of the nine data sets. When the clustering result induced by an algorithm is the closest to the true class distribution, we say that the algorithm has much better clustering performance for a given data set. The other is to randomly generate initial cluster prototypes of each of these algorithms, which is because the clustering performance of the k-modes-type algorithms is dependent on the initial cluster prototypes. Therefore, we need to investigate their performances from the viewpoint of statistics. To solve this issue, we run each of the five algorithms 100 times for each of the nine data sets, in which the initial cluster prototypes of each run are randomly regenerated. Furthermore, we compute the average value (it can be seen as an estimate of the expectation) of the ACs of the 100 clustering results and its standard deviation on each data set. These experimental results are shown in Table VII.

It is easy to see from Table VII that the proposed SBC algorithm is statistically much better than each algorithm compared for clustering the nine categorical data sets, in which the highest value of AC is underlined for each of the nine data sets. In these nine data sets, we can see that the two versions of the SBC algorithm get much higher ACs than the four representative baseline algorithms for seven data sets, while the Chan and Mkm-ndm algorithms only have the highest AC for one data set each, respectively. Even for the two better ones for the baselines, the clustering performance of the SBC algorithms is very close to those of the baseline algorithms. The clustering performance of these representative baseline algorithms has no clear good/bad relationships for the nine data sets. In addition, from Table VII, we also can observe that the SBC algorithms usually have much smaller standard deviation than each of the other four algorithms. This implies that compared with the four representative algorithm, the SBC algorithms have much better robustness for clustering a categorical data set. It is worth pointing out that the SBC algorithms can statistically and markedly improve the AC on most data sets. For example, the average value of ACs can be increased by $0.8878 - 0.7865 = 10.13\%$ on the data set Promoters, and can be increased by $0.7618 - 0.6589 = 10.29\%$ on the data set Lymphography, respectively.

In what follows, we consider another widely used index (adjusted rand index) for evaluating the performance of clustering algorithms from a different point of view, which is also suitable for both the numeric data and the categorical data. The adjusted rand index is again an external criterion that attempts to measure the similarity between two clustering results of objects in the same data set. The adjusted rand index (ARI) is formally defined as [18]

$$\mathrm{ARI} = \frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i \binom{b_i}{2}\sum_j \binom{d_j}{2}}{\frac{1}{2}\binom{n}{2}\left[\sum_i \binom{b_i}{2} + \sum_j \binom{d_j}{2}\right] - \sum_i \binom{b_i}{2}\sum_j \binom{d_j}{2}} \tag{13}$$

where the n_{ij}, b_i and d_j values from the contingency table (see Table VI). Like the index AC, if the clustering result is close to the true class distribution on the same data set, then the value of the index (ARI) is high. The values of the index ARI in these experimental results are shown in Table VIII.

It is easy to see from Table VIII that the proposed SBC algorithm is also statistically much better than each of the four algorithms compared for clustering the nine categorical data sets, in which the highest value of ARI is underlined. In these nine data sets, we can see that the two versions of the SBC algorithm all get much higher ARI values than the other four representative baseline algorithms in seven data sets, while each of algorithms Chan and Mkm-ndm get only the best in one data set each, respectively. It is worth noting that the SBC algorithm can statistically and clearly improve the ARI index on most of the data sets. For example, compared with the best results clustered by the four baseline clustering algorithms, the average value of ARI in the SBC-1 algorithm can be increased by $0.6072 - 0.3334 = 27.38\%$ on the data set Promoters, and in the SBC-2 algorithm can be increased by $0.4590 - 0.1987 = 26.03\%$ on the data set Balloon. For the two worse results, the clustering performance of the SBC algorithms is still very close to that of the best one of the other four algorithms. In addition, from Table VIII, it can be seen that the SBC algorithm usually has much smaller standard deviations than each of the other four algorithms. This implies that compared with the four representative baseline algorithm, the SBC algorithm has much better robustness for clustering a categorical data set.

C. Computational Performance Analysis

The purpose of this experiment is to test the computational efficiency of the SBC algorithm.

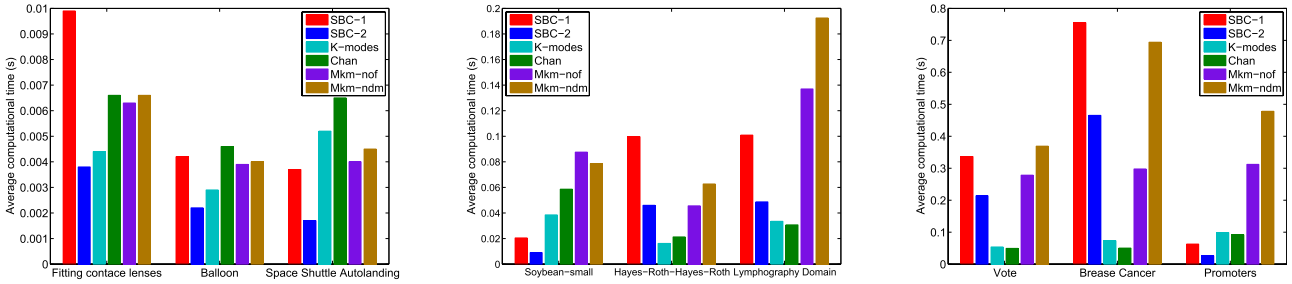
We first tested the average time of 100 runs of the six algorithms on the nine data sets, which are shown

TABLE VII
INDEX AC FROM FIVE DIFFERENT CLUSTERING METHODS FOR THE CATEGORICAL DATA ON THE NINE DATA SETS

Data sets	**SBC-1**	**SBC-2**	K-modes	Chan	Mkm-nof	Mkm-ndm
Fitting contact lenses	0.7288 ± 0.0016	0.7458 ± 0.0005	0.6417 ± 0.0013	0.6588 ± 0.0016	0.6813 ± 0.0026	0.6587 ± 0.0020
Balloon	0.7940 ± 0.0034	0.8485 ± 0.0001	0.6910 ± 0.0083	0.7045 ± 0.0080	0.7310 ± 0.0070	0.6710 ± 0.0071
Space Shuttle Autolanding	0.7140 ± 0.0074	0.6720 ± 0.0007	0.6240 ± 0.0010	0.6293 ± 0.0011	0.6140 ± 0.0007	0.6367 ± 0.0040
Soybean-small	0.9660 ± 0.0061	0.9589 ± 0.0079	0.9185 ± 0.0087	0.8353 ± 0.0096	0.7626 ± 0.0051	0.9483 ± 0.0078
Hayes-Roth-Hayes-Roth	0.4566 ± 0.0004	0.4630 ± 0.0009	0.4256 ± 0.0004	0.4782 ± 0.0020	0.4550 ± 0.0019	0.4329 ± 0.0026
Lymphography Domain	0.7618 ± 0.0009	0.7217 ± 0.0008	0.6252 ± 0.0030	0.5808 ± 0.0007	0.5801 ± 0.0000	0.6589 ± 0.0046
Vote	0.8783 ± 0.0001	0.8759 ± 0.0000	0.8604 ± 0.0001	0.8094 ± 0.0088	0.8715 ± 0.0027	0.8715 ± 0.0027
Breast Cancer	0.9293 ± 0.0000	0.9413 ± 0.0000	0.8608 ± 0.0112	0.7717 ± 0.0022	0.7697 ± 0.0000	0.9464 ± 0.0000
Promoters	0.8878 ± 0.0023	0.8106 ± 0.0001	0.6335 ± 0.0057	0.7865 ± 0.0028	0.7500 ± 0.0026	0.7043 ± 0.0121
Average value	0.7907	0.7820	0.6979	0.6949	0.6906	0.7254

TABLE VIII
INDEX ARI FROM FIVE DIFFERENT CLUSTERING METHODS FOR THE NINE DATA SETS

Data sets	**SBC-1**	**SBC-2**	K-modes	Chan	Mkm-nof	Mkm-ndm
Fitting contact lenses	0.2897 ± 0.0347	0.3582 ± 0.0207	0.0169 ± 0.0088	0.1009 ± 0.0232	0.1232 ± 0.0198	0.0621 ± 0.0118
Balloon	0.3262 ± 0.0215	0.4590 ± 0.0010	0.1356 ± 0.0247	0.1536 ± 0.0253	0.1987 ± 0.0234	0.0981 ± 0.0209
Space Shuttle Autolanding	0.1556 ± 0.0248	0.0566 ± 0.0028	-0.0050 ± 0.0017	0.0064 ± 0.0020	-0.0155 ± 0.0014	0.0262 ± 0.0110
Soybean-small	0.9400 ± 0.0193	0.9410 ± 0.0159	0.8247 ± 0.0288	0.6956 ± 0.0214	0.6330 ± 0.0038	0.9111 ± 0.0233
Hayes-Roth-Hayes-Roth	0.0133 ± 0.0001	0.0369 ± 0.0005	-0.0018 ± 0.0001	0.0377 ± 0.0010	0.0240 ± 0.0007	0.0071 ± 0.0011
Lymphography Domain	0.2721 ± 0.0020	0.1934 ± 0.0015	0.0627 ± 0.0053	0.0106 ± 0.0011	0.0080 ± 0.0000	0.1109 ± 0.0106
Vote	0.5715 ± 0.0001	0.5641 ± 0.0000	0.5187 ± 0.0005	0.4123 ± 0.0455	0.5599 ± 0.0127	0.5599 ± 0.0127
Breast Cancer	0.7331 ± 0.0001	0.7780 ± 0.0000	0.5395 ± 0.0792	0.2636 ± 0.0126	0.2487 ± 0.0000	0.7959 ± 0.0000
Promoters	0.6072 ± 0.0082	0.3802 ± 0.0003	0.0859 ± 0.0065	0.3334 ± 0.0061	0.2545 ± 0.0060	0.2084 ± 0.0310
Average value	0.4343	0.4186	0.2419	0.2238	0.2261	0.3089



(a) (b) (c)

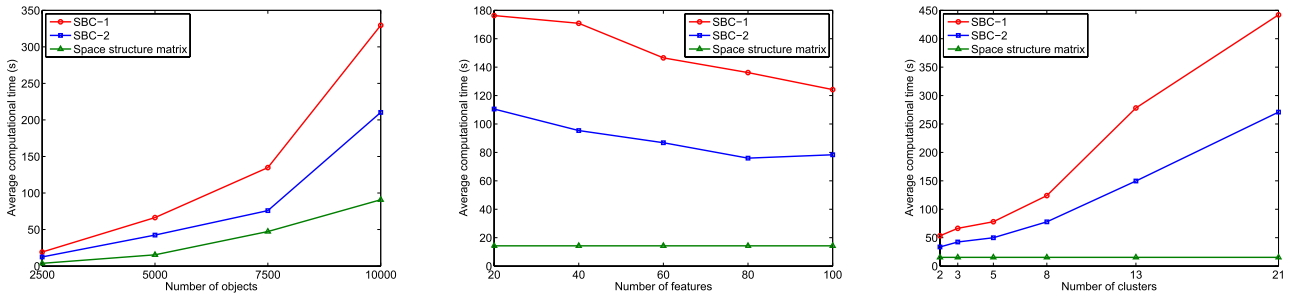
Fig. 4. Running time comparison of the six algorithms on the nine data sets. (a) Fitting contact lenses, Balloon and Space Shuttle Autolanding. (b) Soybean-small, Hayes-Roth-Hayes-Roth and Lymphography Domain. (c) Vote, Breast Cancer and Promoters.

in Fig. 4(a)–(c), respectively. Each column refers to the average time of 100 runs of an algorithm.

It can be seen from Fig. 4 that the SBC-1 algorithm with the cosine distance consumes much longer computational time than the four representative baseline algorithms for three data sets, and the SBC-2 algorithm with the Euclidean distance consumes statistically much shorter computational time for all the nine data sets. In particular, the SBC-2 algorithm spends the shortest computational time for five data sets. Although the time complexity of the SBC algorithm is $O(n^2)$, it is satisfying and acceptable from computational efficiency, especially for small-scale data sets.

In what follows, we tested three types of scalability of the SBC algorithm through artificial data sets. The first is the scalability against the number of objects (where the number of features is 30, and the number of clusters is 3), the second is the scalability against the number of features (where the number of objects is 5000, and the number of clusters is 3), and the third is the scalability against the number of clusters (where the number of objects is 5000, and the number of features is 30). Fig. 5 shows the computational time of the SBC algorithm for every clustering task.

From Fig. 5(a)–(c), one can see that the SBC-1 algorithm takes much more computational time than the SBC-2 algorithm, which is caused by distance calculations using cosine distance metric. From (6), one can see that the cosine distance metric needs to compute three Euclidean distances $\|\mathbf{x}\|$, $\|\mathbf{y}\|$ and $\mathbf{x}^T\mathbf{y}$, while (8) only computes



(a) (b) (c)

Fig. 5. Scalability test of the SBC algorithm. (a) Number of objects. (b) Number of features. (c) Number of clusters.

one Euclidean distance. For a given categorical data set, the computational time of mapping it into the corresponding Euclidean space is unchanged, which is only a very small percentage of the entire computational time. For example, in Fig. 5(c), mapping 5000 objects spend only almost 15 s, while the SBC-1 version and the SBC-2 version almost spent 440 and 270 s. Each of these two versions is one of the SBC family for the clustering categorical data.

The above experimental results thus tell us that the proposed algorithms cannot only guarantee to be convergent but also have better clustering results. The novel point of the SBC algorithm is the finding of the space structure for categorical data sets, which is very suitable for discovering the grouping structure inherent in the categorical data.

D. Relative Discussions

In this section, we summarize the advantages of the space structure-based clustering algorithm and analyze their reasons. From the experimental analyses in Sections V-B and V-C, one can see several advantages of the proposed SBC-type algorithms, which are listed as follows.

1) Each of the two versions of the SBC-type algorithms statistically obtains much better clustering results than the k-modes-type algorithms. In Section III, we have given the data-representation scheme of categorical data, which maps a set of categorical objects into a Euclidean space. This makes the difference between two categorical objects much finer-gained to measure than the 0–1 distance, which can be done by either a Euclidean distance or a cosine distance. In addition, the new cluster center induced by the mean of all objects in a cluster may have much better representativeness than the prototype updated by frequent feature values in each feature domain in a cluster. Hence, the SBC-type algorithms significantly improve the clustering performance of the k-modes-type algorithms when dealing with the categorical data sets.
2) The SBC algorithms possess much better clustering performance through selecting a corresponding dissimilarity measure according to the scale of a categorical data set. From the data-representation scheme of categorical data in Section III, we know that the number of dimensions of the Euclidean space mapped is equal to the number of objects for a given categorical data set. When the categorical data set is of large scale, the SBC algorithm with a cosine distance usually has a much better clustering performance. When the scale of the data set is small, the SBC algorithm with a Euclidean distance may be a good choice. Some modified versions of the original k-means algorithm have much better performances. Even so, the proposed SBC algorithms still show a very good performance for clustering a categorical data set.
3) The SBC-type algorithms can provide a semantically unified framework for clustering a mixed data set with the numeric data and the categorical data. In most existing algorithms, for clustering mixed data, the usual method is to fuse these two types of features with a variable weight. In essence, the numeric data and the categorical data are processed separately, which is not really unified semantically. In the proposed SBC algorithms in this paper, through using the new data-representation scheme, we can map all objects with the numeric and the categorical features into the same Euclidean space, and then use the same dissimilarity measure to cluster the mixed data set. In this paper, our objective is to solve the problem of the categorical data not having a clear space structure. We will study the interesting problem of clustering a mixed data set in our further work.

VI. Conclusion

Clustering categorical data are an important problem in such areas as pattern recognition, machine learning, data mining, and knowledge discovery. Many categorical clustering algorithms have been proposed, in which the k-modes-type algorithms are very representative because of their good performance. Due to the fact that the categorical data have no clear space structure like numeric data, these algorithms still have a great room for improving the clustering performance.

For clustering categorical data, we have proposed a new data-representation scheme for the categorical data, which is used to transform categorical objects into a Euclidean space with new dimensions. In this Euclidean space, each of the original features constructs one of the new dimensions. Based on the data-representation scheme, we have also given a general framework of the categorical data clustering algorithms (SBC). Through selecting various dissimilarity measures, the different versions of the SBC-type algorithms can be obtained.

In this paper, we have selected the Euclidean distance and the cosine distance to construct two versions. To verify the performance of the SBC-type algorithms, we have employed four representative algorithms of the k-modes-type algorithms as references or baselines. Experiments on the nine public categorical data sets show that the SBC-type algorithms have much better clustering performance than the k-modes-type algorithms. In addition, it is worth pointing out that it will be an interesting issue to cluster a data set mixed with the numeric data and the categorical data by exploiting the proposed SBC-type algorithms.

REFERENCES

[1] C. C. Aggarwal, C. Procopiuc, and P. S. Yu, "Finding localized associations in market basket data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 51–62, Jan./Feb. 2002.

[2] M. Alamuri, B. R. Surampudi, and A. Negi, "A survey of distance/similarity measures for categorical data," in *Proc. Int. Joint Conf. Neural Netw.*, 2014, pp. 1907–1914.

[3] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "LIMBO: Scalable clustering of categorical data," in *Proc. 9th Int. Conf. Extending Database Technol.*, 2004, pp. 123–146.

[4] L. Bai, J. Liang, C. Dang, and F. Cao, "The impact of cluster representatives on the convergence of the K-modes type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1509–1522, Jun. 2013.

[5] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel attribute weighting algorithm for clustering high-dimensional categorical data," *Pattern Recognit.*, vol. 44, no. 12, pp. 2843–2861, 2011.

[6] D. Barbara, Y. Li, and J. Couto, "COOLCAT: An entropy-based algorithm for categorical clustering," in *Proc. 11th Int. Conf. Inf. Knowl.*, 2002, pp. 582–589.

[7] D. Barbara and S. Jajodia, Eds., *Applications of Data Mining in Computer Security*. Dordrecht, The Netherlands: Kluwer, 2002.

[8] A. Baxevanis and F. Ouellette, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd ed. New York, NY, USA: Wiley, 2001.

[9] F. Cao, J. Liang, L. Bai, X. Zhao, and C. Dang, "A framework for clustering categorical time-evolving data," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 5, pp. 872–885, Oct. 2010.

[10] M. A. Carreira-Perpinan and W. Wang. (2013). "The K-modes algorithm for clustering." [Online]. Available: http://arxiv.org/abs/1304.6478

[11] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern Recognit.*, vol. 37, no. 5, pp. 943–952, 2004.

[12] E. Cesario, G. Manco, and R. Ortale, "Top-down parameter-free clustering of high-dimensional categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 12, pp. 1607–1624, Dec. 2007.

[13] H.-L. Chen, K.-T. Chuang, and M.-S. Chen, "On data labeling for clustering categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 11, pp. 1458–1472, Nov. 2008.

[14] M. R. Chmielewski and J. W. Grzymala-Busse, "Global discretization of continuous attributes as preprocessing for machine learning," *Int. J. Approx. Reasoning*, vol. 15, no. 4, pp. 319–331, 1996.

[15] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, nos. 1–2, pp. 155–176, 2003.

[16] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Mach. Learn.*, vol. 2, no. 2, pp. 139–172, 1987.

[17] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS—Clustering categorical data using summaries," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 73–83.

[18] M. A. Gluck and J. E. Corter, "Information uncertainty and the utility of categories," in *Proc. 7th Annu. Conf. Cognit. Sci. Soc.*, 1985, pp. 283–287.

[19] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognit.*, vol. 24, no. 6, pp. 567–578, 1991.

[20] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Inf. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[22] Z. He, S. Deng, and X. Xu, "Improving K-modes algorithm considering frequencies of attribute values in mode," in *Proc. Int. Conf. Comput. Intell. Security*, 2005, pp. 157–162.

[23] C.-C. Hsu, C.-L. Chen, and Y.-W. Su, "Hierarchical clustering of mixed data based on distance hierarchy," *Inf. Sci.*, vol. 177, no. 20, pp. 4474–4492, 2007.

[24] Q. Hu, Z. Xie, and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognit.*, vol. 40, no. 12, pp. 3509–3521, 2007.

[25] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in *Proc. SIGMOG Workshop Res. Issues Data Mining Knowl. Discovery*, 1997, pp. 1–8.

[26] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[27] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 446–452, Aug. 1999.

[28] X. Huang, Y. Ye, and H. Zhang, "Extensions of kmeans-type algorithms: A new clustering framework by integrating intracluster compactness and intercluster separation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1433–1446, Aug. 2014.

[29] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[30] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.

[31] D.-W. Kim, K. Y. Lee, D. Lee, and K. H. Lee, "A k-populations algorithm for clustering categorical data," *Pattern Recognit.*, vol. 38, no. 7, pp. 1131–1134, 2005.

[32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[33] S.-G. Lee and D.-K. Yun, "Clustering categorical and numerical data: A new procedure using multidimensional scaling," *Int. J. Inf. Technol. Decision Making*, vol. 2, no. 1, pp. 135–159, 2003.

[34] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul./Aug. 2002.

[35] M. Lee and W. Pedrycz, "The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features," *Fuzzy Sets Syst.*, vol. 160, no. 24, pp. 3590–3600, 2009.

[36] J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang, "Determining the number of clusters using information entropy for mixed data," *Pattern Recognit.*, vol. 45, no. 6, pp. 2251–2265, 2012.

[37] J. Liang, L. Bai, C. Dang, and F. Cao, "The K-means-type algorithms versus imbalanced data distributions," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 4, pp. 728–745, Aug. 2012.

[38] J. Liang, F. Wang, C. Dang, and Y. Qian, "A group incremental approach to feature selection applying rough set technique," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 294–308, Feb. 2014.

[39] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 503–507, Mar. 2007.

[40] M. K. Ng and L. Jing, "A new fuzzy k-modes clustering algorithm for categorical data," *Int. J. Granular Comput., Rough Sets Intell. Syst.*, vol. 1, no. 1, pp. 105–119, 2009.

[41] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, Dec. 2014.

[42] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Boston, MA, USA: Kluwer, 1991.

[43] W. Pedrycz and G. Vukovich, "Feature analysis through information granulation and fuzzy sets," *Pattern Recognit.*, vol. 35, no. 4, pp. 825–834, 2002.

[44] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[45] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation: An accelerator for attribute reduction in rough set theory," *Artif. Intell.*, vol. 174, nos. 9–10, pp. 597–618, 2010.

[46] Y. H. Qian, J. Y. Liang, W. Z. Wu, and C. Y. Dang, "Information granularity in fuzzy binary GrC model," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 2, pp. 253–264, Apr. 2011.

[47] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework," *Pattern Recognit.*, vol. 44, no. 8, pp. 1658–1670, 2011.

[48] Y. H. Qian, J. Y. Liang, and C. Y. Dang, "Incomplete multigranulation rough set," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 2, pp. 420–431, Mar. 2010.
[49] Y. Qian, H. Zhang, F. Li, Q. Hu, and J. Liang, "Set-based granular computing: A lattice model," *Int. J. Approx. Reasoning*, vol. 55, no. 3, pp. 834–852, 2014.
[50] D. N. Reshef *et al.*, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.
[51] O. M. San, V.-N. Huynh, and Y. Nakamori, "An alternative extension of the *k*-means algorithm for clustering categorical data," *Int. J. Appl. Math. Comput. Sci.*, vol. 14, no. 2, pp. 241–247, 2004.
[52] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Ordered Sets*, I. Rival, Ed. Dordrecht, The Netherlands: Reidel, 1982, pp. 445–470.
[53] N. Wrigley, *Categorical Data Analysis for Geographers and Environmental Scientists*. London, U.K.: Longman, 1985.
[54] W. Z. Wu, M. Zhang, H. Z. Li, and J. S. Mi, "Knowledge reduction in random information systems via Dempster–Shafer theory of evidence," *Inf. Sci.*, vol. 174, no. 3–4, pp. 143–164, 2005.
[55] J. Xie, B. Szymanski, and M. J. Zaki, "Learning dissimilarities for categorical symbols," in *Proc. 4th Workshop Feature Sel. Data Mining*, 2009, pp. 1058–1063.
[56] T. Xiong, S. Wang, A. Mayers, and E. Monga, "A new MCA-based divisive hierarchical algorithm for clustering categorical data," in *Proc. 9th IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 1058–1063.
[57] T. Xiong, S. Wang, A. Mayers, and E. Monga, "DHCC: Divisive hierarchical clustering of categorical data," *Data Mining Knowl. Discovery*, vol. 24, no. 1, pp. 103–135, 2012.
[58] Y. Yang, "An evaluation of statistical approaches to text categorization," *Inf. Retr.*, vol. 1, nos. 1–2, pp. 69–90, 1999.
[59] Y. Y. Yao, "Information granulation and rough set approximation," *Int. J. Intell. Syst.*, vol. 16, no. 1, pp. 87–104, 2001.
[60] J. Yu, "General C-means clustering model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1197–1211, Aug. 2005.
[61] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 90, no. 2, pp. 111–127, 1997.
[62] Z.-H. Zhou, "Three perspectives of data mining," *Artif. Intell.*, vol. 143, no. 1, pp. 139–146, 2003.
[63] (2010). *UCI Machine Learning Repository*. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html
[64] (2015). *Space structure and clustering of categorical data*. [Online]. Available: http://www.yuhuaqian.net/Cms_Data/Contents/SXU_YHQ/Media/code/ code_space_structure_and_clustering_of_categorical_data.zip

Yuhua Qian (M'10) received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multigranulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has authored over 50 articles on these topics in international journals.

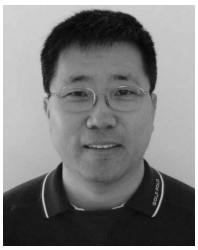
Prof. Qian served on the Editorial Board of the *International Journal of Knowledge-Based Organizations* and *Artificial Intelligence Research*. He has served as the Program Chair or Special Issue Chair of the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and the Conference on Industrial Instrumentation and Control, and a PC Member of many machine learning, data mining, and granular computing conferences.

Feijiang Li (S'14) received the B.S. degree from the School of Computer Science and Technology, Northeast University, Shenyang, China, in 2012. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Shanxi University, Taiyuan, China.

His current research interests include data mining and knowledge discovery.

Jiye Liang received the B.S. degree in computational mathematics and the Ph.D. degree in information science from Xi'an Jiaotong University, Xi'an, China.

He is currently a Professor with the School of Computer and Information Technology and the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University, Taiyuan, China. He has authored over 60 articles in international journals. His current research interests include artificial intelligence, granular computing, data mining, and knowledge discovery.



Bing Liu (F'14) received the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K.

He was with the National University of Singapore, Singapore. He is currently a Professor of Computer Science with the University of Illinois at Chicago, Chicago, IL, USA. He has authored extensively in his research topics in top conferences and journals. He has also authored two books entitled *Sentiment Analysis and Opinion Mining* (Morgan and Claypool Publishers) and *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data* (Springer). His current research interests include sentiment analysis and opinion mining, opinion spam detection, machine learning, data mining, and natural language processing.

Dr. Liu's work has been widely reported in the international press, including a front-page article in *The New York Times*. He has served as the Program Chair of the Conference on Knowledge Discovery and Data Mining, the International Conference on Data Mining, the Conference on Information and Knowledge Management, the Conference on Web Search and Data Mining, the Structures, Structural Dynamics, and Materials Conference, and the Pacific-Asia Conference on Knowledge Discovery and Data Mining, and the Area/Track Chair or Senior PC Member of many data mining, natural language processing, and Web technology conferences. He serves as the Chair of the ACM Special Interest Group on Knowledge Discovery and Data Mining.

Chuangyin Dang (SM'03) received the B.S. degree in computational mathematics from Shanxi University, Taiyuan, China, in 1983, the M.S. degree in applied mathematics from Xidian University, Xi'an, China, in 1986, and the Ph.D. degree in operations research/economics from the University of Tilburg, Tilburg, The Netherlands, in 1991.

He is currently a Professor with the City University of Hong Kong, Hong Kong. He is best known for the development of the D1-triangulation of the Euclidean space and the simplicial method for integer programming. His current research interests include computational intelligence, optimization theory and techniques, and applied general equilibrium modeling and computation.

Prof. Dang is a member of the Econometric Society, the Institute for Operations Research and the Management Sciences, and the Medical Protection Society.