Hindawi
Computational Intelligence and Neuroscience
Volume 2017, Article ID 3691316, 7 pages
https://doi.org/10.1155/2017/3691316

Research Article

A Global-Relationship Dissimilarity Measure for the *k*-Modes Clustering Algorithm

Hongfang Zhou, Yihui Zhang, and Yibin Liu

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

Correspondence should be addressed to Hongfang Zhou; zhouhf@xaut.edu.cn

Received 19 January 2017; Revised 4 March 2017; Accepted 19 March 2017; Published 28 March 2017

Academic Editor: Elio Masciari

The *k*-modes clustering algorithm has been widely used to cluster categorical data. In this paper, we firstly analyzed the *k*-modes algorithm and its dissimilarity measure. Based on this, we then proposed a novel dissimilarity measure, which is named as GRD. GRD considers not only the relationships between the object and all cluster modes but also the differences of different attributes. Finally the experiments were made on four real data sets from UCI. And the corresponding results show that GRD achieves better performance than two existing dissimilarity measures used in *k*-modes and Cao's algorithms.

1. Introduction

Clustering is an important technique in data mining, and its main task is to group the given data based on some similarity/dissimilarity measures [1]. Most clustering techniques use distances largely to measure the dissimilarity between different objects [2–4]. However, these methods work only on the data sets with numeric attributes, which limits their uses in solving categorical data clustering problems [5].

Some researchers have made great efforts to quantize relationships among different categorical attributes. Guha et al. [6] proposed a hierarchical clustering method termed ROCK, which can measure the similarity between a pair of objects [7]. In ROCK, the number of *Link* is computed as the number of common neighbors between two objects [8]. However, the following two deficiencies still exist: (1) two involved parameters (θ, k) must be assigned in advance and (2) the mass calculation is involved [9]. For these reasons, some researchers have generated some new algorithms like QROCK [10], DNNS [11], and GE-ROCK [12] to modify or improve the ROCK algorithm. To remove the numeric-only limitation of *k*-means algorithm, Huang et al. [13, 14] proposed the *k*-modes algorithm, which extends the *k*-means algorithm by using (1) a simple matching dissimilarity measure for categorical attributes; (2) modes in place of means for clustering; and (3) a frequency-related strategy to update modes to minimize the clustering costs [15]. In fact, the idea of simple matching has been used in many clustering algorithms, such as fuzzy *k*-modes algorithm [16], fuzzy *k*-modes algorithm with fuzzy centroid [17], and *k*-prototype algorithm [14]. However, simple matching often results in some low intradissimilarity clusters [18] and disregards of the dissimilarity hidden between the categorical values [19].

In this paper, a Global-Relationship Dissimilarity (GRD) measure for the *k*-modes clustering algorithm is proposed. This dissimilarity measure considers not only the relationships between the object and all cluster modes but also the differences of various attributes instead of simple matching. The clustering effectiveness of *k*-modes based on GRD (KBGRD) is demonstrated on four standard data sets from the UCI Machine Learning Repository [20].

The remainder of this paper is organized as follows: a detailed review of the dissimilarity measure used in *k*-modes is presented and analyzed in Section 2. In Section 3, the new dissimilarity measure GRD is proposed. Section 4 describes the details of KBGRD algorithm. Section 5 illustrates the performance and stability of KBGRD. Finally, a concluding remark is given in Section 6.

2. Related Works

2.1. Categorical Data. As is known to all, the structural data can be stored in a table, where each row represents a fact about

an object. And the practical data usually contains categorical attributes [21]. We firstly define the term "data set" [22].

Definition 1 (data set). A data set information system can be expressed as a quadruple IS = $\{U, A, V, f\}$, which is satisfied with

(1) $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty set of n data objects, which is named as a universe;

(2) $A = \{a_1, a_2, \ldots, a_m\}$ is a nonempty set of m categorical attributes;

(3) V is the union of all attribute domains, that is, $V = \bigcup_{j=1}^{m} V_{a_j}$, where $V_{a_j} = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)}\}$ is the value domain of attribute a_j, and it is finite and unordered; n_j is the number of categories of attribute a_j for $1 \le j \le m$;

(4) f is a mapping function $U \times A \to V$ which can be formally expressed as $(\forall x)(\forall y)((x \in U) \wedge (y \in A) \to f(x, y) \in V_{a_j})$ $(j = 1, 2, \ldots, m)$.

2.2. k-Modes Dissimilarity Measure. The k-modes clustering algorithm is an improvement of the k-means algorithm [4] by using a simple dissimilarity measure for categorical data. And it adopts a frequency-related strategy to update modes in the clustering to minimize the clustering costs. These extensions have excluded the numeric-only limitation existed in k-means algorithm and enable the clustering process to be used on large-size categorical data sets from real world database [22].

Definition 2. Let IS = $\{U, A, V, f\}$ be a categorical data set information system which is defined in Definition 1 and $a_j \in A$. For any object $x_i \in U$ and cluster mode z_l for $1 \le l \le k$, $\mathrm{Dis}_0(z_l, x_i)$ is the simple matching dissimilarity measure between object x_i and the mode z_l of the lth cluster which is defined as follows:

$$\mathrm{Dis}_0(z_l, x_i) = \sum_{j=1}^{m} \delta^{a_j}(z_l, x_i) \tag{1}$$

In (1), $\delta^{a_j}(z_l, x_i)$ can be expressed as $\delta^{a_j}(z_l, x_i) = \{1,\ f(z_l, a_j) \ne f(x_i, a_j);\ 0,\ f(z_l, a_j) = f(x_i, a_j)\}$.

There are nine objects $\{x_1, x_2, \ldots, x_9\}$ with four attributes $\{A_1, A_2, A_3, A_4\}$ and three initial cluster modes as shown in Table 1. For determining the appropriate cluster of x_1, it is required to compute the dissimilarity of x_1 and the three cluster modes. According to (1), $\mathrm{Dis}_0(c_1, x_1) = \mathrm{Dis}_0(c_2, x_1) = \mathrm{Dis}_0(c_3, x_1) = 1$. Therefore, it is impossible to determine exactly to which cluster the object x_1 should be assigned.

The dissimilarity between an object and a cluster mode should consider the relationships between the object and all cluster modes as well as the differences of various attributes. When the k-modes dissimilarity measure is computing dissimilarity of a certain attribute, it only simply matches this object with this mode and ignores the differences of various attributes. Such as attribute "A4" in Table 1, almost all of objects and cluster modes is "E"; "A4" should contribute more to dissimilarity than other attributes. However, the k-modes dissimilarity treats all attributes equally.

Table 1: An artificial data set.

Objects	A_1	A_2	A_3	A_4
x_1	A	B	A	E
x_2	A	A	B	D
x_3	C	A	A	E
Cluster 1 (c_1)	A	A	A	E
x_4	A	B	B	E
x_5	B	A	C	E
x_6	A	B	C	E
Cluster 2 (c_2)	A	B	C	E
x_7	A	A	A	E
x_8	D	C	B	E
x_9	C	C	A	E
Cluster 3 (c_3)	A	C	A	E

3. Global-Relationship Dissimilarity Measure

Definition 3. Let IS = $\{U, A, V, f\}$ be a categorical data set information system which is defined in Definition 1 and $a_j \in A$. For any object $x_i \in U$ and cluster mode z_l for $1 \le l \le k$, $\mathrm{Dis}(z_l, x_i)$ is the new dissimilarity measure between object x_i and the mode z_l of the lth cluster which is defined as

$$\mathrm{Dis}(z_l, x_i) = 1 - \frac{\mathrm{Sim}(z_l, x_i)}{m}. \tag{2}$$

In (2), m is the dimension number of data set and the similarity function $\mathrm{Sim}(z_l, x_i)$ is defined as follows:

$$\mathrm{Sim}(z_l, x_i) = \sum_{j=1}^{m} \varphi^{a_j}(z_l, x_i), \tag{3}$$

subject to

$$\varphi^{a_j}(z_l, x_i) = \begin{cases} 1 - \dfrac{S-1}{k}, & f(z_l, a_j) = f(x_i, a_j) \\ 0, & f(z_l, a_j) \ne f(x_i, a_j), \end{cases} \tag{4}$$

where k is the number of cluster modes, and

$$S = \sum_{l=1}^{k} s^{a_j}(z_l, x_i); \tag{5}$$

here $s^{a_j}(z_l, x_i)$ is satisfied with

$$s^{a_j}(z_l, x_i) = \begin{cases} 1, & f(z_l, a_j) = f(x_i, a_j) \\ 0, & f(z_l, a_j) \ne f(x_i, a_j). \end{cases} \tag{6}$$

As shown in Table 1, it is required to compute the dissimilarity of x_1 with three cluster modes for determining which cluster x_1 should be assigned to. According to (2)–(6), the following three ones can be got:

(1) $\mathrm{Dis}(c_1, x_1) = 1-(1/4)(1-(3-1)/3+0+1-(2-1)/3+1-(3-1)/3) = 8/12$.

(2) $\mathrm{Dis}(c_2, x_1) = 1-(1/4)(1-(3-1)/3+1-(1-1)/3+0+1-(3-1)/3) = 7/12$.

(3) $\mathrm{Dis}(c_3, x_1) = 1-(1/4)(1-(3-1)/3+0+1-(2-1)/3+1-(3-1)/3) = 8/12$.

Hence, x_1 can be assigned to cluster "2" definitely.

4. KBGRD Algorithm

In this section, we give the concrete procedure of the k-modes based on GRD (KBGRD) algorithm. In addition, the computational complexity of KBGRD is analyzed.

4.1. KBGRD Algorithm Description

Definition 4. Let IS $= \{U, A, V, f\}$ be a categorical data set information system which is defined in Definition 1 and $a_j \in A$. The k-modes algorithm uses the k-means paradigm to cluster categorical data. The objective function of the k-modes algorithm is defined as follows:

$$F(W, Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\mathrm{Dis}\left(z_l, x_i\right). \tag{7}$$

In (7), $\{w_{li} \in \{0,1\},\ 1 \le l \le k,\ 1 \le i \le n;\ \sum_{l=1}^{k} w_{li} = 1,\ 1 \le i \le n;\ 0 < \sum_{i=1}^{n} w_{li} < n,\ 1 \le l \le k\}$. Here $k(\le n)$ is a known cluster number; $W = [w_{li}]$ 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 is the lth cluster mode with categorical attributes $a_1, a_2, \ldots, a_m$.

4.2. Update and Convergence Analysis. The steps of the KBGRD algorithm are presented below. Here $Z^{(t)}$ and $W^{(t)}$ denote cluster modes and membership matrix at tth iteration, respectively.

(1) Randomly select k distinct objects from U as initial mode $Z^{(1)} = [z_1, z_2, \ldots, z_k]$. Determine $W^{(1)}$ such that $F(W^{(1)}, Z^{(1)})$ is minimized according to (8). Set $t = 1$.

(2) Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized according to (9). If $F(W^{(t)}, Z^{(t+1)}) == F(W^{(t)}, Z^{(t)})$, then stop; otherwise, go to step (3).

(3) Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized according to (8). 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).

In each iteration, W and Z are updated by the following formulae.

When Z is given, W is updated by (8) for $1 \le i \le n$ and $1 \le l \le k$.

$$w_{li} = \begin{cases} 1, & \mathrm{Dis}\left(\hat{z}_l, x_i\right) \le \mathrm{Dis}\left(\hat{z}_h, x_i\right),\ 1 \le h \le k \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

And when W is given, Z is updated as follows:

$$f\left(z_l, a_j\right) = a_j^{(r)} \in V_{a_j}, \tag{9}$$

where $\sum_{i=1, x_{ij}=a_j^{(r)}}^{n} w_{li}\varphi^{a_j}(z_l, x_i) \ge \sum_{i=1, x_{ij}=a_j^{(h)}}^{n} w_{li}\varphi^{a_j}(z_l, x_i)$, $1 \le h \le n_j$. Here, $V_{a_j} = \{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)}\}$; n_j is the number of categorical of attribute a_j for $1 \le j \le m$.

Now we consider the convergence of the KBGRD algorithm.

Theorem 5. *$F(W, \widehat{Z})$ is minimized when $Z = \widehat{Z}$ and W is updated by (8).*

Proof. For a given Z, we have $F(W, \widehat{Z}) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\mathrm{Dis}(z_l, x_i)$. The updating method of W is computing the minimized dissimilarity between objects and modes according to (8), and the dissimilarities of objects and modes are independent. So W is updated by (8) such that $F(W, \widehat{Z})$ is minimized. □

Theorem 6. *$F(\widehat{W}, Z)$ is minimized when $W = \widehat{W}$ and Z is updated by (9).*

Proof. For a given W, we have

$$\begin{aligned} F\left(\widehat{W}, Z\right) &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\mathrm{Dis}\left(z_l, x_i\right) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\left(1 - \frac{1}{m}\mathrm{Sim}\left(z_l, x_i\right)\right) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\mathrm{Sim}\left(z_l, x_i\right) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{li}\varphi^{a_j}\left(z_l, x_i\right) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{li}\varphi^{a_j}\left(z_l, x_i\right) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{j=1}^{m}\phi_{li}, \end{aligned} \tag{10}$$

where $\phi_{li} = \sum_{i=1}^{n} w_{li}\varphi^{a_j}(z_l, x_i)$. Note that all inner sums ϕ_{li} are nonnegative and independent. Then minimizing $F(\widehat{W}, Z)$ is equivalent to maximizing each inner sum. When $z_l = a_j^{(r)}$, according to (9), ϕ_{li} is maximized. So Z is updated by (9) such that $F(\widehat{W}, Z)$ is minimized. □

Input: data set U and initial cluster number k;
Output: clusters.
Sub function Cluster(U, modes)
Begin:
(1) **for** $l = 0$ to k // k is the number of clusters.
(2) **for** $i = 1$ to n // n is the number of objects.
(3) Calculating Dis(c_l, x_i) according to Eqs. (2)–(6);
(4) **end for**
(5) **end for**
(6) **if** (Dis(c_l, x_i) ≤ Dis(c_h, x_i)($0 \le h \le k$)){
(7) Classify ith object x_i into into lth cluster;}
End
Sub function Fun()
Begin:
(1) **for** $l = 0$ to k // k is the number of clusters.
(2) **for** $i = 1$ to n // n is the number of objects.
(3) Calculating SumDissimilarity according to Eq. (7);
(4) **return** SumDissimilarity;
end
Main function
Begin:
(1) Randomly choose k distinct objects $x_1, x_2, \ldots, x_k$ as initial modes from U;
(2) Cluster(U, modes);
(3) newDissimilarity = Fun(); // calculating the value of $F(W, Z)$.
(4) **Do**{
(5) oldDissimilarity = newDissimilarity;
(6) Update modes according to Eq. (9);
(7) Cluster(U, modes);
(8) newDissimilarity = Fun();}
(9) **while**{newDissimilarity != oldDissimilarity};
End

Pseudocode 1: Pseudocodes of KBGRD algorithm.

Theorem 7. *The KBGRD algorithm converges in a finite number of iterations.*

Proof. Firstly, we note that there are only a finite number ($N = \prod_{j=1}^{m} n_j$) of potential cluster mode. There are k^N possible kinds for k cluster modes; it is a finite number too.

Secondly, each possible mode appears at most once in the iteration process of KBGRD algorithm. If not, there exist t_1, t_2 ($t_1 < t_2$) such that $Z^{(t_1)} = Z^{(t_2)}$. According to Theorem 6, a given Z can obtain a certain W, that is, $Z^{(t_1)} \Rightarrow W^{(t_1)}$, $Z^{(t_2)} \Rightarrow W^{(t_2)}$. When $Z^{(t_1)} = Z^{(t_2)}$, we have $W^{(t_1)} = W^{(t_2)}$, that is, in the iteration of algorithm, occurring $F(W^{(t_1)}, Z^{(t_1)}) = F(W^{(t_1)}, Z^{(t_2)}) = F(W^{(t_2)}, Z^{(t_2)})$ at $t_1 < t_2$. However, if $F(W^{(t)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t)})$ or $(W^{(t+1)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t+1)})$, algorithm is stopped according to steps (2) and (3) of the KBGRD algorithm, that is, $F(W^{(t_1)}, Z^{(t_1)}) = F(W^{(t_1)}, Z^{(t_2)}) = F(W^{(t_2)}, Z^{(t_2)})$ never occurs.

So the KBGRD algorithm converges in a finite number of iterations. □

4.3. Pseudocodes and Complexity Analysis. The pseudocodes of KBGRD algorithm are presented in Pseudocode 1.

The major function of subfunction Cluster() is computing the dissimilarity between object and cluster mode and classifying the objects into the clusters whose dissimilarity is the minimum. The function of subfunction Fun() is computing the value of objective function.

In fact, main function is a controller, which controls the iterations of algorithm. We first choose k distinct objects as initial modes. Line 2 is the initialization of cluster; Line

Table 2: Data sets.

Data set	Attribute characteristics	#of data objects	# of attributes	# of class	Missing values
QSAR	Integer/real	1055	41	2	No
Chess	Categorical	3196	36	2	No
Mushroom	Categorical	8142	22	2	Yes (very few)
Nursery	Categorical	12960	8	5	No

3 computes original cluster result and "new Dissimilarity." Lines 4–9 are to iteratively update modes and clusters. And when "new Dissimilarity" is invariant, the iteration stops.

Referring to the pseudocodes as shown in Pseudocode 1, the computational complexity of KBGRD algorithm is analyzed as follows. We only consider the major computational steps.

We firstly consider the computational complexity of two subfunctions. The computational complexity for computing the dissimilarity is $O(k{\cdot}n{\cdot}m)$, where k is the number of modes, n is the number of objects in data set, and m is the dimension of data set. The computational complexity for assigning the ith object into the lth cluster is $O(k \cdot n)$. So the computational complexity for updating all clusters is $O(k \cdot n \cdot (m + 1))$, that is, $O(k \cdot n \cdot m)$. The computational complexity of computing objective function is $O(k \cdot n \cdot m)$.

Suppose that the iteration time is t and the whole computational cost of KBGRD algorithm is $t(O(k \cdot n \cdot m) + O(k \cdot n \cdot m)) = 2O(t \cdot k \cdot n \cdot m)$, that is, $O(t \cdot k \cdot n \cdot m)$. This shows that the computational cost is linearly scalable with the number of objects, the number of attributes, and the number of clusters.

5. Experimental Analysis

5.1. Experimental Environment and Evaluation Indexes. The experiments are conducted on a PC with an Intel i3 processor and 4 G byte memory running the Windows 7 operating system. All algorithms are coded by JAVA on Eclipse.

To evaluate the efficiency of clustering algorithm, the evaluation indexes *Accuracy (AC)* and *RandIndex* are employed in the experiments.

Let $C = \{C_1, C_2, C_3\}$ be the set of three classes in the data set and $C' = \{C'_1, C'_2, C'_3\}$ be the set of three clusters generated by the clustering algorithm. Given a pair of objects (X_i, X_j) in the data set, we refer to it as

(1) a if both objects belong to the same cluster in C and the same cluster in C';

(2) b if the two objects belong to the same cluster in C and two different clusters in C';

(3) c if the two objects belong to two different clusters in C and to the same cluster in C';

(4) d if both objects belong to two different clusters in C and two different clusters in C'.

Let S_1, S_2, S_3, and S_4 be the number of a, b, c, and d, *RandIndex* [23] is defined as follows:

$$RandIndex = \frac{S_1 + S_4}{S_1 + S_2 + S_3 + S_4}. \tag{11}$$

Table 3: Average *RandIndex* on four data sets for three algorithms.

	QSAR	Chess	Mushroom	Nursery
k-modes	0.513	0.5102	0.5101	0.6908
Cao's	0.5106	0.5136	0.5251	0.7895
KBGRD	0.5153	0.5229	0.5543	0.7933

Table 4: Average AC on four data sets for three algorithms.

	QSAR	Chess	Mushroom	Nursery
k-modes	0.5820	0.5720	0.5701	0.4786
Cao's	0.5944	0.5432	0.5895	0.5897
KBGRD	0.6042	0.6073	0.6634	0.5938

Accuracy (AC) is defined as follows:

$$AC = \frac{\sum_{i=1}^{k} a_i}{n}, \tag{12}$$

where k is the number of clusters, n is the number of objects, and a_i is the number of objects that are correctly assigned to the cluster C_i $(1 \le i \le k)$.

Four categorical data sets from the UCI Machine Learning Repository are used to evaluate the clustering performance, including QSAR Biodegradation (QSAR), Chess, Mushroom, and Nursery. The relative information about the data sets is tabulated in Table 2.

5.2. Experimental Results and Analysis. In the experiments, we compare KBGRD algorithm with the original k-modes and Cao's algorithm [24]. Three algorithms are sequentially run on all data sets. Each algorithm requires the number of modes *(ClusterNum)* as an input parameter. We randomly select distinct *ClusterNum* objects as initial cluster modes. The number of iteration of all algorithms is no more than 500.

Note that there are very few missing values in the Mushroom data set; we use optimal completion strategy to deal with missing values. In the optimal completion strategy, the missing values in data set are viewed as additional variables [25, 26].

Firstly, we set *ClusterNum* as the classes' number of the data set. The average *RandIndex* of ten times' experiments on four data sets for three algorithms is summarized in Table 3. The average *AC* of ten times' experiments on four data sets for three algorithms is summarized in Table 4. As shown in Tables 3 and 4, KBGRD achieves the highest *RandIndex* and *AC*. That is, it performs better than other algorithms under the same conditions.

Table 5: Average *RandIndex* of three algorithms on QSAR data set.

	10	15	20	25	30	35	Average
k-modes	0.4613	0.4608	0.4611	0.4596	0.4603	0.4584	0.4603
Cao's	0.4650	0.4610	0.4625	0.4608	0.4611	0.4593	0.4616
KBGRD	0.4658	0.4634	0.4628	0.4612	0.4620	0.4605	0.4626

Table 6: Average *RandIndex* of three algorithms on Chess data set.

	10	15	20	25	30	35	Average
k-modes	0.5016	0.5011	0.5008	0.5024	0.5032	0.5027	0.5020
Cao's	0.5041	0.5023	0.5014	0.5064	0.5045	0.5044	0.5039
KBGRD	0.5060	0.5090	0.5073	0.5072	0.5070	0.5075	0.5074

Table 7: Average *RandIndex* of three algorithms on Mushroom data set.

	10	15	20	25	30	35	Average
k-modes	0.5771	0.5641	0.5611	0.5622	0.5443	0.5404	0.5582
Cao's	0.5925	0.5644	0.5679	0.5790	0.5558	0.5638	0.5706
KBGRD	0.5932	0.5648	0.5731	0.5834	0.5678	0.5730	0.5759

Table 8: Average *RandIndex* of three algorithms on Nursery data set.

	10	15	20	25	30	35	Average
k-modes	0.6839	0.7061	0.6963	0.6875	0.6815	0.6942	0.6916
Cao's	0.7188	0.7071	0.6956	0.6989	0.6847	0.6982	0.7006
KBGRD	0.7195	0.7073	0.6967	0.7022	0.6957	0.6988	0.7034

In real world applications, the number of initial cluster modes is unknown. We evaluated clustering stability by setting different *ClusterNum* (10, 15, 20, 25, 30, and 35) for each data set and used *RandIndex* to evaluate clustering results. The average *RandIndex* of ten times' experiments on four data sets for three algorithms is summarized in Tables 5–8. And the last column shows the average clustering *RandIndex* of each algorithm on six *ClusterNum*. As shown in Tables 5–8, KBGRD achieves the highest *RandIndex*. That is to say, it performs better than other algorithms on four data sets. Additionally, KBGRD has the highest stability compared with other algorithms.

6. Conclusion

This paper analyzes the advantages and disadvantages of *k*-modes algorithms for categorical data. Based on this, we propose a novel dissimilarity measure (GRD) for clustering categorical data. This measure is used to improve the performance of the existing *k*-modes algorithm. The computational complexity of KBGRD algorithm has been analyzed which is linear with the number of data objects, attributes, and clusters. We have tested KBGRD algorithm on four real data sets from UCI. Experimental results have shown that KBGRD algorithm is effective and stable in clustering categorical data sets.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Science Foundation of China under the Grants of 61402363 and 61472319, Education Department of Shaanxi Province Key Laboratory Project under the Grant of 15JS079, Xi'an Science Program Project under the Grant of CXY1509(7), Beilin district of Xi'an Science and Technology Project under the Grant of GX1625, and CERNET Innovation Project under the Grant of NGLL20150707.

References

[1] A. Saha and S. Das, "Categorical fuzzy *k*-modes clustering with automated feature weight learning," *Neurocomputing*, vol. 166, pp. 422–435, 2015.

[2] H. Zhou, J. Guo, and Y. Wang, "A feature selection approach based on term distributions," *SpringerPlus*, vol. 5, no. 1, pp. 1–14, 2016.

[3] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Las Vegas, Nev, USA, August 2008.

[4] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, Berkeley, Calif, USA, 1967.

[5] F. Jiang, G. Liu, J. Du, and Y. Sui, "Initialization of K-modes clustering using outlier detection techniques," *Information Sciences*, vol. 332, pp. 167–183, 2016.

[6] S. Guha, R. Rastogi, and K. Shim, "Rock: a robust clustering algorithm for categorical attributes," *Information Systems*, vol. 25, no. 5, pp. 345–366, 2000.

[7] H. Zhou, J. Guo, Y. Wang, and M. Zhao, "A feature selection approach based on interclass and intraclass relative contributions of terms," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 1715780, 8 pages, 2016.

[8] I.-K. Park and G.-S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289–295, 2015.

[9] H. Zhou, X. Zhao, and X. Wang, "An effective ensemble pruning algorithm based on frequent patterns," *Knowledge-Based Systems*, vol. 56, no. 3, pp. 79–85, 2014.

[10] M. Dutta, A. K. Mahanta, and A. K. Pujari, "QROCK: a quick version of the ROCK algorithm for clustering of categorical data," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2364–2373, 2005.

[11] J. Yang, "A clustering algorithm using dynamic nearest neighbors selection model," *Chinese Journal of Computers*, vol. 30, no. 5, pp. 756–762, 2007.

[12] Q. Zhang, L. Ding, and S. Zhang, "A genetic evolutionary ROCK algorithm," in *Proceedings of the International Conference on Computer Application and System Modeling (ICCASM '10)*, pp. V12-347–V12-351, IEEE, Taiyuan, China, October 2010.

[13] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in *Proceedings of the SIGMOD*

Workshop Research Issues on Data Mining & Knowledge Discovery, pp. 1–8, 1998.

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

[15] Z. He, X. Xu, and S. Deng, "Attribute value weighting in k-modes clustering," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15365–15369, 2011.

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

[17] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1263–1271, 2004.

[18] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, "On the impact of dissimilarity measure in *k*-modes clustering algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 503–507, 2007.

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

[20] UCI Machine Learning Repository, 2016, https://archive.ics.uci.edu/ml/datasets.html.

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

[22] L. Bai and J. Liang, "The k-modes type clustering plus between-cluster information for categorical data," *Neurocomputing*, vol. 133, pp. 111–121, 2014.

[23] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 657–668, 2005.

[24] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the *k*-modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, no. 9, pp. 120–127, 2012.

[25] L. Zhang, W. Lu, X. Liu, W. Pedrycz, and C. Zhong, "Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values," *Knowledge-Based Systems*, vol. 99, pp. 51–70, 2016.

[26] H. Zhou, J. Li, J. Li, F. Zhang, and Y. Cui, "A graph clustering method for community detection in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 469, pp. 551–562, 2017.