ATLANTIS PRESS
Advances in Computer Science Research, volume 74
2nd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2017)

An Improved K-modes Clustering Algorithm Based on Intra-cluster and Inter-cluster Dissimilarity Measure

Hongfang Zhou [a], Yihui Zhang [b], Yibin Liu [c]

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

[a]zhouhf@xaut.edu.cn, [b]yifuzuonan@163.com, [c]ivuu329@126.com

Keywords: Clustering, categorical data, dissimilarity measure, k-modes algorithm.

Abstract. Categorical data clustering has attracted much attentions recently because most practical data contains categorical attributes. The k-modes algorithm, as the extension of the k-means algorithm, is one of the most widely used clustering algorithms for categorical data. In this paper, we firstly analyzed the limitations of two existing dissimilarity measures. Based on this, we proposed a novel dissimilarity measure--IID. IID considers the relationship between the object and all clusters as well as that within clusters. Finally the experiments are made on six benchmark data sets from UCI. And the corresponding results show that IID achieves better performance than two existing ones used in k-modes and KBGRD algorithms.

1. Introduction

Clustering is one of the most important techniques in data mining which partitions the given data into clusters on the basis of some similarity/dissimilarity measures. A number of clustering algorithms have been proposed over the past few decades [1-21]. Most clustering algorithms focus on numeric data [1-6]. However, they cannot be used in solving categorical data clustering problems [7] which is widely existed in real life.

A lot of algorithms have been proposed for clustering categorical data [7-21] in recent years. The *k*-modes [8] algorithm, as the extension of the *k*-means algorithm, is one of the most famous algorithms. And it extends the *k*-means algorithm by the following strategies [9]: (1) a simple matching dissimilarity measure for categorical data; (2) modes instead of means for clusters; and (3) a frequency-based method to update modes to minimize the clustering costs.

The *k*-modes clustering algorithm begins with an initial set of cluster modes and uses the alternating minimization method to minimize the clustering costs in finding clustering solutions [8]. The idea of simple matching dissimilarity has been widely used in many clustering algorithms [10-12]. In fact, the *k*-modes algorithm and its modified versions are well-known for their clustering efficiency and stability. However, the simple matching dissimilarity treats all categorical attributes equally, and it disregards the hidden dissimilarity between the categorical data [13]. Based on the idea of biological and genetic taxonomy and rough membership function, Cao et al. proposed a new dissimilarity measure for the *k*-modes algorithm [14]. Bai et al. used the between-cluster information to improve the effectiveness of *k*-modes type algorithms [8]. Qin et al. applied information theory into clustering in literature [15]. Zhou et al. proposed a global-relationship dissimilarity measure for the k-modes algorithm, which considers the relationships between the object and all cluster centers as well as the differences of various attributes [16].

In this paper, an intra-cluster and inter-cluster dissimilarity (IID) measure for *k*-modes clustering algorithm is proposed. It combines the relationship between the object and all clusters (inter-cluster) with that within clusters (intra-cluster) instead of simple matching. And then, a new clustering algorithm, KBIID, is proposed based on IID. Finally, experimental results on six standard data sets from the UCI Machine Learning Repository show that KBIID achieves better performance than two existing ones.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related works. In section 3, the new dissimilarity measure, IID, is proposed. Section 4 describes the details of KBIID.

Section 5 illustrates the performance and stability of KBIID. Finally, Section 6 presents a concluding remark.

2. Related Works

2.1 Categorical Data.

As everyone knows, the structural data is usually stored in a table, where each row represents a fact about an object. And categorical attributes are important parts of practical data set [22]. A categorical data set is defined as follows [16].

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

(1) $U = \{x_1, x_2, ..., x_n\}$ is a nonempty set of n data objects, called the universe;

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

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

(4) f is a mapping function $U \times A \rightarrow V$ such that $f(x_i, a_j) \in V_{a_j}$ for any $x \in U$ and $a_j \in A$.

2.2 K-modes Dissimilarity Measure.

The *k*-modes clustering algorithm is a well-known partition-based categorical data clustering algorithm. It improves the *k*-means algorithm by three extensions, which enable *k*-modes algorithm to process large-size categorical data sets from real world database [16].

Definition 2. Let $IS = \{U, A, V, f\}$ be a data set which is defined in Definition 1. For an object $x_i \in U$ and a cluster mode z_l of *l*th cluster for $1 \le l \le k$, $Dis_0(z_l, x_i)$ is the simple matching dissimilarity between the object x_i and the mode z_l which is defined as Eq.(1).

$$Dis_0(z_l, x_i) = \sum_{j=1}^{m} \delta^{a_j}(z_l, x_i) \tag{1}$$

In Eq.(1), $\delta^{a_j}(z_l, x_i)$ can be further expressed as $\delta^{a_j}(z_l, x_i) = \begin{cases} 1, f(z_l, a_j) \ne f(x_i, a_j) \\ 0, f(z_l, a_j) = f(x_i, a_j) \end{cases}$.

Table 1 shows a categorical data set with six objects $\{x_1, x_2, ..., x_6\}$ and three attributes $\{A_1, A_2, A_3\}$, and there are two initial cluster modes. Let us determine the object x_1 should be assigned to which clusters by simple matching dissimilarity measure. According to Eq.(1), we can get $Dis_0(z_1, x_1) = Dis_0(z_2, x_1) = 1$, which means x_1 has an undetermined assignment.

Table 1. An artificial data set

Objects	A1	A2	A3
x1	A	B	A
x2	A	A	B
x3	C	A	A
cluster1(z1)	A	A	A
x4	A	B	B
x5	C	A	C
x6	C	B	A
cluster2(z2)	C	B	A

2.3 Global-relationship Dissimilarity Measure.

The global-relationship dissimilarity (GRD) measure is an extension of simple matching dissimilarity measure, it thinks about not only the relationships between the object and all cluster modes but the differences of various attributes [16].

Definition 3. Let $IS=\{U,A,V,f\}$ be a data set which is defined in Definition 1. For an object $x_i \in U$ and a cluster mode z_l of lth cluster for $1 \le l \le k$, $Dis_1(z_l,x_i)$ is the global-relationship dissimilarity measure between the object x_i and the mode z_l which is defined as Eq.(2).

$$Dis_1(z_l,x_i)=1-\frac{Sim_1(z_l,x_i)}{m}. \tag{2}$$

In Eq.(2), m is the dimensional number of data set and the similarity function $Sim_1(z_l,x_i)$ is defined as follow.

$$Sim_1(z_l,x_i)=\sum_{j=1}^{m}\varphi^{a_j}(z_l,x_i), \tag{3}$$

subject to

$$\varphi^{a_j}(z_l,x_i)=\begin{cases}1-\dfrac{\left|\{z_h \mid f(z_h,a_j)=f(z_l,a_j), z_h \in Z\}\right|-1}{k}, & f(z_l,a_j)=f(x_i,a_j)\\ 0, & f(z_l,a_j)\neq f(x_i,a_j)\end{cases}, \tag{4}$$

where k is the number of cluster modes, Z is the set of cluster modes, and |*| is the number of *.

As shown in Table 1, let us determine the object x_1 should be assigned to which clusters by global-relationship dissimilarity measure. According to Eq.(2), we can get $Dis_1(z_1,x_1)=Dis_1(z_2,x_1)=\frac{1}{2}$, which means x_1 has an undetermined assignment.

3. Intra-cluster and Inter-cluster Dissimilarity Measure

K-modes simple matching dissimilarity measure ignores the differences between various attributes. GRD has made some improvements compared to the simple matching dissimilarity measure. However, these two dissimilarities ignore the intra-cluster information. As we all know, cluster mode only represents partial information of a cluster. Based on this, we proposed a novel dissimilarity measure termed the intra-cluster and inter-cluster dissimilarity measure (IID).

Definition 4. Let $IS=\{U,A,V,f\}$ be a data set which is defined in Definition 1. For an object $x_i \in U$ and a cluster mode z_l of lth cluster for $1 \le l \le k$, $Dis(z_l,x_i)$ is the intra-cluster and inter-cluster dissimilarity measure between the object x_i and the mode z_l which is defined as Eq.(5).

$$Dis(z_l,x_i)=1-\frac{Sim(z_l,x_i)}{m} \tag{5}$$

In Eq.(5), m is the dimensional number of data set and the similarity function $Sim(z_l,x_i)$ is defined as follows.

$$Sim(z_l,x_i)=\sum_{j=1}^{m}\vartheta^{a_j}(z_l,x_i), \tag{6}$$

subject to

$$\vartheta^{a_j}(z_l,x_i)=\begin{cases}\alpha_{l_i}^{a_j}\cdot\beta_{l_i}^{a_j}, & f(z_l,a_j)=f(x_i,a_j)\\ \dfrac{1}{k}\cdot\beta_{l_i}^{a_j}, & f(z_l,a_j)\neq f(x_i,a_j)\end{cases}, \tag{7}$$

here

$$\alpha_{l_i}^{a_j}=1-\frac{\left|\{z_h \mid f(z_h,a_j)=f(z_l,a_j), z_h \in Z\}\right|-1}{k}, \tag{8}$$

$$\beta_{l_i}^{a_j}=\frac{\left|\{x_h \mid f(x_h,a_j)=f(x_i,a_j), x_h \in c_l\}\right|}{|c_l|}, \tag{9}$$

where k is the number of cluster modes, Z is the set of cluster modes, c_l is the lth cluster, and |*| is the number of *. In fact, $\alpha_{l_i}^{a_j} = \frac{1}{k}$ if and only if $x_i^{a_j}$ is equal to all of $z_h^{a_j}$ ($z_h \in Z$); $\alpha_{l_i}^{a_j} = 1$ if and only if $x_i^{a_j}$ is equal to z_l and not equal to all the other modes. β is the probability of $x_i^{a_j}$ and $x_h^{a_j}$ is equal in c_l. The function of β is to display all of internal information of c_l.

As shown in Table 1, let us determine the object x_1 should be assigned to which clusters by intra-cluster and inter-cluster dissimilarity measure. According to Eq.(5)-Eq.(9), we can get $Dis(z_1, x_1) = \frac{11}{18}$, $Dis(z_2, x_1) = \frac{12}{18}$. Hence, x_1 can be assigned to cluster '1' definitely.

4. KBIID Algorithm

4.1 KBIID Algorithm Description.

Definition *5.* Let $IS = \{U, A, V, f\}$ be a data set which is defined in Definition 1. The objective function of the k-modes algorithm is defined as follows.

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

In Eq.(10), it is subjected to $\begin{cases} w_{li} \in \{0,1\}, & 1 \le l \le k, 1 \le i \le n \\ \sum_{l=1}^{k} w_{li} = 1, & 1 \le i \le n \\ 0 < \sum_{i=1}^{n} w_{li} < n, & 1 \le l \le k \end{cases}$. Here $k(\le n)$ is a known cluster number; $W = [w_{li}]$ is a $n \times k$ matrix. w_{li} is a binary variable, which presents whether the object x_i belongs to the lth cluster; $w_{li} = 1$ if x_i belongs to the lth cluster and $w_{li} = 0$ otherwise; $Z = [z_1, z_2, ..., z_k]$, and z_l is the lth cluster mode.

4.2 Updating and Convergence Analysis.

The steps of the KBIID algorithm are shown below. Here $Z^{(t)}$ and $W^{(t)}$ denote the cluster modes and the membership matrix at the tth iteration respectively.

Randomly choose k distinct objects $z_1, z_2, ..., z_k$ from U as initial modes $Z^{(1)}$; Determine $W^{(1)}$ such that $F(W^{(1)}, Z^{(1)})$ is minimized according to Eq.(11); Set t =1.

Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized according to Eq.(12). If $F(W^{(t)}, Z^{(t+1)}) == F(W^{(t)}, Z^{(t)})$, then stop; otherwise, go to step (3).

Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized according to Eq.(11). 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, the membership matrix W and the modes Z are updated by the following formulae.

When Z is given, W is updated by Eq.(11) for $1 \le i \le n$ and $1 \le l \le k$.

$$w_{li} = \begin{cases} 1, & Dis(\hat{z}_l, x_i) \le Dis(\hat{z}_h, x_i), 1 \le h \le k \\ 0, & otherwise \end{cases} \tag{11}$$

And when W is given, Z is updated by Eq.(12).

$$f(z_l, a_j) = a_j^{(r)} \in V_{a_j} \tag{12}$$

where $\sum_{\substack{i=1 \\ x_{ij}=a_j^{(r)}}}^{n} w_{li}\alpha^{a_j}(z_l,x_i) \geq \sum_{\substack{i=1 \\ x_{ij}=a_j^{(h)}}}^{n} w_{li}\alpha^{a_j}(z_l,x_i), 1 \leq h \leq n_j$. Here, $V_{a_j}=\{a_j^{(1)},a_j^{(2)},...,a_j^{(n_j)}\}$, n_j is the possible value number of the attribute a_j for $1 \leq j \leq m$.

Now we consider the convergence of the KBIID algorithm.

Theorem *1.* $F(W,\hat{Z})$ is minimized when $Z=\hat{Z}$ is fixed and W is updated by Eq.(11).

Proof. For a given Z, we have $F(W,\hat{Z})=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}Dis(z_l,x_i)$. The updating strategy of membership matrix W is calculating the minimized dissimilarity between objects and modes according to Eq.(11), and the dissimilarities of objects and modes are independent. So W is updated by Eq.(11) such that $F(W,\hat{Z})$ is minimized.

Theorem 2. $F(\hat{W},Z)$ is minimized when $W=\hat{W}$ is fixed and Z is updated by Eq.(12).

Proof. For a given W, we have:

$$\begin{aligned}
F(\hat{W},Z) &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}Dis(z_l,x_i) \\
&= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}(1-\frac{1}{m}Sim(z_l,x_i)) \\
&= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}Sim(z_l,x_i) \\
&= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{li}\vartheta^{a_j}(z_l,x_i) \\
&= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{li}\vartheta^{a_j}(z_l,x_i) \\
&= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li} - \frac{1}{m}\sum_{l=1}^{k}\sum_{j=1}^{m}\phi_{li}
\end{aligned}$$

where $\phi_{li}=\sum_{i=1}^{n} w_{li}\vartheta^{a_j}(z_l,x_i)$. Note that all inner sums ϕ_{li} are nonnegative and independent. So minimizing $F(\hat{W},Z)$ is equivalent to maximizing each inner sum ϕ_{li}.

We have $\vartheta^{a_j}(z_l,x_i)=\begin{cases}\alpha_{l_i}^{a_j}\cdot\beta_{l_i}^{a_j}, & f(z_l,a_j)=f(x_i,a_j) \\ \frac{1}{k}\cdot\beta_{l_i}^{a_j}, & f(z_l,a_j)\neq f(x_i,a_j)\end{cases}$. According to Eq.(9), β is fixed when $W=\hat{W}$ is fixed. Therefore, maximizing ϕ_{li} is equivalent to maximizing α. When $z_l=a_j^{(r)}$, ϕ_{li} is maximized according to Eq.(12). So Z is updated by Eq.(12) such that $F(\hat{W},Z)$ is minimized.

Theorem 3. The KBIID 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 possible cluster modes.

Secondly, each possible mode appears at most once in the iteration process of KBIID algorithm. If not, there exist t_1,t_2 ($t_1<t_2$) such that $Z^{(t_1)}=Z^{(t_2)}$. According to Theorem 2, a given Z can obtain a certain W, i.e. $Z^{(t_1)}\Rightarrow W^{(t_1)}, Z^{(t_2)}\Rightarrow W^{(t_2)}$. 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 $F(W^{(t+1)},Z^{(t+1)})=F(W^{(t)},Z^{(t+1)})$, algorithm is stopped according

to the step (2) and step (3) of the KBIID algorithm, i.e. $F(W^{(t_1)}, Z^{(t_1)}) = F(W^{(t_1)}, Z^{(t_2)}) = F(W^{(t_2)}, Z^{(t_2)})$ never occurs.

So the KBIID algorithm converges in a finite number of iterations.

5. Experimental Analysis

5.1 Experimental Environment and Evaluation Metrics.

The experiments are carried on a PC with an Intel i3 processor and 4G byte memory running the Windows 7 Operating System. All algorithms are coded by JAVA.

The evaluation indexes *Accuracy* (*AC*) and *RandIndex* are employed to evaluate the performance of clustering algorithm 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 of the 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 the same cluster in C';

(4) *d* if the 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* respectively, *RandIndex* [23] is defined by Eq.(13).

$$RandIndex = \frac{S_1 + S_4}{S_1 + S_2 + S_3 + S_4} \tag{13}$$

Accuracy(*AC*) is defined by Eq.(14).

$$AC = \frac{\sum_{i=1}^{k} a_i}{n} \tag{14}$$

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

Six categorical data sets from the UCI Machine Learning Repository are employed to evaluate the clustering performance, which are Zoo, Hayes-Roth (Hayes), Breast-Cancer (Breast), Chess, Mushroom and Nursery. The relative information about the data sets is tabulated in Table 3.

Table 3. Data sets

Data Set	#of Data Objects	# of Attributes	# of Class	Missing Values
Zoo	101	16	7	No
Hayes	132	4	3	No
Breast	286	9	2	No
Chess	3196	36	2	No
Mushroom	8142	22	2	Yes(very few)
Nursery	12960	8	5	No

5.2 Experimental Results and Analysis.

In the experiments, we compare KBIID algorithm with the original *k*-modes and KBGRD algorithm [16]. Three algorithms are run on all data sets respectively. We randomly select *ClusterNum* different objects as initial cluster modes, and set the number of iterations of all algorithms is no more than 500. We use optimal completion strategy [24] to deal with the very few missing values in Mushroom data set.

Firstly, we set ClusterNum as the classes' number of the data set. The mean *RandIndex* of ten times experiments on six data sets for three algorithms are summarized in Table 4. The mean *AC* of ten times experiments on six data sets for three algorithms are summarized in Table 5. As shown in Table 4 and Table 5, KBIID achieves the highest *RandIndex* and *AC.* That is, KBIID performs better than other two algorithms under the same conditions.

In practice, the number of initial cluster modes is unknown. We evaluated clustering stability by setting different *ClusterNum* (10, 15, 20, 25, 30, 35) for each data set, and used *RandIndex* to evaluate clustering results. The mean *RandIndex* of ten times experiments on six data sets for three algorithms are summarized in Table 6-Table 11. And the last column shows the mean *RandIndex* of each algorithm on six *ClusterNum.* As shown in Table 6-Table 11, KBIID achieves the highest *RandIndex.* That is to say, it performs better than other algorithms on six data sets. Therefore, KBIID has the highest stability when compared with other algorithms.

Table 4. Mean *RandIndex* on six data sets for three algorithms

	Zoo	Hayes	Breast	Chess	Mushroom	Nursery
k-modes	0.8186	0.5437	0.4989	0.5102	0.5101	0.6908
KBGRD	0.8341	0.5452	0.4989	0.5230	0.5544	0.7895
KBIID	0.8447	0.5465	0.4991	0.5254	0.5552	0.7896

Table 5. Mean *AC* on four data sets for three algorithms

	Zoo	Hayes	Breast	Chess	Mushroom	Nursery
k-modes	0.5842	0.3712	0.5105	0.5720	0.5701	0.4786
KBGRD	0.5842	0.4242	0.5175	0.6076	0.6635	0.5897
KBIID	0.6040	0.4355	0.5210	0.6130	0.6712	0.5897

Table 6. Mean *RandIndex* of three algorithms on Zoo data set

	10	15	20	25	30	35	mean
k-modes	0.8368	0.8160	0.8194	0.8137	0.8011	0.7968	0.8140
KBGRD	0.8475	0.8224	0.8218	0.8139	0.8014	0.7999	0.8178
KBIID	0.8507	0.8329	0.8255	0.8147	0.8032	0.8018	0.8215

Table 7. Mean *RandIndex* of three algorithms on Hayes data set

	10	15	20	25	30	35	mean
k-modes	0.6201	0.6445	0.6506	0.6514	0.6534	0.6537	0.6456
KBGRD	0.6207	0.6467	0.6532	0.6528	0.6551	0.6588	0.6479
KBIID	0.6208	0.6473	0.6564	0.6548	0.6579	0.6590	0.6494

Table 8. Mean *RandIndex* of three algorithms on Breast data set

	10	15	20	25	30	35	mean
k-modes	0.4388	0.4339	0.4336	0.4296	0.4285	0.4267	0.4319
KBGRD	0.4506	0.4343	0.4388	0.4313	0.4365	0.4319	0.4372
KBIID	0.4597	0.4381	0.4395	0.4341	0.4369	0.4324	0.4401

Table 9. Mean *RandIndex* of three algorithms on Chess data set

	10	15	20	25	30	35	mean
k-modes	0.5016	0.5011	0.5008	0.5024	0.5032	0.5027	0.5020
KBGRD	0.5031	0.5051	0.5069	0.5072	0.5070	0.5096	0.5065
KBIID	0.5128	0.5067	0.5098	0.5134	0.5075	0.5101	0.5101

Table 10. Mean *RandIndex* of three algorithms on Mushroom data set

	10	15	20	25	30	35	mean
k-modes	0.5771	0.5641	0.5611	0.5622	0.5443	0.5404	0.5582
KBGRD	0.5933	0.5655	0.5731	0.5829	0.5683	0.5724	0.5759
KBIID	0.6308	0.5728	0.5873	0.6061	0.5852	0.5733	0.5926

Table 11. Mean *RandIndex* of three algorithms on Nursery data set

	10	15	20	25	30	35	mean
k-modes	0.6839	0.7061	0.6963	0.6876	0.6834	0.6942	0.6919
KBGRD	0.7195	0.7073	0.6967	0.6936	0.6957	0.6942	0.7012
KBIID	0.7195	0.7073	0.6967	0.6936	0.6957	0.6942	0.7012

6. Conclusion

This paper analyzes the advantages and disadvantages of the simple matching dissimilarity measure in *k*-modes algorithm and the global-relationship dissimilarity measure for categorical data. Based on this, we propose a novel dissimilarity measure (IID) for clustering categorical data. This measure is used to improve the performance of the existing *k*-modes algorithm. We have tested

KBIID algorithm on six real data sets from UCI. Experimental results show that KBIID algorithm is effective and stable in clustering categorical data.

Acknowledgements

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. Jain, R. Dubes. Algorithms for clustering data. Prentice Hall. 1988.
[2]. A. Likas, N. Vlassis, J. J. Verbeek. The global k-means clustering algorithm. Pattern Recognition. Vol. 36(2003) No. 2, p. 451-461.
[3]. A. Y. Ng , M. I. Jordan, Y. Weiss. On Spectral Clustering: Analysis and an algorithm. Proceedings of Advances in Neural Information Processing Systems. (2002), p. 849-856.
[4]. M. A. Rahman, M. Z. Islam. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Systems. Vol. 71 (2014), p. 345-365.
[5]. J. Xie, H. Gao, W. Xie, et al. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K -nearest neighbors. Information Sciences. Vol. 354(2016), p. 19-40.
[6]. D. Huang, J. H. Lai, C. D. Wang. Ensemble clustering using factor graph. Pattern Recognition. Vol. 50 (2015) No. C, p. 131-142.
[7]. L. Bai, J. Liang. The k -modes type clustering plus between-cluster information for categorical data. Neurocomputing. Vol. 133 (2014), p. 111-121.
[8]. M. K. Ng, M. J. Li, J. Z. Huang, et al. On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Transactions on Pattern Analysis & Machine Intelligence. Vol. 29 (2007) No 3, p. 503-507.
[9]. Z. He, X. Xu, S. Den. Attribute value weighting in k-modes clustering. Expert Systems with Applications. Vol. 38 (2011) No. 12, p. 15365-15369.
[10]. Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining & Knowledge Discovery. Vol. 2 (1998) No. 3, p. 283-304.
[11]. Z. Huang, M. K. Ng. A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems. Vol. 7 (1999) No. 4, p. 446-452.
[12]. D. W. Kim, K. H. Lee, D. Lee. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters. Vol. 25 (2004) No. 11, p. 1263-1271.
[13]. C. C. Hsu, C. L. Chen, Y. W. Su. Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences. Vol. 177 (2007) No. 20, p. 4474-4492.
[14]. F. Cao, J. Liang, D. Li, et al. A dissimilarity measure for the k-modes clustering algorithm. Knowledge-Based Systems. Vol. 26 (2011) No. 9, p.120-127.
[15]. H. Qin, X. Ma, T. Herawan, et al. MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. Knowledge-Based Systems. Vol. 67 (2014) No. 3, p. 401-411.
[16]. H. Zhou, Y. Zhang, Y. Liu. A Global-Relationship Dissimilarity Measure for the k-Modes Clustering Algorithm. Computational Intelligence and Neuroscience. Vol. 2017 (2017).
[17]. T. R. L. D. Santos, L. E. Zárate. Categorical data clustering: What similarity measure to recommend?. Expert Systems with Applications. Vol. 42 (2015) No. 3, p. 1247-1260.
[18]. H. Zhou, J. Li, J. Li, et al. A graph clustering method for community detection in complex networks. Physica A Statistical Mechanics & Its Applications. Vol. 469 (2017), p. 551-562.
[19]. I. Saha, J. P. Sarkar, U. Maulik. Ensemble based rough fuzzy clustering for categorical data. Knowledge-Based Systems. Vol. 77 (2015), p. 114-127.

[20]. A. Saha, S. Das. Categorical fuzzy k-modes clustering with automated feature weight learning. Neurocomputing. Vol. 166 (2015) No. C, p. 422-435.
[21]. K. C. Gowda, E. Diday. Symbolic clustering using a new dissimilarity measure. Pattern Recognition. Vol. 24 (1991) No. 6, p. 567-578.
[22]. J. Z. Huang, M. K. Ng, H. Rong, et al. Automated variable weighting in k-means type clustering. IEEE Transactions on Pattern Analysis & Machine Intelligence. Vol. 27 (2005) No. 5, p. 657-668.
[23]. H. Zhou, J. Guo, Y. Wang. A feature selection approach based on term distributions. Springerplus. Vol. 5 (2016) No. 1, p. 1-14.
[24]. L. Zhang, W. Lu, X. Liu, et al. Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. Knowledge-Based Systems. Vol. 99 (2016) No. C, p. 51-70.
[25]. H. Zhou, J. Guo, Y. Wang, et al. A feature selection approach based on interclass and intraclass relative contributions of terms. Computational Intelligence & Neuroscience. Vol. 2016 (2016) No. 6, p. 1-8.
[26]. H. Zhou, X. Zhao, X. Wang. An effective ensemble pruning algorithm based on frequent patterns. Knowledge-Based Systems. Vol. 56 (2014) No. 3, p. 79-85.