Knowledge-Based Systems 26 (2012) 120–127

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A dissimilarity measure for the *k*-Modes clustering algorithm

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

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
[b] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 1 December 2010
Received in revised form 20 July 2011
Accepted 20 July 2011
Available online 27 July 2011

Keywords:
Categorical data clustering
k-Modes algorithm
Rough membership function
Dissimilarity measure
Genetic taxonomy

ABSTRACT

Clustering is one of the most important data mining techniques that partitions data according to some similarity criterion. The problems of clustering categorical data have attracted much attention from the data mining research community recently. As the extension of the *k*-Means algorithm, the *k*-Modes algorithm has been widely applied to categorical data clustering by replacing means with modes. In this paper, the limitations of the simple matching dissimilarity measure and Ng's dissimilarity measure are analyzed using some illustrative examples. Based on the idea of biological and genetic taxonomy and rough membership function, a new dissimilarity measure for the *k*-Modes algorithm is defined. A distinct characteristic of the new dissimilarity measure is to take account of the distribution of attribute values on the whole universe. A convergence study and time complexity of the *k*-Modes algorithm based on new dissimilarity measure indicates that it can be effectively used for large data sets. The results of comparative experiments on synthetic data sets and five real data sets from UCI show the effectiveness of the new dissimilarity measure, especially on data sets with biological and genetic taxonomy information.

1. Introduction

The widespread use of computer and information technology has made extensive data collection in business, manufacturing and medical organizations a routine task. This explosive growth in stored data has generated an urgent need for new techniques that can transform the vast amounts of data into useful knowledge. Data mining is, perhaps, most suitable for this need [1].

In data mining, clustering is a widely used technique that partitions a data set consisting of n points embedded in an m-dimensional space into k distinct clusters such that the data points within the same cluster are more similar to each other than to data points in other clusters. Essentially, clustering is performed according to the similarity or dissimilarity among objects. The similarity or dissimilarity between two objects is generally based on difference in corresponding attribute values. In a clustering algorithm, the similarity or dissimilarity between objects is usually measured by a distance function. The smaller the distance, the more similar the two objects are considered to be. The most commonly used distance function is the Minkowski metric that includes the Euclidean distance and the Manhattan distance as special cases. However, Minkowski metric is only for numeric data, and it becomes difficult to capture this notion for categorical attributes. Therefore, the computation of similarity or dissimilarity between categorical data objects in unsupervised learning is very important.

Roughly speaking, the current approaches to similarity or dissimilarity measures of categorical values can be classified into the following four categories.

1.1. Simple matching approaches

Simple matching is a common approach in which comparison of two identical categorical values yields a difference of *zero* while comparison of two distinct categorical values yields a difference of *one*. The idea of simple matching has been utilized in many categorical clustering algorithms in [2–4] including the *k*-Modes algorithm and its variants, such as the *k*-Modes algorithm [5], fuzzy *k*-Modes algorithm [6], fuzzy *k*-Modes algorithm with fuzzy centroid [7], and *k*-prototype algorithm [8]. However, simple matching often results in clusters with weak intrasimilarity [9], and disregards the similarity hidden between categorical values [10]. A valuable dissimilarity measure is introduced for *k*-Modes clustering algorithm by Ng et al. [9], that extends the standard simple matching approach by taking account of the frequency of mode components in the current cluster.

1.2. Co-occurrence approaches

Gibson et al. [11] pointed out that the similarity of two categorical values refers to their co-occurrence with a common value or a

* Corresponding author.
E-mail addresses: cfy@sxu.edu.cn (F. Cao), ljy@sxu.edu.cn (J. Liang), lidy@sxu.edu.cn (D. Li), sxbailiang@126.com (L. Bai), mecdang@cityu.edu.hk (C. Dang).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.07.011

set of values. Some algorithms based on the idea of co-occurrence of categorical value are proposed. ROCK [12] uses the concept of a *link* to measure the similarity between categorical patterns. A measure $Link(p_i, p_j)$ is defined as the number of common neighbors between two patterns p_i and p_j for ROCK. The objective of the algorithm is to group together patterns that have a relatively large number of links. CACTUS [13] defines the similarity between patterns by looking at the support of two attribute values, which is the frequency of two values appearing in patterns together. The higher the support is, the more similarity the two attribute values are. Based on the co-occurrence probability of two categorical values, a distance metric is presented by Ahamad and Dey [14] for mixed numeric and categorical data clustering. The significance of an attribute towards the clustering process is also hidden in this distance metric.

1.3. Probabilistic approaches

Conceptual clustering algorithms in [15,16] for handling data with categorical values use conditional probability estimate to define relations between groups or clusters. The *category utility* (CU) measure [17] defines a probability matching strategy to measure the usefulness of a class in correctly predicting feature values, and the idea is also adopted in the system COBWEB [18] and its derivatives [19,20]. AUTOCLASS [21] assumes a classical finite mixture distribution model on the data and uses Bayesian method to derive the most probable class distribution for the data with some prior information. Wong et al. proposed a discrete-valued data clustering algorithm DECA [22], which has been used in bimolecular data clustering [23]. Chiu et al. [24] proposed a distance measure for dealing with mixed-type attributes in large database. Their techniques are derived from a probabilistic model in which the distance between two clusters is equivalent to the decrease in the log-likelihood function as for merging. An entropy-based categorical data clustering algorithm COOLCAT [25] finds a set of initial clusters, and then incrementally adds patterns to the clusters according to the criterion that minimizes the expected entropy of the clusters.

1.4. Distance hierarchy approaches

Distance hierarchy [10,26–28] extends the concept of hierarchy [29] by associating each link with a weight representing a distance to facilitate the computation of distance between categorical values. However, such an approach needs domain experts to incorporate knowledge, e.g., the general-specific relationship, for facilitating further mining of clustering results.

Other similarity or dissimilarity measures for categorical data clustering algorithms include Gower's similarity coefficient [30], Goodall's similarity measure [31,32], and Gowda's dissimilarity measure [33].

Rough set theory introduced by Pawlak [34] is a kind of symbolic machine learning technology for categorical value information systems with uncertainty information [35–37]. In recent years, rough set theory has received a great deal of attention in some of the clustering literature. Parmar et al. [38] proposed a new algorithm min-min-roughness (MMR) for clustering categorical data based on rough set theory, which has the ability to handle the uncertainty in the clustering process. By defining outlying partition similarity based on the concept of rough set, outliers on the key attribute subset rather than on the full dimensional attribute set of data set can be mined [39]. Using the notion of rough membership function from rough set theory, Jiang et al. [40] defined the rough outlier factor for outlier detection. Chen and Wang [41] presented an improved clustering algorithm based on rough set and Shannon's Entropy theory. Herawan et al. [42] proposed a new technique called maximum dependency attributes for selecting clustering attribute based on rough set theory by taking into account the dependency of attributes of the database. Cao et al. [43] proposed a framework for clustering categorical time-evolving data based on rough membership function and sliding window technique.

In this paper, the limitations of simple matching dissimilarity measure and Ng's dissimilarity measure are revealed using some illustrative examples. Based on the idea of biological and genetic taxonomy, we introduce a new rough membership-based dissimilarity measure between two objects by taking into account the distribution of attribute values in the universe. Furthermore, the dissimilarity measure between a mode of a cluster and an object is given by improving Ng's dissimilarity measure. The proposed dissimilarity measure is utilized in the *k*-Modes algorithm, the algorithm convergence is proved and the corresponding time complexity is analyzed as well. The scalability and clustering effectiveness of the *k*-Modes algorithm with the proposed dissimilarity measure are demonstrated on synthetic data sets and five standard data sets downloaded from the UCI Machine Learning Repository [44], respectively.

The organization of the rest of this paper is as follows. In Section 2, two kinds of new dissimilarity measures, between two objects and between a mode and an object, for the *k*-Modes algorithm are defined. Convergence and time complexity of the *k*-Modes algorithm with the proposed measure are analyzed in Section 3. In Section 4, experimental results on the synthetic data sets and five real data sets demonstrate the scalability and effectiveness of the *k*-Modes algorithm based on the new dissimilarity measure by comparison with other dissimilarity measures. Section 5 concludes the paper.

2. New dissimilarity measures for the *k*-Modes algorithm

In this section, we first review some basic concepts of rough set theory, such as categorical information system, indiscernibility relation and rough membership function. Then, a new dissimilarity measure between two objects is defined based on rough membership function. Furthermore, a new dissimilarity measure between the mode of a cluster and an object is introduced for the *k*-Modes algorithm.

The data is assumed to be in a table, where each row (tuple) represents facts about an object. A data table is also called an information system. Objects in the real world are sometimes described by categorical information system.

Definition 1. Formally, a categorical information system is a quadruple $IS = (U, A, V, f)$, where:

U, the nonempty set of objects, called the universe;
A, the nonempty set of attributes;
V, the union of all attribute domains, i.e., $V = \bigcup_{a \in A} V_a$, where V_a is the domain of attribute a and it is finite and unordered;
f: $U \times A \to V$, a mapping called an information function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$.

Definition 2. Let $IS = (U, A, V, f)$ be a categorical information system and $P \subseteq A$, a binary relation $IND(P)$, called indiscernibility relation, is defined as:

$$IND(P) = \{(x, y) \in U \times U | \ \forall\ a \in P,\ f(x, a) = f(y, a)\}.$$

Informally two objects are indiscernible in the context of a set of attributes if they have the same values for those attributes. $IND(P)$ is an equivalence relation on U and $IND(P) = \bigcap_{a \in P} IND(\{a\})$.

The relation $IND(P)$ induces a partition of U, denoted by $U/IND(P) = \{[x]_P | x \in U\}$, where $[x]_P$ denotes the equivalence class determined by x with respect to P, i.e., $[x]_P = \{y \in U | (x,y) \in IND(P)\}$.

Definition 3 [34]. Let $IS = (U, A, V, f)$ be a categorical information system, $P \subseteq A$ and $X \subseteq U$. The rough membership function $\mu_X^P : U \to [0,1]$ is defined as:

$$\mu_X^P(x) = \frac{|[x]_P \bigcap X|}{|[x]_P|}.$$

The rough membership function quantifies the degree of relative overlap between the set X and the equivalence class $[x]_P$ to which x belongs.

In classical set theory, either an element belongs to a set or it does not. The corresponding membership function is the characteristic function of the set, i.e., the function takes values 1 and 0, respectively. However, the rough membership function takes values between 0 and 1.

2.1. A new dissimilarity measure between two objects

Now, we take account of the similarity between two objects with respect to some attribute by virtue of the rough membership function.

Definition 4. Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $a \in P$ and $x, y \in U$, a similarity measure between objects x and y with respect to a is defined as:

$$Sim_a(x,y) = \mu_{\{y\}}^{\{a\}}(x) = \frac{\left|[x]_{\{a\}} \bigcap \{y\}\right|}{\left|[x]_{\{a\}}\right|}.$$

In Definition 4, the domain of the rough membership function is an object y of U, not the universe U. The degree of relative overlap between the object x and the object y means the similarity between the object x and the object y.

The similarity $Sim_a(x,y)$ can be also described as:

$$Sim_a(x,y) = \frac{f(x,a) \equiv f(y,a)}{\sum_{z \in U} f(x,a) \equiv f(z,a)},$$

where

$$f(x,a) \equiv f(y,a) = \begin{cases} 1, & \text{if } f(x,a) \neq f(y,a), \\ 0, & \text{otherwise}. \end{cases}$$

For similarity measure $Sim_a(x,y)$, the following is valid.

Property 1. *Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $a \in P$ and $x, y \in U$, we have:*

(1) *Symmetry $Sim_a(x,y) = Sim_a(y,x)$.*
(2) *Minimum $Sim_a(x,y) = 0$ iff $f(x,a) \neq f(y,a)$.*
(3) *Maximum $Sim_a(x,y) = 1$ iff $x = y$ and $|[x]_{\{a\}}| = 1$.*

By (2) in Property 1, we know that $Sim_a(x,y) = 0$ if and only if the objects x and y belong to different equivalence classes of a, in other words, x and y can be distinguished by a. Given an attribute a, (3) in Property 1 means that an object is only similar to itself and the similarity achieves to the maximum value 1, only when it is different from every other object with respect to attribute a.

Since $Sim_a(x,y) = \frac{1}{|[x]_{\{a\}}|}$ if $y \in [x]_{\{a\}}$, any two objects in the same equivalence class have the same similarity, also the similarity value monotonically decreases with the size of the equivalence class. Formally, we have the following proposition.

Proposition 1. *Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $x, y, u, v \in U$ and $a \in P$, if $f(x,a) = f(y,a)$, $f(u,a) = f(v,a)$ and $|[x]_{\{a\}}| \leqslant |[u]_{\{a\}}|$, then $Sim_a(x,y) \geqslant Sim_a(u,v)$.*

Goodall's similarity measure [31] defined for handling nominal and numeric features, was first proposed for biological and genetic taxonomy problems, where unusual characteristics shared by biological entities is often attributed to closely related genetic information resulting in these entities being classified into the same species. Accordingly, a pair of entities x and y is considered more similar to each other than another pair of entities u and v, if and only if the entities x and y exhibit a greater match in feature values that are less common in the population. In other words, similarity among objects is decided by the uncommonality of their feature value matches.

In clustering problems, the above-mentioned principle can help us define more cohesive, tight clusters where objects grouped into the same cluster are likely to share special and characteristic feature values. By this token, the similarity in Definition 4 is in accord with the principle in Genetic Taxonomy that "the similarity value is a function of the uncommonality of the feature value within the universe".

Following is a definition of dissimilarity between two objects over several attributes defined in terms of similarity between objects.

Definition 5. Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $x, y \in U$, the dissimilarity measure between x and y with respect to P is defined as:

$$d_P(x,y) = \sum_{a \in P} d_a(x,y),$$

where

$$d_a(x,y) = 1 - Sim_a(x,y).$$

For the dissimilarity measure $d_P(x,y)$, it is easy to prove the following properties.

Property 2. *Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $x, y, z \in U$, we have:*

(1) *Symmetry $d_P(x,y) = d_P(y,x)$.*
(2) *Nonnegativity $d_P(x,y) \geqslant 0$.*
(3) *Triangle Inequality $d_P(x,y) + d_P(y,z) \geqslant d_P(x,z)$.*

Property 2 shows that the dissimilarity measure $d_P(x,y)$ is a distance metric.

Since first published in 1997, the k-Modes algorithm[5] has become an important technique for solving categorical data clustering problems in various domains. The k-Modes algorithm extends the k-Means algorithm by use of a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. These extensions have removed the numeric-only limitation of the k-Means algorithm and allows the k-Means clustering process to be used to efficiently cluster large categorical data sets in real world databases.

The simple matching dissimilarity measure has been used frequently. It is defined following.

Definition 6 ([5]). Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. For any $x, y \in U$, the simple matching dissimilarity measure between x and y with respect to P is defined as:

$$D_P(x,y) = \sum_{a \in P} D_a(x,y),$$

where

$$D_a(x,y)=\begin{cases}1, & \text{if } f(x,a)\neq f(y,a),\\ 0, & \text{otherwise.}\end{cases}$$

It is easy to verify that the function D_P defines a metric space on the set of categorical objects. Traditionally, the simple matching approach is often used for binary attributes that are obtained from categorical attributes. However, the distance between objects computed with the simple matching dissimilarity measure often results in clusters with weak intrasimilarity and disregards the similarity embedded in the categorical values.

Let us consider the following example that reveals the limitation of the simple matching dissimilarity measure.

Example 1. In Table 1, five objects x_1, x_2, x_3, x_4, x_5 are randomly selected from the lenses data set [44], where $P=\{$*age of the patient*, *spectacle prescription*, *astigmatic*, *tear production rate*$\}$ is the attribute set, and the distinguished attribute *class* denotes the outcome of classification.

Let x_3 and x_5 be the initial cluster centers. By using the simple matching dissimilarity measure, we have $D_P(x_2,x_3)=D_P(x_2,x_5)=1$. This means that we cannot determine to which cluster x_2 should be assigned.

However, by Definition 5

$$d_P(x_2,x_3)=1-1/3+1-1/3+1+1-1/5=94/30,$$

and

$$d_P(x_2,x_5)=1+1-1/3+1-1/2+1-1/5=89/30.$$

It follows that the object x_2 can be assigned to the second cluster "*class* = 2" properly.

2.2. A new dissimilarity measure between a mode and an object

Taking account of the frequency of mode components in the current cluster, Ng et al. [9] introduced a valuable dissimilarity measure into the *k*-Modes clustering algorithm. Ng's dissimilarity measure $Dis_P(z_l,x_i)$ between a categorical object x_i and the mode of a cluster z_l with respect to P is defined as:

$$Dis_P(z_l,x_i)=\sum_{a\in P}Dis_a(z_l,x_i),$$

where

$$Dis_a(z_l,x_i)=\begin{cases}1, & \text{if } f(z_l,a)\neq f(x_i,a),\\ 1-m_a, & \text{otherwise.}\end{cases}$$

$$m_a=\frac{|\{x_i|f(x_i,a)=f(z_l,a),x_i\in c_l\}|}{|c_l|},$$

and $|c_l|$ is the number of objects in the lth cluster.

For the *k*-Modes algorithm with Ng's dissimilarity measure [9], the simple matching dissimilarity measure is still used in the first iteration.

Now, we introduce a new dissimilarity measure into the *k*-Modes algorithm by using $Sim_a(x,y)$ defined in Definition 4.

Table 1
Some objects of lenses data set.

Objects	Age of the patient	Spectacle prescription	Astigmatic	Tear production rate	Class
x_1	1	1	2	2	1
x_2	1	2	1	2	2
x_3	1	2	2	2	1
x_4	2	1	2	2	1
x_5	2	2	1	2	2

Definition 7. Let $IS=(U,A,V,f)$ be a categorical information system, and $P\subseteq A$. For any $x_i\in U$ and cluster mode z_l for $1\leqslant l\leqslant k$, the new dissimilarity measure between x_i and z_l with respect to P is defined as:

$$NDis_P(z_l,x_i)=\sum_{a\in P}NDis_a(z_l,x_i),$$

where

$$NDis_a(z_l,x_i)=1-Sim_a(z_l,x_i)\times m_a.$$

As opposed to Ng's dissimilarity measure, the similarity $Sim_a(z_l,x_i)$ between x_i and z_l is included in the proposed measure $NDis_a(z_l,x_i)$. Therefore, $NDis_a(z_l,x_i)$ and $NDis_P(z_l,x_i)$ reflect the principle of biological taxonomy. It should be noted that the dissimilarity measure $d_P(x,y)$ of two objects is a degenerate form of $NDis_P(z_l,x_i)$ as the lth cluster only includes one object.

Example 2. An artificial data set containing 9 objects with three clusters is given in Table 2, where $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$ and $P=A=\{a_1,a_2\}$. z_1, z_2, z_3 denote the cluster modes. Let us determine object x_1 should be assigned to which cluster using Ng's dissimilarity measure.

Using Ng's dissimilarity measure it follows that:

$$Dis_P(z_1,x_1)=1-2/3+1-1/3=1,$$
$$Dis_P(z_2,x_1)=1-1/3+1-2/3=1.$$

and

$$Dis_P(z_3,x_1)=1+1-2/3=4/3.$$

This means that x_1 has an undetermined assignment.

Replacing Ng's dissimilarity measure with the proposed dissimilarity measure in Example 2, we have:

$$NDis_P(z_1,x_1)=1-1/3\times 2/3+1-1/5\times 1/3=77/45,$$
$$NDis_P(z_2,x_1)=1-1/3\times 1/3+1-1/5\times 2/3=79/45,$$

and

$$NDis_P(z_3,x_1)=1+1-1/5\times 2/3=84/45.$$

Then the object x_1 can be determinately assigned into the cluster *Cluster* 1.

3. Convergence and complexity analysis

In this section, we give some theorems which guarantee the convergency of the *k*-Modes algorithm with the proposed dissimilarity measure. In addition, the time complexity of the algorithm is analyzed.

Table 2
An artificial data set.

Objects	a_1	a_2
x_1	1	2
x_2	1	3
x_3	2	4
Cluster 1 (z_1)	1	2
x_4	1	4
x_5	2	2
x_6	3	2
Cluster 2 (z_2)	1	2
x_7	4	4
x_8	4	2
x_9	5	2
Cluster 3 (z_3)	4	2

Let $IS = (U, A, V, f)$ be a categorical information system, and $P \subseteq A$. The k-Modes algorithm uses the k-Means paradigm to cluster categorical data. The objective of clustering a set of $n = |U|$ objects into k clusters is to find W and Z that minimize:

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

subject to:

$$\omega_{li} \in \{0,1\}, \quad 1 \leqslant l \leqslant k, \quad 1 \leqslant i \leqslant n, \tag{2}$$

$$\sum_{l=1}^{k} \omega_{li} = 1, \quad 1 \leqslant i \leqslant n, \tag{3}$$

and

$$0 < \sum_{i=1}^{n} \omega_{li} < n, \quad 1 \leqslant l \leqslant k, \tag{4}$$

where $k(<n)$ is a known number of clusters, $W = [\omega_{li}]$ is a k-by-n $\{0,1\}$ matrix, $Z = [z_1, z_2, \ldots, z_k]$, and z_l is the lth cluster mode with the attribute set P.

The minimization of F in (1) with the constraints in (2)–(4) forms a class of constrained nonlinear optimization problems whose solutions are unknown. The usual method towards optimization of F in (1) is to use partial optimization for Z and W. In this method, we first fix Z and find necessary conditions on W to minimize F. Then, we fix W and minimize F with respect to Z. This process can be formulated as the following k-Modes algorithm:

Step 1. Choose k distinct objects $z_1, z_2, \ldots, z_k$ from U as an initial mode $Z^{(1)} = [z_1, z_2, \ldots, z_k] \in U^k$. Determine $W^{(1)}$ such that $F(W, Z^{(1)})$ is minimized. Set $t = 1$.

Step 2. Determine $Z^{(t+1)}$ such that $F(W^{(t)}, Z^{(t+1)})$ is minimized. If $F(W^{(t)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t)})$, then stop.

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.

The convergence of the k-Modes algorithm with the simple matching dissimilarity measure and with Ng's dissimilarity measure is proved in [6,9], respectively. By $F_N(W,Z)$ we denote $\sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li} NDis_P(z_l, x_i)$. Now we consider the convergence of the k-Modes algorithm with the proposed dissimilarity measure $NDis_P(z_l, x_i)$.

It is easy to prove the following theorem.

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

$$\min_{W} F_N\left(W, \widehat{Z}\right),$$

subject to (2)–(4). The minimizer $\widehat{W}$ is given by

$$\hat{\omega}_{li} = \begin{cases} 1, & \text{if } NDis_P(\hat{z}_l, x_i) \leqslant NDis_P(\hat{z}_h, x_i), \quad 1 \leqslant h \leqslant k, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 2. *Let $z_l = [z_{l,1}, z_{l,2}, \ldots, z_{l,|P|}]$ be the mode of the lth $(1 \leqslant l \leqslant k)$ cluster and the domain V_{a_j} of attribute a_j be $\left\{a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)}\right\}$ $(1 \leqslant j \leqslant |P|)$. Denote arbitrary object x_i by $[x_{i,1}, x_{i,2}, \ldots, x_{i,\,|P|}]$. Then $F_N(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li} NDis_P(z_l, x_i)$ is minimized if and only if $z_{l,j} = a_j^{(r)}$, where $a_j^{(r)} \in V_{a_j}$ satisfies:*

$$\left|\left\{\omega_{li} \middle| x_{i,j} = a_j^{(r)},\ \omega_{li} = 1\right\}\right| \geqslant \left|\left\{\omega_{li} \middle| x_{i,j} = a_j^{(t)},\ \omega_{li} = 1\right\}\right|,$$

$1 \leqslant t \leqslant n_j$ for $1 \leqslant j \leqslant |P|$.

Proof. For a given W, we have:

$$\begin{aligned} F_N(W,Z) &= \sum_{l=1}^{k}\sum_{i=1}^{n} \omega_{li} NDis_P(z_l, x_i) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{|P|} \omega_{li} NDis_{a_j}(z_l, x_i) \\ &= \sum_{l=1}^{k}\sum_{j=1}^{|P|}\sum_{i=1}^{n} \omega_{li} NDis_{a_j}(z_l, x_i) = \sum_{l=1}^{k}\sum_{j=1}^{|P|} \psi_{l,j}, \end{aligned}$$

where $\psi_{l,j} = \sum_{i=1}^{n} \omega_{li} NDis_{a_j}(z_l, x_i)$.

Note that all the inner sums $\psi_{l,j}$ of $F_N(W,Z)$ are nonnegative and independent. Then minimizing $F_N(W,Z)$ is equivalent to minimizing each inner sum.

By the Definitions 4 and 7, when $z_{l,j} = a_j^{(t)}$, we have:

$$\begin{aligned} \psi_{l,j} &= \sum_{i=1}^{n} \omega_{li} NDis_{a_j}(z_l, x_i) \\ &= \sum_{x_{i,j} = a_j^{(t)}} \omega_{li}\left(1 - Sim_{a_j}(z_l, x_i) \times \frac{|c_{l,j,t}|}{|c_l|}\right) + \sum_{x_{i,j} \neq a_j^{(t)}} \omega_{li} \\ &= \sum_{x_{i,j} = a_j^{(t)}} \omega_{li}\left(1 - \frac{1}{|[x_i]_{\{a_j\}}|} \times \frac{|c_{l,j,t}|}{|c_l|}\right) + \sum_{x_{i,j} \neq a_j^{(t)}} \omega_{li} \\ &= |c_{l,j,t}|\left(1 - \frac{1}{|[x_i]_{\{a_j\}}|} \times \frac{|c_{l,j,t}|}{|c_l|}\right) + \left(|c_l| - |c_{l,j,t}|\right) \\ &= |c_l| - \frac{1}{|[x_i]_{\{a_j\}}| \times |c_l|} \times |c_{l,j,t}|^2 \end{aligned}$$

where $c_{l,j,t} = \left\{x_i \in c_l | x_{i,j} = a_j^{(t)}\right\}$.

It should be noted that the numbers $|c_l|$ and $|[x_i]_{\{a_j\}}|$ are constant for arbitrary l and j. It means that $\psi_{l,j}$ is minimized *iff* $|c_{l,j,t}|$ is maximal for $1 \leqslant t \leqslant n_j$. In other words, $\psi_{l,j}$ is minimized *iff* $t = r$ i.e., $z_{l,j} = a_j^{(r)}$. The result follows. □

We remark that in each iteration of the k-Modes clustering algorithm using the proposed dissimilarity measure the matrix W and Z can be updated according to the Theorems 1 and 2, respectively, when the initial mode Z is given.

Theorem 3. *The k-Modes algorithm with the proposed dissimilarity measure converges in a finite number of iterations.*

Proof. We first note that there are only a finite number $\left(N = \prod_{j=1}^{|P|} n_j\right)$ of possible cluster modes $Z = (z_1, z_2, \ldots, z_k)$. We now show that each possible mode appears at most once in the iteration process of the k-Modes algorithm. If not, there exist $t_1 \neq t_2$ such that $Z^{(t_1)} = Z^{(t_2)}$. According to Theorem 1, the k-Modes algorithm with the proposed measure computes the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ for $Z = Z^{(t_1)}$ and $Z = Z^{(t_2)}$, respectively. By Theorem 1, it is clear that:

$$F\left(W^{(t_1)}, Z^{(t_1)}\right) = F\left(W^{(t_1)}, Z^{(t_2)}\right) = F\left(W^{(t_2)}, Z^{(t_2)}\right).$$

However, the sequence $F(W^{(t)}, Z^{(t)})$ generated by the k-Modes algorithm with the new dissimilarity measure is strictly decreasing. This is a contradiction. Hence, the result follows. □

Theorem 3 implies that the k-Modes algorithm with the proposed dissimilarity measure can be used safely.

The pseudo code of the k-Modes algorithm with the proposed dissimilarity measure is described in Table 3.

Referring to the pseudo code of Table 3, the time complexity of k-Modes algorithm with the new dissimilarity measure is analyzed as follows. We only consider the four major computational steps:

Table 3
The k-Modes algorithm with the proposed dissimilarity measure.

```
1  Initialize the variable oldmodes as a k × |P|-ary empty array;
2  Randomly choose k distinct objects x_1, x_2, ..., x_k from U
3  and assign [x_1, x_2, ..., x_k] to the k × |P|-ary array variable newmodes;
4  for l = 1 to k
5    for j = 1 to |P|
6      calculate the similarity Sim_{a_j}(x_l, x_l) according to Definition 2;
7    end;
8  end;
9  while oldmodes< >newmodes do
10   oldmodes = newmodes;
11   for i = 1 to |U|
12     for l = 1 to k
13       calculate the dissimilarity between the ith object and
14       the lth mode according to Definition 5, and classify the ith
15       object into the cluster whose mode is closest to it;
16     end;
17   end;
18   for l = 1 to k
19     find the mode z_l of each cluster and assign to newmodes;
20     for j = 1 to |P|
21       calculate the similarity Sim_{a_j}(z_l, z_l) according to Definition 2;
22       calculate m_{a_j} of Definition 5;
23     end;
24   end;
25   if oldmodes==newmodes
26     break;
27   end;
28 end.
```

With respect to a given attribute, the time complexity for computing the similarity of one object with itself according to Definition 4 is $O(|U|)$ (see Lines 5 and 19 in Table 3).

The computational complexity for assigning the ith object into the lth cluster is $O(|U||P|k)$ (see Lines 10–15 in Table 3).

The computational complexity for updating all cluster centers is $O(|U||P|k)$ (see Line 17 in Table 3).

The computational complexity of m_{a_j} is $O(|U|)$ (see Line 20 in Table 3).

Suppose that the iteration times is t, the whole computational cost of the k-Modes algorithm with the proposed dissimilarity measure is $O(|U||P|k) + t(O(|U|k + O(|U||P|k))) = O(t|U||P|k)$. This shows that the computational cost is linearly scalable to the number of objects, the number of attributes and the number of clusters.

4. Experimental analysis

In this section, we demonstrate the scalability and efficiency of the k-Modes algorithms based on three different dissimilarity measures. In Section 4.1, the experimental environment and evaluation indexes are described. Section 4.2 presents the scalability on the evaluation results of the k-Modes algorithms using three different dissimilarity measures, and Section 4.3 presents the efficiency on the evaluation results of the k-Modes algorithms using three different dissimilarity measures.

4.1. Experimental environment and evaluation indexes

The experiments are conducted on a PC with an Intel Pentium D processor (2.8 GHz) and 1 G byte memory running the Windows XP SP3 operating system. The k-Modes algorithms with three different dissimilarity measures are coded in Matlab 7.0 programming language.

To evaluate the efficiency of clustering algorithms, the evaluation indexes accuracy defined as $AC = \frac{\sum_{i=1}^{k} a_i}{|U|}$, is employed in the experiments, where k is the number of classes of the data, which is known; a_i is the number of objects that are correctly assigned to the class $C_i (1 \leqslant i \leqslant k)$.

4.2. Evaluation on scalability

To compare the scalability of the k-Modes algorithms with three different dissimilarity measures, the synthetic data sets generated by the generator described in [45] are used. In the synthetic data sets, the number of objects varies from 10,000 to 100,000, and the dimensionality is in the range of 10–50. In all synthetic data sets, each attribute possesses five different values. To avoid the influence of the randomness arising from initializing of cluster centers, each experiment is executed 10 times on the same data set. Therefore, each value in Figs. 1 and 2 is the average of 10 times runs.

Fig. 1 shows the scalability over data size of the k-Modes algorithms with three different dissimilarity measures. This study sets the dimensionality to 10, and the cluster number to 3, and also varies the data size from 10,000 to 100,000. In Fig. 1(a), it can be seen that the complexity of the k-Modes algorithms with three different dissimilarity measures are linear with respect to the data size. The

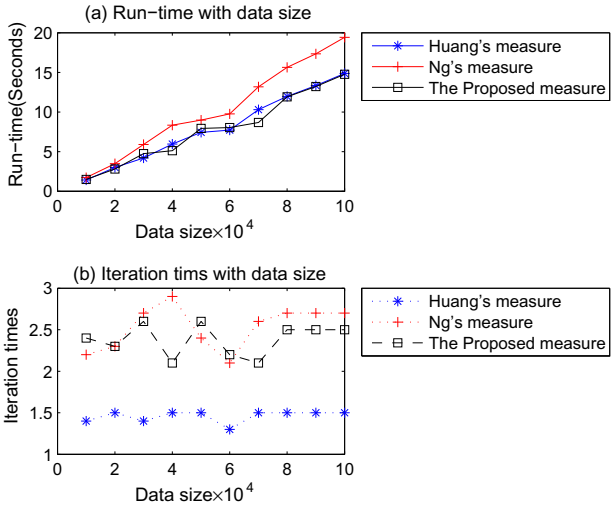

Fig. 1. Execution time comparison using the k-Modes algorithms with three different dissimilarity measures: scalability with data size.

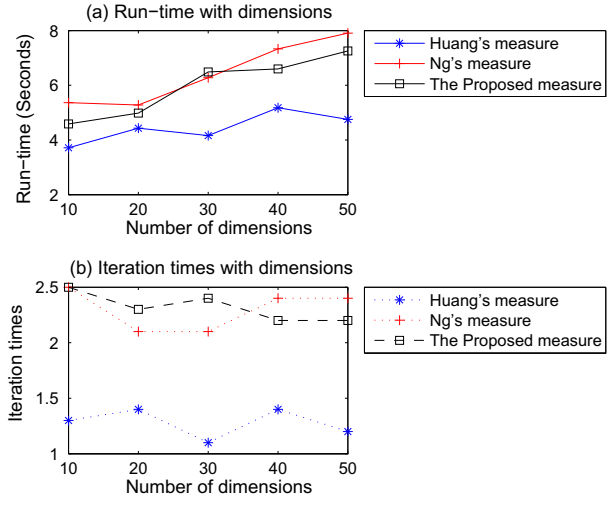

Fig. 2. Execution time comparison using the k-Modes algorithms with three different dissimilarity measures: scalability with data dimensionality.

run-time of the *k*-Modes algorithm using proposed dissimilarity measure is lower than that of the *k*-Modes algorithms using Ng's dissimilarity measure under the same data size. This is because the number of iterations of the *k*-Modes algorithm using the proposed dissimilarity measure are less than that of the *k*-Modes algorithm using Ng's dissimilarity measure (See Fig. 1(b)). The run-time of the *k*-Modes algorithm using the proposed dissimilarity measure is very close to that of the *k*-Modes algorithm using Huang's dissimilarity measure. Therefore, the *k*-Modes algorithm using the proposed dissimilarity measure is also scalable to large data sets.

Fig. 2 shows the scalability with data dimensions of the *k*-Modes algorithms using three different dissimilarity measures. In the experiment, we fix the data size at 30,000, and the cluster number at 3, and vary the number of dimensions from 10 to 50. From the run-time and iteration times, the *k*-Modes algorithm with the proposed dissimilarity measure is nearly close that of the *k*-Modes algorithm with Ng's dissimilarity measure, and is slightly inferior to that of the *k*-Modes algorithm with Huang's dissimilarity measure.

4.3. Evaluation on clustering efficiency

In this subsection, to compare the effectiveness of the *k*-Modes algorithms with three different dissimilarity measures given by simple matching dissimilarity measure, Ng's dissimilarity measure and the proposed dissimilarity measure, five standard data sets are downloaded from the UCI Machine Learning Repository [44]. The data sets' characteristics are summarized in Table 4.

In the experiment, the objects with missing attribute values are removed in Lung-cancer data and Breast-cancer data, respectively. And the 330 outlier records are also removed in Nursery data. We have carried out 100 runs of the *k*-Modes algorithm with each measure on the five standard data sets, respectively. In each run, the same initial cluster modes were used for all the three different measures. And the comparison results of the *k*-Modes algorithm with each of the three measures on the five data sets are shown in Table 5. Each value in the table is the average of 100 times experiments.

According to Table 5, we can find that the clustering performance of the *k*-Modes algorithm using the proposed dissimilarity measure is superior to the others in the case of the five data sets, and especially in case of Lung-cancer data, Breast-cancer data and Mushroom data. In fact, in case of Lung-cancer data, Breast-cancer data and Mushroom data, some biological and genetic taxonomy information is hidden in the attributes, such as Bland Chromatin in case of Breast-cancer data, Cap-surface and Cap-color in case of Mushroom data. This just justifies use the proposed dissimilarity measure. It seems that the proposed dissimilarity measure is more suitable for the data with biological and genetic taxonomy information.

Table 4
Summary of real data sets' characteristics.

Data set	Objects	Attributes	Classes
Lung-cancer	32	56	3
Breast-cancer	683	9	2
Zoo	101	17	7
Mushroom	8140	22	2
Nursery	12690	8	3

Table 5
The *AC* from the three different dissimilarity measure on five data sets.

Data set	Huang's dissimilarity	Ng's dissimilarity	Proposed dissimilarity
Lung-cancer	0.5656	0.5678	0.5841
Breast-cancer	0.8471	0.8581	0.9146
Zoo	0.8382	0.8426	0.8650
Mushroom	0.7863	0.7937	0.8243
Nursery	0.6817	0.6997	0.7010

5. Conclusions

The *k*-Modes algorithm is widely used for clustering categorical data. Dissimilarity or similarity measures play a crucial role for clustering effectiveness. In this paper, the limitations of the simple matching dissimilarity measure and Ng's dissimilarity measure have been analyzed by some illustrative examples. Based on the idea of biological and genetic taxonomy and rough membership function, a new dissimilarity measure for the *k*-Modes algorithm has been defined. A distinct characteristic of the new dissimilarity measure takes into account the distribution of attribute values over the whole universe. As opposed to Ng's dissimilarity measure, it unifies the dissimilarity measures between two objects and between an object and a mode as well. The convergence theorem and time complexity analysis that indicate that the *k*-Modes algorithm using the new dissimilarity measure can be safely and effectively used in case of large data sets. The results of experiments using synthetic data sets and five real data sets from UCI show the effectiveness of the new dissimilarity measure, especially in case of the data sets with biological and genetic taxonomy information. It is our wish that this study provides a new view and thinking on clustering biosystematics data in applications.

Acknowledgements

This work was also supported by the National Natural Science Foundation of China (Nos. 71031006, 70971080, 60970014), the Natural Science Foundation of Shanxi (Nos. 2010021016-2, 2010011021-1), the National Key Basic Research and Developmen Program of China (973) (No. 2011CB311805), the Key Problems in Science and Technology of Shanxi (No. 20110321027-01) and the Foundation of Doctoral Program Research of Ministry of Education of China (No. 200801080006).

References

[1] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufman, San Francisco, 2001.
[2] Z. He, X. Xu, S. Deng, Squeezer: an efficient algorithm for clustering categorical data, Journal of Computational Science and Technology 17 (5) (2002) 611–624.
[3] Z. He, X. Xu, S. Deng, Scalable algorithms for clustering large datasets with mixed type attributes, International Journal of Intelligent Systems 20 (10) (2005) 1077–1089.
[4] G.E. Tsekouras, D. Papageorgiou, S. Kotsiantis, C. Kalloniatis, P. Pintelas, Fuzzy clustering of categorical attributes and its use in analyzing cultural data, International Journal of Computational Intelligence 1 (2) (2004) 147–151.
[5] Z.X. Huang, Clustering large datasets with mixed numeric and categorical values, in: Proceedings of the 1st Pacific Asia Knowledge Discovery and Data Mining Conference, World Scientific, Singapore, 1997, pp. 21–34.
[6] Z.X. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Transactions on Fuzzy Systems 7 (4) (1999) 446–452.
[7] D.W. Kim, K.H. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids, Pattern Recognition Letters 25 (2004) 1263–1271.
[8] Z.X. Huang, Extensions to the *k*-Means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998) 283–304.
[9] M.K. Ng, M.J. Li, Z.X. Huang, Z.Y. He, On the impact of dissimilarity measure in *k*-Modes clustering algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 503–507.
[10] C.C. Hsu, C.L. Chen, Y.W. Su, Hierarchical clustering of mixed data based on distance hierarchy, Information Sciences 177 (20) (2007) 474–4492.
[11] D. Gibson, J. Kleinberg, P. Raghavan, Clustering categorical data: an approach based on dynamical systems, in: Proceedings of the 24th VLDB Conference, New York, 1998, pp. 311–322.

[12] S. Guha, R. Rastogi, K. Shim, Rock: a robust clustering algorithm for categorical attributes, in: Proceedings of the IEEE International Conference on Data Engineering, Sydney, Australia, 1999, pp. 512–521.
[13] V. Ganti, J. Gehrke, R. Ramakrishnan, Cactus-clustering categorical data using summaries, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999, pp. 73–84.
[14] A. Ahamad, L. Dey, A *k*-mean clustering algorithm for mixed numeric and categorical data, Data & Knowledge Engineering 63 (2007) 503–527.
[15] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning 2 (2) (1987) 139–172.
[16] M. Lebowitz, Experiments with incremental concept formation, Machine Learning 2 (2) (1987) 103–138.
[17] M. Gluck, J. Corter, Information, uncertainty, and the utility of categories, in: Proceedings of the Seventh Annual Conference in Cognitive Society, 1985, pp. 283–287.
[18] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning 2 (1987) 139–172.
[19] K. McKusick, K. Thomson, COBWEB/3: a portable implementation, Technical report FIA-90-6-18-2, NASA Ames Research Center, 1990.
[20] Y. Reich, S.J. Fenves, The formation and use of abstract concepts in design, in: D.H. Fisher, M.J. Pazzani, P. Langley (Eds.), Concept Formation: Knowledge and Experience in Unsupervised Learning, Morgan Kaufman, Los Altos, CA, 1991, pp. 323–352.
[21] P. Cheesman, J. Stutz, Bayesian classification (AUTO-CLASS): theory and results, in: Advances in Knowledge Discovery and Data Mining, 1995.
[22] A.K.C. Wong, D.C.C. Wang, DECA: a discrete-valued data clustering algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (4) (1979) 342–349.
[23] A.K.C. Wong, K.Y. Chiu, W. Huang, A discrete-valued clustering algorithm with applications to bimolecular data, Information Sciences 139 (2001) 97–112.
[24] T. Chiu, D. Fang, J. Chen, Y. Wang, C. Jeris, A robust and scalable clustering algorithm for mixed type attributes in large database environment, in: Proceedings of the 2001 International Conference on Knowledge Discovery and Data Mining (KDD01), San Francisco, CA, 2001, pp. 263–268.
[25] D. Barbara, J. Couto, Y. Li, COOLCAT: an entropy-based algorithm for categorical clustering, in: Proceedings of the 7th International Conference on Information and Knowledge Management, McLean, VI, USA, 2002, pp. 582–589.
[26] C.C. Hsu, Y.C. Chen, Mining of mixed data with application to catalog marketing, Expert Systems with Applications 32 (2007) 12–23.
[27] C.C. Hsu, Extending attribute-oriented induction algorithm for major values and numeric values, Expert Systems with Applications 27 (2004) 187–202.
[28] C.C. Hsu, Generalizing self-organizing map for categorical data, IEEE Transactions on Neural Network 17 (2) (2006) 294–304.
[29] R.S. Michalski, P.E. Stepp, Automated construction of classification: conceptual clustering versus numeric taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1983) 396–410.
[30] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics 27 (1971) 857–874.
[31] D.W. Goodall, A new similarity index based on probability, Biometrics 22 (1966) 882–907.
[32] C. li, G. Biswas, Unsupervied learning with mixed numeric and nominal data, IEEE Transactions on Knowledge and Data Engineering 14 (4) (2002) 673–690.
[33] K.C. Gowda, E. Diday, Symbolic clustering using a new dissimilarity measure, Pattern Recognition 24 (6) (1986) 567–578.
[34] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.
[35] J.Y. Liang, C.Y. Dang, K.S. Chin, C.M. Yam Richard, A new method for measuring uncertainty and fuzziness in rough set theory, International Journal of General Systems 31 (4) (2002) 331–342.
[36] Y. Leung, D.Y. Li, Maximal consistent block technique for rule acquisition in incomplete information systems, Information Sciences 153 (2003) 85–106.
[37] Y.H. Qian, J.Y. Liang, D.Y. Li, F. Wang, N.N. Ma, Approximation reduction in inconsistent incomplete decision tables, Knowledge-Based Systems 23 (5) (2010) 427–433.
[38] D. Parmar, T. Wu, J. Blackhurst, MMR: an algorithm for clustering data using rough set theory, Data & Knowledge Engineering 63 (3) (2007) 879–893.
[39] P. Yang, Q.S. Zhu, Finding key attribute subset in dataset for outlier detection, Knowledge-Based Systems 24 (2) (2010) 269–274.
[40] F. Jiang, Y.F. Sui, C.G. Cao, A rough set approach to outlier detection, International Journal of General Systems 37 (5) (2008) 519–536.
[41] C.B. Chen, L.Y. Wang, Rough set-based clustering with refinement using Shannon's Entropy theory, Computer and Mathematics with Application 52 (2006) 1563–1576.
[42] T. Herawan, M.M. Deris, Jemal H. Abawajy, A rough set approach for selecting clustering attribute, Knowledge-Based Systems 23 (3) (2010) 220–231.
[43] F.Y. Cao, J.Y. Liang, L. Bai, C.Y. Dang, A framework for clustering categorical time-evolving data, IEEE Transactions on Fuzzy Systems 18 (5) (2010) 872–885.
[44] UCI Machine Learning Repository. <http://www.ics.uci.edu/mlearn/MLRepository.html>, 2009.
[45] Data Generator: Perfect Data for an Imperfect World. <http://www.generatedata.com>, 2010.