A New Distance Metric for Unsupervised Learning of Categorical Data

Hong Jia, Yiu-ming Cheung, *Senior Member, IEEE*, and Jiming Liu, *Fellow, IEEE*

***Abstract*— Distance metric is the basis of many learning algorithms, and its effectiveness usually has a significant influence on the learning results. In general, measuring distance for numerical data is a tractable task, but it could be a nontrivial problem for categorical data sets. This paper, therefore, presents a new distance metric for categorical data based on the characteristics of categorical values. In particular, the distance between two values from one attribute measured by this metric is determined by both the frequency probabilities of these two values and the values of other attributes that have high interdependence with the calculated one. Dynamic attribute weight is further designed to adjust the contribution of each attribute-distance to the distance between the whole data objects. Promising experimental results on different real data sets have shown the effectiveness of the proposed distance metric.**

***Index Terms*— Attribute interdependence, categorical attribute, clustering analysis, distance metric, unsupervised learning.**

I. Introduction

MEASURING the distance or dissimilarity between two data objects plays an important role in many data mining and machine learning tasks, such as clustering, classification, recommendation system, outlier detection, and so on. In general, distance computation is an embedded step for these learning algorithms, and different metrics can be conveniently utilized. However, the effectiveness of adopted distance metric usually has a significant influence on the performance of the whole learning method [1]–[3]. Therefore, it becomes a key research issue to present more appropriate distance metrics for the various learning tasks.

For purely numerical data sets, the distance computation is a tractable problem as any numerical operation can be directly applied. In the literature, a number of distance metrics and metric learning methods have been proposed for numerical data. The most widely used metrics in practice should be the Manhattan distance, Euclidean distance, and Mahalanobis distance [4]. By contrast, measuring distance for categorical data can be much more challenging. For example, Table I shows a fragment of the Mushroom data set from UCI Machine Learning Repository. It is known that some of these mushroom samples are edible but some are poisonous. If we have no information about their categories and would like to conduct a well partition on them, or given a wild mushroom, we want to find out which sample from the database is most similar to it and we have to measure the difference between these mushroom samples. However, as the attribute values of this data set are unordered nominal values rather than numerical ones, the only numerical operation that can be straightforwardly applied is the identical comparison operation [5]. Under these circumstances, the popular distance metrics defined for numerical data cannot work anymore. In general, the simplest way to overcome this problem is to transform the categorical values into numerical ones, e.g., the binary strings [6]–[8], and then, the existing numerical-value-based distance metrics can be utilized. Nevertheless, such a kind of method has ignored the information embedded in the categorical values and cannot faithfully reveal the relationship structure of the data sets [9], [10]. Therefore, it is desirable to solve this problem by proposing a distance metric for categorical data based on the characteristics of categorical values.

TABLE I
Fragment of Mushroom Data Set

No.	Cap-shape	Cap-surface	Gill-attachment	Gill-spacing
1	convex	smooth	free	close
2	bell	smooth	free	close
3	convex	scaly	free	close
4	flat	fibrous	free	crowded
5	flat	smooth	attached	close
6	knobbed	scaly	free	close
7	knobbed	fibrous	free	crowded
8	convex	smooth	attached	close
9	knobbed	smooth	attached	close

Among the existing works, the most straightforward and widely used distance metric for categorical data is the Hamming distance [4], in which the distance between different categorical values is set at 1, while a distance of 0 is assigned to identical values. Although the Hamming distance is easy to understand and convenient for computation, the main drawback of this metric is that all attribute values have been

Manuscript received July 31, 2014; revised May 4, 2015; accepted May 16, 2015. Date of publication June 9, 2015; date of current version April 15, 2016. This work was supported in part by the Faculty Research Grant of Hong Kong Baptist University (HKBU) under Project FRG2/14-15/075 and Project FRG1/14-15/041, in part by the National Science Foundation of China under Grant 61272366, and in part by the Strategic Development Fund under Grant HKBU: 03-17-033. *(Corresponding author: Yiu-ming Cheung.)*

H. Jia and J. Liu are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: hjia@comp.hkbu.edu.hk; jiming@comp.hkbu.edu.hk).

Y.-m. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, and also with the United International College, Beijing Normal University–Hong Kong Baptist University, Zhuhai 519085, China (e-mail: ymc@comp.hkbu.edu.hk).

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

Digital Object Identifier 10.1109/TNNLS.2015.2436432

considered equally and the statistical properties of different values have not been distinguished [1]. For this reason, more researchers have attempted to measure the distance for categorical data by considering the distribution characteristics of categorical values. For example, Cost and Salzberg [11] proposed a distance metric, namely, modified value difference metric (MVDM) for supervised learning task. Since true class label is required by the MVDM, it cannot work in an unsupervised learning environment. Under the circumstances, for unsupervised distance measure of categorical data, Le and Ho [5] presented an indirect method that defines the distance between two values from one attribute as the sum of the Kullback–Leibler divergence between conditional probability distributions of other attributes, given these two values. Similar idea has also been adopted in [12]. These two methods have assumed that each attribute can be jointly expressed with all the other attributes without considering the relevancy between different attribute pairs. Therefore, Ienco *et al.* [13], [14] proposed the concept of context, which is a subset that contains the attributes that are relevant to the given one. Subsequently, the distance between two values of an attribute is measured based on the values of the attributes from current attribute's context. Although the relationships between different attribute pairs have been well taken into account, numerical experiments and analysis have found that these three kinds of indirectly defined distance metrics [5], [12], [14] cannot work if the attributes of the given data set are totally independent of each other.

Besides the aforementioned methods, which directly propose special distance metric for categorical data sets, some similarity measures [15]–[23] presented for categorical or mixed data can also be utilized to quantify the relationship between different categorical data objects. For example, the Goodall similarity metric proposed in [15] assigns a greater weight to the matching of uncommon attribute values than common values in similarity computation without assuming the underlying distributions of categorical values. This method has paid attention to the occurrence frequency of different values. However, the similarity between two different values has been defined as a constant 0, which has ignored the characters of these attribute values. Moreover, Gowda and Diday [16]–[18] proposed an algebraic method to measure the similarity between complex data objects. In this method, the similarity between two attribute values is defined based on three components: 1) position; 2) span; and 3) content. This similarity measure is applicable for both of numerical attributes and categorical attributes. Nevertheless, this method is more suitable for interval-type attribute values. For a pair of different values with absolute type, which is common in practice, the similarity between them will always be quantified by 0. In addition, all these similarity measures treated the categorical attributes individually and have ignored the variant attribute relationships.

Furthermore, other than the previously introduced deterministic methods, metric learning is also a useful approach to acquire a good distance metric for the given data objects. This technique was first proposed in [24]. Ever since then, extensive research has been conducted in this area, and variant approaches have been presented in the literature. To overview the details of popular metric learning methods, one can refer to [25] and [26]. Roughly, the existing metric learning methods can be grouped into two categories: 1) linear methods and 2) nonlinear ones. Representative linear methods include global distance metric learning [24], neighborhood components analysis [27], large margin nearest neighbors (LMNN) [28], information theoretic metric learning [29], and semisupervised metric learning paradigm with hypersparsity [30]. Besides, typical examples of nonlinear approaches include the kernelization methods [31]–[33], neural network-based methods [34], and the χ^2-LMNN and GB-LMNN proposed in [35]. In general, the aforementioned metric learning methods are proposed for purely numerical data and cannot be directly applied to the data with categorical attributes. To solve this problem, He *et al.* [36] recently proposed the kernel density metric learning (KDML), which provides nonlinear, probability-based distance measure, and can handle not only numerical attributes but also categorical ones. However, KDML is a supervised method, and the true class labels should be provided in advance, which restricts its application in the unsupervised learning environment.

In this paper, we further study the distance measure for categorical data objects and propose a new distance metric, which can well quantify the distance between categorical values in the unsupervised learning environment. This distance metric takes into account the characteristics of the categorical values. The core idea is to measure the distance with the frequency probability of each attribute value in the whole data set. Moreover, in order to utilize the useful relationship information accompanying with each pair of attributes well, the interdependence redundancy measure [37] has been introduced to evaluate the dependence degree between different attributes. Subsequently, the distance between two values from one attribute is not only measured by their own frequency probabilities but also determined by the values of other attributes that are highly correlated with this one. In addition, a new kind of weight named dynamic attribute weight has been presented to adjust the contribution of distance along each attribute to the whole object distance. The effectiveness of the proposed metric has been experimentally investigated on different real data sets in terms of cluster discrimination and clustering analysis. The promising results indicate that the proposed distance metric is appropriate for the unsupervised learning on categorical data as it can well reveal the true relationship between categorical objects. The main contributions of our work can be summarized as follows.

1) A frequency-probability-based distance measure is proposed for the categorical values. This method regards the situation with 0 distance as distance origin, and the distances between other pairs of values are quantified by comparing with this origin situation.
2) A dynamic weighting scheme for categorical attributes is presented, which assigns larger weights to the attributes with infrequent matching or mismatching value pairs as they can provide more important information.
3) The dependence degree between each pair of attributes is introduced. The complete distance between

two categorical values from one attribute is estimated with not only their own frequency probability but also the co-occurrent probability of them with other values from highly correlated attributes.

The rest of this paper is organized as follows. In Section II, we will overview some existing distance metrics for categorical data. Section III proposes a new distance metric for the unsupervised learning of categorical data. Then, Section IV shows the experimental results on real data sets. Finally, the conclusion is drawn in Section V.

II. Overview of Existing Distance Metric

In the literature, researchers have proposed some distance metrics to quantify the distance between categorical data. This section will present an overview of them as follows.

A. Hamming Distance

Suppose we have a data set with n objects, expressed as $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, represented by a set of categorical attributes $\{A_1, A_2, \ldots, A_d\}$, where d is the dimensionality of the data. Each attribute A_r can be accompanied by a value domain $\text{dom}(A_r)$ ($r = 1, 2, \ldots, d$), which contains all the possible values that can be chosen by this attribute. Since the value domains of the categorical attributes are finite and unordered, the domain of A_r with m_r elements can be expressed as $\text{dom}(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$ and for any $a, b \in \text{dom}(A_r)$, either $a = b$ or $a \neq b$ [38]. Subsequently, each object $\mathbf{x}_i$ can be denoted by a vector $(x_{i1}, x_{i2}, \ldots, x_{id})^T$, where $x_{ir} \in \text{dom}(A_r)$ and T is the transpose operator of a matrix.

For each pair of categorical data object $\mathbf{x}_i$ and $\mathbf{x}_j$, $i, j \in \{1, 2, \ldots, n\}$, the Hamming distance [4] between them is defined as

$$D(x_i, x_j) = \sum_{r=1}^{d} \delta(x_{ir}, x_{jr}) \tag{1}$$

with

$$\delta(x_{ir}, x_{jr}) = \begin{cases} 1, & \text{if } x_{ir} \neq x_{jr} \\ 0, & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{2}$$

That is, the distance between two different categorical values is fixed at 1 and the distance between identical values is regarded as 0. Thus, the Hamming distance between a pair of categorical data objects will be equal to the number of attributes in which they mismatch.

B. Modified Value Difference Metric

Cost and Salzberg [11] modified Stanfill and Waltz's VDM [39] with a new weighting scheme to make the symmetric condition hold. This distance metric is applicable for supervised learning on categorical data. Based on a training set X with samples from k different clusters, the distance between two categorical values of a specific attribute A_r is defined as

$$D(a_{ri}, a_{rj}) = \sum_{t=1}^{k} |p(C = t|A_r = a_{ri}) - p(C = t|A_r = a_{rj})|^{\alpha} \tag{3}$$

where $r \in \{1, 2, \ldots, d\}$, $i, j \in \{1, 2, \ldots, m_r\}$, C stands for the class label and α is a constant, which is usually set to 1. $p(C = t|A_r = a_{ri})$ calculates the conditional probability of $C = t$ given that $A_r = a_{ri}$. Under this metric, two values are regarded as similar if they occur with the same relative frequency for all categories. Subsequently, the distance between two data samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated by

$$D(\mathbf{x}_i, \mathbf{x}_j) = w_{\mathbf{x}_i} w_{\mathbf{x}_j} \sum_{r=1}^{d} D(x_{ir}, x_{jr})^{\gamma} \tag{4}$$

where $\mathbf{x}_i$ is a data sample from the training set and $\mathbf{x}_j$ is a new sample. The distance between categorical values x_{ir} and x_{jr} is given by (3) and the constant γ was set at 2 in most cases. $w_{\mathbf{x}_i}$ and $w_{\mathbf{x}_j}$ are the weights assigned to the samples $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. In practice, for the new sample $\mathbf{x}_j$, $w_{\mathbf{x}_j}$ is set at 1. For a training sample $\mathbf{x}_i$, the weight is assigned according to its performance history. Usually, accurate sample will have $w_{\mathbf{x}_i} \approx 1$, while unreliable sample will have $w_{\mathbf{x}_i} > 1$ to make it appear further away from a new sample.

C. Ahmad's Distance Metric

Ahmad and Dey [12] proposed to calculate the distance between any two categorical values from one attribute with respect to all the other attributes. In particular, given the categorical data set X, the distance between a pair of categorical values a_{ri} and a_{rj} from attribute A_r is defined as

$$D(a_{ri}, a_{rj}) = \frac{1}{d-1} \left(\sum_{t=1}^{d} D(a_{ri}, a_{rj}, A_t) \right) \quad (t \neq r) \tag{5}$$

where $r \in \{1, 2, \ldots, d\}$, $i, j \in \{1, 2, \ldots, m_r\}$, and $D(a_{ri}, a_{rj}, A_t)$ stands for the distance between a_{ri} and a_{rj} with respect to attribute A_t, which is defined as

$$D(a_{ri}, a_{rj}, A_t) = \max \left(p_r^{a_{ri}}(\Omega) + p_r^{a_{rj}}(\sim\Omega) - 1 \right). \tag{6}$$

Here, Ω is a subset of $\text{dom}(A_t)$. $p_r^{a_{ri}}(\Omega)$ denotes the probability that data objects in X with rth attribute value equal to a_{ri} has the value contained in Ω for attribute A_t. $p_r^{a_{rj}}(\sim\Omega)$ denotes the probability that data objects in X with the rth attribute value equal to a_{rj} has the value not contained in Ω for attribute A_t. In practice, $p_r^{a_{ri}}(\Omega)$ is calculated by

$$p_r^{a_{ri}}(\Omega) = \sum_{a_{\text{th}} \in \Omega} p(A_t = a_{\text{th}}|A_r = a_{ri}) \tag{7}$$

and $p_r^{a_{rj}}(\sim\Omega)$ is given by

$$p_r^{a_{rj}}(\sim\Omega) = \sum_{a_{\text{th}} \in (\sim\Omega)} p(A_t = a_{\text{th}}|A_r = a_{rj}). \tag{8}$$

Let $\vartheta_t = \max(p_r^{a_{ri}}(\Omega) + p_r^{a_{rj}}(\sim\Omega))$, according to [40], the value of ϑ_t and the corresponding value set Ω can be calculated using Algorithm 1. Subsequently, we have $D(a_{ri}, a_{rj}, A_t) = \vartheta_t - 1$.

However, numerical studies have found that if the attributes of given data set are totally independent of each other according to the mutual information criterion [41], this distance metric cannot work well as the distance between each pair of attribute values will be quantified as 0. For example, suppose

Algorithm 1 Calculate ϑ_t and Ω

```
Let $\vartheta_t = 0$ and $\Omega = \phi$.
for $h = 1$ to $m_t$ do
    if $p(A_t = a_{th}|A_r = a_{ri}) > p(A_t = a_{th}|A_r = a_{rj})$ then
        Add $a_{th}$ to $\Omega$
        $\vartheta_t = \vartheta_t + p(A_t = a_{th}|A_r = a_{ri})$
    else
        Add $a_{th}$ to $\sim\Omega$
        $\vartheta_t = \vartheta_t + p(A_t = a_{th}|A_r = a_{rj})$
    end if
end for
```

TABLE II

EXAMPLE OF DATA SET X

Data	A_1	A_2	A_3
$\mathbf{x}_1$	E	F	L
$\mathbf{x}_2$	E	F	M
$\mathbf{x}_3$	E	H	L
$\mathbf{x}_4$	E	H	M
$\mathbf{x}_5$	G	F	L
$\mathbf{x}_6$	G	F	M
$\mathbf{x}_7$	G	H	L
$\mathbf{x}_8$	G	H	M

the data set X is given as Table II. The distance between categorical values E and G will be calculated as follows.

1) Get the distance between E and G with respect to attribute A_2

$$\begin{aligned} D(E, G, A_2) &= \max\left(p_1^E(\Omega) + p_1^G(\sim\Omega) - 1\right) \\ &= p_1^E(\{F\}) + p_1^G(\{H\}) - 1 \\ &= \frac{1}{2} + \frac{1}{2} - 1 = 0. \end{aligned}$$

2) Get the distance between E and G with respect to attribute A_3

$$\begin{aligned} D(E, G, A_3) &= \max\left(p_1^E(\Omega) + p_1^G(\sim\Omega) - 1\right) \\ &= p_1^E(\{L\}) + p_1^G(\{M\}) - 1 \\ &= \frac{1}{2} + \frac{1}{2} - 1 = 0. \end{aligned}$$

3) Based on the previous results, we have

$$D(E, G) = \frac{1}{2} \cdot (0 + 0) = 0.$$

Following the similar procedure, we can further get $D(F, H) = 0$ and $D(L, M) = 0$. These results indicate that Ahmad's distance metric is not applicable to this kind of data set as it cannot distinguish different categorical values and data objects.

D. Association-Based Distance Metric

The association-based distance metric presented in [5] also utilizes an indirect method to estimate the distance between categorical values. According to this metric, the distance between two categorical values of one attribute is indirectly estimated by the sum of dissimilarities between conditional probability distributions of other attributes given these two values. In particular, for the categorical data set X, the distance between two values a_{ri} and a_{rj} of attribute A_r is defined as

$$D(a_{ri}, a_{rj}) = \sum_{t=1}^{d} \psi(\text{cpd}(A_t|A_r = a_{ri}), \quad \text{cpd}(A_t|A_r = a_{rj})) \tag{9}$$

where $r \in \{1, 2, \ldots, d\}$, $i, j \in \{1, 2, \ldots, m_r\}$, and $t \neq r$. $\text{cpd}(A_t|A_r = a_{ri})$ and $\text{cpd}(A_t|A_r = a_{rj})$ are conditional probability distributions. $\psi(.,.)$ is a dissimilarity function for two probability distributions. In practice, if the Kullback–Leibler divergence method [42], [43] is utilized as the dissimilarity measure, $D(a_{ri}, a_{rj})$ can be calculated by

$$\begin{aligned} D(a_{ri}, a_{rj}) = \sum_{t, t \neq r} \sum_{h=1}^{m_t} \Bigg(& p(a_{\text{th}}|a_{ri}) \frac{p(a_{\text{th}}|a_{ri})}{p(a_{\text{th}}|a_{rj})} \\ & + p(a_{\text{th}}|a_{rj}) \frac{p(a_{\text{th}}|a_{rj})}{p(a_{\text{th}}|a_{ri})} \Bigg) \end{aligned} \tag{10}$$

where $p(a_{\text{th}}|a_{ri})$ stands for the conditional probability $p(A_t = a_{\text{th}}|A_r = a_{ri})$. Similar to Ahmad's distance metric, if all attributes are independent of each other, the distance between any pair of values will be estimated as 0.

E. Context-Based Distance Metric

To distinguish the different relationships between attributes, Ienco *et al.* [13], [14] have proposed the concept of context, which is a subset of relevant attributes. In practice, the context of an attribute A_r, denoted as $\text{context}(A_r)$, is determined by a measure named symmetrical uncertainty (SU). In particular, for two attributes A_r and A_t, the SU is calculated by

$$\text{SU}(A_r, A_t) = 2 \cdot \frac{\text{IG}(A_r|A_t)}{H(A_r) + H(A_t)} \tag{11}$$

where $H(A_r)$ and $H(A_t)$ are the entropy of attributes A_r and A_t, respectively. $\text{IG}(A_r|A_t)$ is the information gain, which is given by

$$\text{IG}(A_r|A_t) = H(A_r) - H(A_r|A_t) \tag{12}$$

with

$$\begin{aligned} H(A_r) &= -\sum_{i=1}^{m_r} p(a_{ri}) \log(p(a_{ri})) \\ H(A_r|A_t) &= -\sum_{j=1}^{m_t} p(a_{tj}) \sum_{i=1}^{m_r} p(a_{ri}|a_{tj}) \log(p(a_{ri}|a_{tj})). \end{aligned}$$

Subsequently, the context of attribute A_r is determined by

$$\text{context}(A_r) = \{A_t | t \neq r, \text{SU}(A_r, A_t) \geq \sigma E[\text{SU}_{A_r}]\} \tag{13}$$

with

$$E[\text{SU}_{A_r}] = \frac{\sum_{t, t \neq r} \text{SU}(A_r, A_t)}{d - 1}$$

where $\sigma \in [0,1]$ is a tradeoff parameter. Then, the distance between two values a_{ri} and a_{rj} of attribute A_r is defined as

$$D(a_{ri}, a_{rj}) = \sqrt{\sum_{A_t \in \text{context}(A_r)} \sum_{h=1}^{m_t} (p(a_{ri}|a_{\text{th}}) - p(a_{rj}|a_{\text{th}}))^2}. \tag{14}$$

Since this method is also an indirect one, it still cannot work well on data with totally independent attributes.

III. Proposed Distance Metric for Categorical Data

This section will propose a metric to quantify the distance between categorical data for unsupervised learning well. In this new distance metric, the relationship between different attributes, as well as the characteristics of categorical value, will be considered.

A. Frequency Probability-Based Distance Metric

Given the data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ with n objects represented by d categorical attributes $\{A_1, A_2, \ldots, A_d\}$, the distance between two data objects $\mathbf{x}_i$ and $\mathbf{x}_j$ can generally be calculated by

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{d} D(x_{ir}, x_{jr}). \tag{15}$$

Therefore, the key point is to define the distance between two categorical values. Given two categorical values x_{ir} and x_{jr} of attribute A_r from data objects $\mathbf{x}_i$ and $\mathbf{x}_j$, we define $D(x_{ir}, x_{jr}) = 0$ if $x_{ir} = x_{jr}$. This is regarded as the distance origin and the distance of other situation is estimated by comparing with this one. This perspective of distance consists with the common distance definition between numerical objects. In particular, in the numerical space, calculating the distance between two objects in different positions can be regarded as to measure how much distance will be taken by the two objects to get to the same position. Here, the situation that the two objects reach the same position can also be regarded as the distance origin, and the difference between the different-position situation and the distance-origin situation is another explanation of the distance between the two objects.

Subsequently, we define the distance or the difference between the situations $x_{ir} \neq x_{jr}$ and $x_{ir} = x_{jr}$ as the frequency of situation $x_{ir} = x_{jr}$ in the whole data set. For two different categorical values x_{ir} and x_{jr}, the corresponding equal situation has two cases: 1) both with value x_{ir} and 2) both with value x_{jr}. For given data set X, the frequency of the first case is calculated by

$$\begin{aligned} u(\{x_{ir}, x_{ir}\}) &= n \cdot p(\{x_{ir}, x_{ir}\}) \\ &= n \cdot p(A_r = x_{ir}|X) \cdot p^{-}(A_r = x_{ir}|X) \end{aligned} \tag{16}$$

where the frequency probability $p(A_r = x_{ir}|X)$ is given by

$$p(A_r = x_{ir}|X) = \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq \text{NULL}}(X)} \tag{17}$$

and $p^{-}(A_r = x_{ir}|X)$ is calculated by

$$p^{-}(A_r = x_{ir}|X) = \frac{\sigma_{A_r = x_{ir}}(X) - 1}{\sigma_{A_r \neq \text{NULL}}(X) - 1}. \tag{18}$$

Here, the operation $\sigma_{A_r = x_{ir}}(X)$ counts the number of objects in the data set X that have the value x_{ir} for attribute A_r and the symbol NULL refers to the empty. Analogously, the frequency of the other case is calculated by

$$\begin{aligned} u(\{x_{jr}, x_{jr}\}) &= n \cdot p(\{x_{jr}, x_{jr}\}) \\ &= n \cdot p(A_r = x_{jr}|X) \cdot p^{-}(A_r = x_{jr}|X). \end{aligned} \tag{19}$$

Since these two cases are exclusive events, we can get that the frequency of the corresponding equal situation for the pair of categorical values $\{x_{ir}, x_{jr}\}$ is as follows:

$$\begin{aligned} u(x_{ir} = x_{jr}) &= u(\{x_{ir}, x_{ir}\}) + u(\{x_{jr}, x_{jr}\}) \\ &= n \cdot [p(A_r = x_{ir}|X) \cdot p^{-}(A_r = x_{ir}|X) \\ &\quad + p(A_r = x_{jr}|X) \cdot p^{-}(A_r = x_{jr}|X)]. \end{aligned} \tag{20}$$

Let $p(x_{ir} = x_{jr})$ denote the probability of the equal situation. Subsequently, we have

$$\begin{aligned} p(x_{ir} = x_{jr}) = \; & p(A_r = x_{ir}|X) \cdot p^{-}(A_r = x_{ir}|X) \\ & + p(A_r = x_{jr}|X) \cdot p^{-}(A_r = x_{jr}|X) \end{aligned} \tag{21}$$

and

$$u(x_{ir} = x_{jr}) = n \cdot p(x_{ir} = x_{jr}). \tag{22}$$

Since $u(x_{ir} = x_{jr})$ is regarded as the difference between the situation $x_{ir} \neq x_{jr}$ and the distance origin, i.e., the situation $x_{ir} = x_{jr}$, the distance between the two different categorical values x_{ir} and x_{jr} can be estimated by it. That is

$$D(x_{ir}, x_{jr}) = n \cdot p(x_{ir} = x_{jr}), \quad \text{if } x_{ir} \neq x_{jr}. \tag{23}$$

Consequently, the distance between categorical values from one attribute can be defined based on frequency probability as follows:

$$D(x_{ir}, x_{jr}) = \begin{cases} n \cdot p(x_{ir} = x_{jr}), & \text{if } x_{ir} \neq x_{jr} \\ 0, & \text{if } x_{ir} = x_{jr} \end{cases} \tag{24}$$

where $i, j \in \{1, 2, \ldots, n\}$, $r \in \{1, 2, \ldots, d\}$, and $p(x_{ir} = x_{jr})$ are calculated by (21). Moreover, to avoid the situation that data sets with different sample sizes will have different distance scales, we can delete the constant n in (24) without influencing the distance ranking. Therefore, the simplified distance metric for categorical values can be given by

$$D(x_{ir}, x_{jr}) = \begin{cases} p(x_{ir} = x_{jr}), & \text{if } x_{ir} \neq x_{jr} \\ 0, & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{25}$$

Subsequently, the expression of distance between categorical data $\mathbf{x}_i$ and $\mathbf{x}_j$ can be written as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{d} [\delta(x_{ir}, x_{jr}) p(x_{ir} = x_{jr})] \tag{26}$$

where the definition of $\delta(x_{ir}, x_{jr})$ is given by (2).

In addition, it can be easily derived that the distance metric defined by (26) satisfies the following conditions.

1) $D(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
2) $D(\mathbf{x}_i, \mathbf{x}_j) = 0$ if and only if $\mathbf{x}_i = \mathbf{x}_j$.
3) $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i)$.
4) $D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_l) + D(\mathbf{x}_l, \mathbf{x}_j)$, where $i, j, l \in \{1, 2, \ldots, n\}$.

B. Dynamic Attribute Weight

Most existing distance or similarity metrics for categorical data treat each attribute equally in a data set. However, this is not always reasonable in practice. For example, when we compare two objects, we usually pay more attention to the special features they have. That is, unusual features generally can provide more important information for the comparison between objects. Considering this phenomenon, we can further adjust the previous distance metric according to the following criterion. The contribution of the distance between two attribute values to the whole object distance is inverse to the probability of these two values' situation in the whole data set. That is, if two data objects have different values along one attribute, then the contribution of the distance between these two values to the whole data distance is inverse to the probability that two data objects have different values along this attribute in the data set, and vice versa. Therefore, this kind of probability can be utilized as a dynamic weight of attribute distance. Comparing with the existing attribute weighting methods, the proposed one has at least three advantages.

1) It is defined based on an individual situation of attribute value pair, but not the general information of the whole attribute. Therefore, it has better adjusting ability in practice.
2) It highlights the infrequent but important matching or mismatching information, which is consist with the general criterion in practice.
3) As the distance metric is defined with the frequency probability of attribute values, this kind of weight can avoid the domination of values with high frequency.

For an attribute A_r with m_r possible values, the probability that two data objects from X have the same value along A_r is calculated by

$$p_s(A_r) = \sum_{j=1}^{m_r} p(A_r = a_{rj}|X)p^{-}(A_r = a_{rj}|X). \tag{27}$$

Correspondingly, the probability that two data objects from X have different values along A_r is given by

$$p_f(A_r) = 1 - p_s(A_r). \tag{28}$$

Subsequently, following the proposed criterion, the dynamically weighted distance metric should be:

$$\begin{aligned} D(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{r=1}^{d} [\widetilde{w}(A_r) D(x_{ir}, x_{jr})] \\ &= \sum_{r=1}^{d} [\widetilde{w}(A_r)\delta(x_{ir}, x_{jr}) p(x_{ir} = x_{jr})] \end{aligned} \tag{29}$$

where $\widetilde{w}(A_r)$ is the dynamic weight of attribute A_r, which is calculated by

$$\widetilde{w}(A_r) = \frac{w(A_r)}{\sum_{r=1}^{d} w(A_r)} \tag{30}$$

with

$$w(A_r) = \begin{cases} 1 - p_f(A_r), & \text{if } x_{ir} \neq x_{jr} \\ 1 - p_s(A_r), & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{31}$$

That is, if $x_{ir} \neq x_{jr}$, the larger the probability $p_f(A_r)$ is, the smaller weight will be assigned to the distance along the attribute A_r. Similarly, if $x_{ir} = x_{jr}$, the contribution of the distance between them to the data object distance will decrease as the value of $p_s(A_r)$ increases. Moreover, since $p_f(A_r) + p_s(A_r) = 1$, (31) can be rewritten as

$$w(A_r) = \begin{cases} p_s(A_r), & \text{if } x_{ir} \neq x_{jr} \\ p_f(A_r), & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{32}$$

C. Relationship Between Categorical Attributes

In the previous distance metric, the distance along each attribute has been computed individually. However, in practice, we often have some attributes that are highly dependent on each other. Under the circumstances, the computation of similarity or dissimilarity for categorical attributes in the unsupervised learning task should be considered based on frequently co-occurring items [44]. That is, the distance between two values from one attribute should be calculated by considering the other attributes that are highly correlated with this one. In particular, given the data set X, the dependence degree between each pair of attributes A_i and A_j ($i, j \in \{1, 2, \ldots, d\}$) can be quantified based on the mutual information [41] between them, which is defined as

$$I(A_i; A_j) = \sum_{r=1}^{m_i} \sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log\left(\frac{p(a_{ir}, a_{jl})}{p(a_{ir})p(a_{jl})}\right). \tag{33}$$

Here, the items $p(a_{ir})$ and $p(a_{jl})$ stand for the frequency probability of the two attribute values in the whole data set, which are calculated by

$$p(a_{ir}) = p(A_i = a_{ir}|X) = \frac{\sigma_{A_i = a_{ir}}(X)}{\sigma_{A_i \neq \text{NULL}}(X)} \tag{34}$$

$$p(a_{jl}) = p(A_j = a_{jl}|X) = \frac{\sigma_{A_j = a_{jl}}(X)}{\sigma_{A_j \neq \text{NULL}}(X)}. \tag{35}$$

The expression $p(a_{ir}, a_{jl})$ is to calculate the joint probability of these two attribute values, i.e., the frequency probability of objects in X having $A_i = a_{ir}$ and $A_j = a_{jl}$, which is given by

$$\begin{aligned} p(a_{ir}, a_{jl}) &= p(A_i = a_{ir} \wedge A_j = a_{jl}|X) \\ &= \frac{\sigma_{A_i = a_{ir} \wedge A_j = a_{jl}}(X)}{\sigma_{A_i \neq \text{NULL} \wedge A_j \neq \text{NULL}}(X)}. \end{aligned} \tag{36}$$

The mutual information between two attributes actually measures the average reduction in uncertainty about one attribute that results from learning the value of the other [41]. A larger value of mutual information usually indicates greater dependence. However, a disadvantage of using this index is that its value increases with the number of possible values that can be chosen by each attribute. Therefore, Au *et al.* [37] proposed to normalize the mutual information with a joint entropy, which yields the interdependence redundancy measure denoted as

$$R(A_i; A_j) = \frac{I(A_i; A_j)}{H(A_i, A_j)} \tag{37}$$

where the joint entropy $H(A_i, A_j)$ is calculated by

$$H(A_i, A_j) = -\sum_{r=1}^{m_i}\sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log[p(a_{ir}, a_{jl})]. \tag{38}$$

This interdependence redundancy measure evaluates the degree of deviation from independence between two attributes [37]. In particular, $R(A_i; A_j) = 1$ means that the attributes A_i and A_j are strictly dependent on each other while $R(A_i; A_j) = 0$ indicates that they are statistically independent. If the value of $R(A_i; A_j)$ is between 0 and 1, we can say that these two attributes are partially dependent. Since the number of attribute values has no effect on the result of interdependence redundancy measure, it is perceived as a more ideal index to measure the dependence degree between different categorical attributes.

Utilizing the interdependence measure, we can maintain a $d \times d$ relationship matrix $\mathcal{R}$ to store the dependence degree of each pair of attributes. Each element $\mathcal{R}(i, j)$ of this matrix is given by $\mathcal{R}(i, j) = R(A_i; A_j)$. It is obvious that $\mathcal{R}$ is a symmetric matrix with all diagonal elements equal to 1. To consider the interdependent attributes simultaneously in distance measure, for each attribute A_r, we find out all the attributes that have obvious interdependence with it and store them in a set denoted as S_r. In particular, the set S_r is constructed by

$$S_r = \{A_i | \mathcal{R}(A_r; A_i) > \beta, 1 \le i \le d\} \tag{39}$$

where β is a specific threshold. Subsequently, the distance metric for categorical values in considering the dependence relationship between different attributes can be defined as

$$D(x_{ir}, x_{jr}) = \sum_{A_l \in S_r} \mathcal{R}(r, l) D((x_{ir}, x_{il}), (x_{jr}, x_{jl})) \tag{40}$$

where

$$D((x_{ir}, x_{il}), (x_{jr}, x_{jl})) = \begin{cases} p((x_{ir}, x_{il}) = (x_{jr}, x_{jl})), & \text{if } x_{ir} \neq x_{jr} \\ \delta(x_{il}, x_{jl}) p((x_{ir}, x_{il}) = (x_{jr}, x_{jl})), & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{41}$$

The vector equality probability $p((x_{ir}, x_{il}) = (x_{jr}, x_{jl}))$ here is calculated by

$$\begin{aligned} p((x_{ir}, x_{il}) = (x_{jr}, x_{jl})) &= p(x_{ir}, x_{il}) \cdot p^{-}(x_{ir}, x_{il}) \\ &\quad + p(x_{jr}, x_{jl}) \cdot p^{-}(x_{jr}, x_{jl}) \\ &= p(A_r = x_{ir} \wedge A_l = x_{il} | X) \\ &\quad \cdot p^{-}(A_r = x_{ir} \wedge A_l = x_{il} | X) \\ &\quad + p(A_r = x_{jr} \wedge A_l = x_{jl} | X) \\ &\quad \cdot p^{-}(A_r = x_{jr} \wedge A_l = x_{jl} | X). \end{aligned} \tag{42}$$

Specially, when $A_l = A_r$, we have $\mathcal{R}(r, l) = 1$ and $D((x_{ir}, x_{il}), (x_{jr}, x_{jl})) = \delta(x_{ir}, x_{jr}) p(x_{ir} = x_{jr})$.

It can be observed that when we utilize the further defined metric to measure the distance between two categorical values from one attribute, not only the frequency probability of these two values, but also the co-occurrent probability of them with other values from highly correlated attributes are investigated. Moreover, if we assume that all the attributes are totally independent of each other, $\mathcal{R}$ will become an identity matrix and the set S_r will only contain one item A_r for all $r \in \{1, 2, \ldots, d\}$. Under the circumstances, (40) will degenerate to (25). That is, the distance metric defined by (25) is actually a special case of the one given by (40).

D. Algorithm for Distance Computation and Time Complexity Analysis

Based on the proposed distance metric, for the given categorical data set X, the algorithm to calculate the distance between each pair of objects can be summarized as Algorithm 2.

Next, we further analyze the time complexity of this algorithm. Since the proposed distance metric needs to calculate the joint probability of attribute values, as suggested in [14], we can utilize $(1/2)d(d-1)$ matrices to store the co-occurrence of the values between any pair of attributes. A complete scan of the entire data set is needed to built these matrices and the time cost is $O(d^2m^2n)$, where m is the average number of different values that can be chosen by each attribute. Subsequently, the computational cost of constructing the relationship matrix $\mathcal{R}$ is $O(d^2m^2)$. Given two data objects, calculating the distance along one attribute A_r according to (40) needs $O(d_r)$ time, where $d_r = |S_r|$ stands for the number of elements in the set S_r. Therefore, the computational cost needed to calculate the distance between any pair of objects is $O(d\widetilde{d})$, where $\widetilde{d} = (1/d)\sum_{r=1}^{d} |S_r|$. In conclusion, the time cost of Algorithm 2 is $O(d^2m^2n + d^2m^2 + d\widetilde{d})$. From the practical view point, we have $\widetilde{d} < d$, and m usually is a constant. Consequently, the time complexity of proposed method is $O(d^2n)$. This is the same as the metrics proposed in [5], [12], and [14].

IV. Experiments

To investigate the effectiveness of the unsupervised distance metric for the categorical data proposed in this paper, two different kinds of experiments have been conducted on six real data sets in comparison with the existing distance metrics. The first experiment is to validate the ability of the

Algorithm 2 Distance Calculation for Categorical Data

```
Input: data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
Output: $D(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j \in \{1, 2, \ldots, n\}$
Calculate $p_s(A_r)$ and $p_f(A_r)$ for each attribute $A_r$ according to Eq. (27) and Eq. (28).
For each pair of attributes $(A_r, A_l)$ $(r, l \in \{1, 2, \ldots, d\})$, calculate $R(A_r; A_l)$ according to Eq. (37).
Construct the relationship matrix $\mathcal{R}$.
Get the index set $S_r$ for each attribute $A_r$ by $S_r = \{l | \mathcal{R}(r, l) > \beta, 1 \le l \le d\}$.
Choose two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ from $X$.
Let $D(\mathbf{x}_i, \mathbf{x}_j) = 0$ and $w = 0$.
for $r = 1$ to $d$ do
    if $x_{ir} \neq x_{jr}$ then
        $D(x_{ir}, x_{jr}) = \sum_{l \in S_r} \mathcal{R}(r, l) p((x_{ir}, x_{il}) = (x_{jr}, x_{jl}))$
        $w_r = p_s(A_r)$
    else
        $D(x_{ir}, x_{jr}) = \sum_{l \in S_r} R(r, l) \delta(x_{il}, x_{jl}) p((x_{ir}, x_{il}) = (x_{jr}, x_{jl}))$
        $w_r = p_f(A_r)$
    end if
    $w = w + w_r$
    $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_i, \mathbf{x}_j) + w_r D(x_{ir}, x_{jr})$
end for
$D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_i, \mathbf{x}_j)/w$
```

proposed distance metric in discriminating different clusters and the other one is to investigate its effectiveness in the unsupervised clustering analysis.

A. Cluster Discrimination

It is known that a cluster partition on a data set is to make sure that the similarities between objects in the same cluster are high while the similarities between objects in different clusters are low. As distance metric is a kind of important and frequently used dissimilarity metric, its ability in the cluster discrimination is a significant criterion to evaluate its effectiveness in the data analysis. That is, given a data set with true class labels, a good distance metric should make the intracluster distances as small as possible and the intercluster distances as large as possible. Therefore, to investigate the cluster discrimination ability of proposed distance metric, we utilized it to calculate the average intracluster and intercluster distances for some categorical data sets from the UCI Machine Learning Data Repository (URL: http://archive.ics.uci.edu/ml/). According to [12], for a cluster C_r of data set X with n_r objects, the average intracluster distance is calculated by

$$\text{AAD}(C_r) = \frac{\sum_{\mathbf{x}_i \in C_r} \sum_{\mathbf{x}_j \in C_r} D(\mathbf{x}_i, \mathbf{x}_j)}{n_r^2}.$$

Moreover, for every two clusters C_r with n_r objects and C_t with n_t objects, the average intercluster distance is given by

$$\text{AED}(C_r, C_t) = \frac{\sum_{\mathbf{x}_i \in C_r} \sum_{\mathbf{x}_j \in C_t} D(\mathbf{x}_i, \mathbf{x}_j)}{n_r n_t}.$$

In addition, since the distances calculated with the different metrics usually have the different scales, it is better to normalize the result with the maximum distance value obtained on the data set to ensure a fair comparison. Furthermore, in our experiments, the value of the threshold parameter β in the proposed metric was set equal to the average interdependence redundancy of all attribute pairs. That is, we let $\beta = \beta_0$, where β_0 is calculated by

$$\beta = \frac{1}{d^2} \sum_{i=1}^{d} \sum_{j=1}^{d} R(A_i; A_j).$$

The information of the data sets we utilized is as follows.

1) *Congressional Voting Records Data Set:* There are 435 votes based on 16 key features and each vote comes from one of the two different party affiliations: 1) *democrat* (267 votes) and 2) *republican* (168 votes).
2) *Wisconsin Breast Cancer Database (WBCD):* This data set has 699 instances described by nine categorical attributes with the values from 1 to 10. Each instance belongs to one of the two clusters labeled by *benign* (contains 458 instances) and *malignant* (contains 241 instances).
3) *Mushroom Data Set:* It contains hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota family. Each sample is described by 22 attributes and labeled with edible or poisonous. In total, 8124 samples are included, and 2480 of them have missing attribute values. In our experiments, the 5644 samples without missing values have been adopted.
4) *Small Soybean Database:* There are 47 instances characterized by 35 multivalued categorical attributes. According to the different kinds of diseases, all the instances should be divided into four groups.
5) *Car Evaluation Database:* It contains 1728 car samples derived from a simple decision model that evaluates cars according to six different aspects. Each sample is

TABLE III

AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE VOTING DATA SET

Hamming distance metric			Proposed distance metric		
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.4330	0.6757	C_1	0.3542	0.6380
C_2	0.6757	0.3125	C_2	0.6380	0.2237

TABLE IV

AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE WBCD DATA SET

Hamming distance metric			Proposed distance metric		
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.3796	0.8716	C_1	0.1699	0.6380
C_2	0.8716	0.8128	C_2	0.6380	0.2655

TABLE V

AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE MUSHROOM DATA SET

Hamming distance metric			Proposed distance metric		
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.4700	0.6414	C_1	0.3882	0.5774
C_2	0.6414	0.4945	C_2	0.5774	0.3876

labeled with one of the four categories: 1) unacceptable; 2) acceptable; 3) good; and 4) very good.

6) *Zoo Data Set:* This data set consists of 101 instances represented by 16 attributes, in which each instance belongs to one of the seven animal categories.

The average intracluster distance of each cluster and the average intercluster distance between each pair of clusters obtained by the proposed distance metric on the six data sets have been presented in Tables III–VIII. For a comparative study, the results obtained by the Hamming distance metric have also been listed in the tables. It can be roughly observed from these tables that, for the Soybean and Zoo data sets, the average intracluster distances calculated based on the proposed distance metric have a significant decrease in comparison with those obtained by the Hamming distance, while the intercluster distances obtained by these two metrics are comparable. Moreover, although the intercluster and intracluster distances obtained by proposed metric are all smaller than that obtained by the Hamming distance on the other four data sets, the differences between intracluster and intercluster distances in the result of the proposed metric are larger than that of the Hamming distance. This indicates that the proposed distance metric can better distinguish the different clusters in these data sets.

Furthermore, to present the experimental result simply and clearly, we proposed a new criterion, namely, cluster-discrimination index (CDI) based on the average intracluster and intercluster distance. For a data set with k clusters, the value of this index was calculated by

$$\text{CDI} = \frac{1}{k}\sum_{r=1}^{k}\frac{\text{AAD}(C_r)}{\frac{1}{k-1}\sum_{t\neq r}\text{AED}(C_r, C_t)}.$$

That is, the value of CDI is determined by the average ratio of intracluster distance to the intercluster distance. In general, a smaller value of CDI indicates a better discrimination on the cluster structure of the data set. Table IX records the CDI values obtained by different distance metrics on each data set. In Table IX, Ahmad's distance, ABDM, and DILCA stand for the distance metrics presented in [12], [5], and [14], respectively. DM1, DM2, and DM3 are the three cases of proposed distance metric. In particular, DM1 means the distance metric defined by (26), DM2 means the distance metric expressed by (29), which is adjusted by the dynamic weights without considering the relationship between attributes, and DM3 stands for the complete distance metric calculated by Algorithm 2. The values highlighted in bold imply the best results among the seven metrics on each data set.

It can be found from the table that the DM3 metric has obtained the best result on five tested data sets and the average improvement is over 24% in comparison with the Hamming distance metric. Both without considering the attribute interdependence and weights, the average performance of DM1 metric is still over 12% better than the Hamming distance. This result indicates that quantifying distance between categorical values with frequency probability rather than constant is more reasonable for the analysis of relationship between categorical objects. Comparing the performance of DM2 and DM3, we can find that the information of interdependence between attributes is important for distance measurement. Making a good use of this information can significantly improve the effectiveness of the learning method on categorical data. It can also be observed that the DM2 and DM3 metrics have very similar results on the WBCD data set and have the same performance on the Car data sets. This is because the dependence degree between attributes in these two data sets is very low such that there is only one pair of attributes, whose value of the interdependence redundancy measure has exceeded the threshold β_0 in the WBCD data set and the dependence degree of every pair of different attributes in the Car data set is smaller than the setting value of β.

Moreover, the performance of Ahmad's distance metric also has a significant improvement compared with the Hamming distance on Voting, WBCD, Soybean, and Zoo data sets, and it has obtained the best result on the Zoo data set, while the result of DM3 is in the second place on it. However, for the Car data set whose relationship matrix $\mathcal{R}$ is an identity matrix, i.e., each attribute is statistically independent of all the other attributes, Ahmad's distance as well as ABDM and DILCA metrics cannot get the CDI result as they quantified the distance between every pair of data samples as 0. This result is consistent with the analysis in Section II-C. In addition, for the Mushroom data set, the average intracluster and intercluster distances obtained by Ahmad's metric are as follows:

$$\begin{aligned}\text{AAD}(C_1) = 0.2831, \quad \text{AED}(C_1, C_2) = 0.2536\\ \text{AED}(C_1, C_2) = 0.2536, \quad \text{AAD}(C_2) = 0.2944.\end{aligned}$$

It can be observed that the average intracluster distances are larger than the intercluster distance. Therefore, the obtained

TABLE VI
AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE SOYBEAN DATA SET

Hamming distance metric					Proposed distance metric				
Clusters	C_1	C_2	C_3	C_4	Clusters	C_1	C_2	C_3	C_4
C_1	0.2368	0.6632	0.6011	0.6421	C_1	0.1095	0.6149	0.5233	0.5651
C_2	0.6632	0.2379	0.8237	0.7616	C_2	0.6149	0.0744	0.8877	0.8287
C_3	0.6011	0.8237	0.2463	0.4985	C_3	0.5233	0.8877	0.1392	0.3752
C_4	0.6421	0.7616	0.4985	0.2968	C_4	0.5651	0.8287	0.3752	0.1839

TABLE VII
AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE CAR DATA SET

Hamming distance metric					Proposed distance metric				
Clusters	C_1	C_2	C_3	C_4	Clusters	C_1	C_2	C_3	C_4
C_1	0.6958	0.7433	0.7423	0.7507	C_1	0.5165	0.5688	0.5753	0.5881
C_2	0.7344	0.6486	0.6503	0.6366	C_2	0.5688	0.4526	0.4526	0.4306
C_3	0.7423	0.6503	0.5489	0.5886	C_3	0.5753	0.4526	0.3721	0.3961
C_4	0.7507	0.6366	0.5886	0.4707	C_4	0.5881	0.4306	0.3961	0.2612

TABLE VIII
AVERAGE INTRACLUSTER/INTERCLUSTER DISTANCE OBTAINED BY THE DIFFERENT METRICS ON THE ZOO DATA SET

Hamming distance metric								Proposed distance metric							
Clusters	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Clusters	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	0.18	0.60	0.44	0.59	0.47	0.65	0.68	C_1	0.17	0.73	0.52	0.67	0.57	0.72	0.77
C_2	0.60	0.14	0.42	0.55	0.46	0.46	0.52	C_2	0.73	0.11	0.44	0.55	0.51	0.42	0.48
C_3	0.44	0.42	0.21	0.33	0.27	0.51	0.42	C_3	0.52	0.44	0.23	0.33	0.30	0.51	0.44
C_4	0.59	0.55	0.33	0.08	0.34	0.70	0.45	C_4	0.67	0.55	0.33	0.06	0.34	0.68	0.46
C_5	0.47	0.46	0.27	0.34	0.08	0.48	0.12	C_5	0.57	0.51	0.30	0.34	0.07	0.52	0.41
C_6	0.65	0.46	0.51	0.69	0.48	0.12	0.35	C_6	0.72	0.42	0.51	0.68	0.52	0.12	0.32
C_7	0.68	0.52	0.42	0.45	0.37	0.35	0.21	C_7	0.77	0.48	0.44	0.46	0.41	0.32	0.17

TABLE IX
CDI OBTAINED BY THE DIFFERENT METRICS ON FOUR REAL DATA SETS

Data sets	Hamming Distance	Ahmad's Distance	ABDM	DILCA	DM1	DM2	DM3
Voting	0.5517	0.4660	0.7235	0.4920	0.5198	0.5030	**0.4529**
WBCD	0.6840	0.5031	1.6688	0.3437	0.3734	0.3424	**0.3374**
Mushroom	0.7519	1.1388	1.7336	0.6800	0.7150	0.6914	**0.6717**
Soybean	0.3856	0.2511	0.2930	0.2531	0.3174	0.2545	**0.2086**
Car	0.8614	-	-	-	0.8477	**0.7919**	**0.7919**
Zoo	0.3045	**0.2618**	0.4849	0.2809	0.3091	0.2960	0.2656

CDI value is larger than 1, which is inconsistent with the property of clusters. Similar results have also been obtained by the ABDM method on the WBCD and Mushroom data sets. By contrast, the DILCA metric, which is also an indirect measure, had much better performance on these two data sets. The reason may be that the DILCA method does not use all the other attributes to represent the current one, but only selects the most relevant attributes. This implies that considering irrelevant attributes together may degrade the performance of distance metric.

B. Study of the Threshold Parameter

In the proposed distance metric, we have a threshold parameter β to be set in advance. In general, the value of β has effect on the number of attributes that should be jointly considered in the distance calculation. In particular, a too small β will result in many attributes with insignificant interdependence relationship being jointly considered. The dependence information between these attributes actually has negligible contribution to the distance measure, and will lead an unnecessarily increase in the computation load. By contrast, a too large value of β will lead to the loss of useful dependence information and degrade the contribution of correlated attributes to the distance measure. In practice, we find that an appropriate selection is to let β be equal to the average interdependence redundancy of all attribute pairs, i.e., β_0.

Furthermore, to experimentally investigate the impact of the threshold parameter β on the effectiveness of proposed distance metric, we have utilized the DM3 metric with the different values of β to calculate the intracluster and intercluster distances for the six data sets. The curves that depict the changing trend of obtained CDI values with increasing β have been shown in Fig. 1. From Fig. 1, we can find that, when β is very small (e.g., $\beta < 0.1$), the performance of

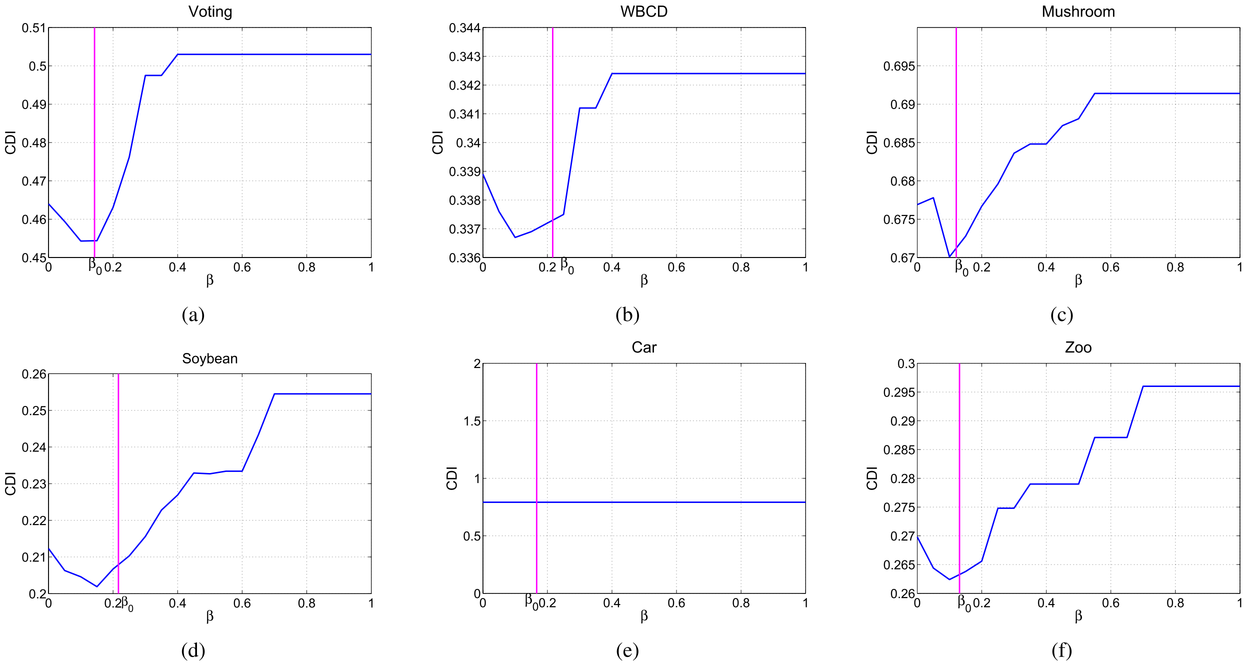


Fig. 1. CDI obtained by the proposed metric with the different values of β on (a) Voting data set, (b) WBCD data set, (c) Mushroom data set, (d) Soybean data set, (e) Car data set, and (f) Zoo data set.

TABLE X

CLUSTERING ERRORS OBTAINED BY k-MODES ALGORITHM WITH THE DIFFERENT DISTANCE METRICS

Data sets	Hamming Distance	Ahmad's Distance	ABDM	DILCA	DM3
Voting	0.1387±0.0065	0.1250±0.0022	0.1864±0.1180	**0.1195**±0.0032	0.1217±0.0026
WBCD	0.1650±0.1594	0.1264±0.0890	0.1230±0.0590	0.1164±0.1151	**0.0902**±0.0891
Mushroom	0.2932±0.1596	0.4595±0.0406	0.3258±0.1284	0.2450±0.1359	**0.2357**±0.0851
Soybean	0.1830±0.1430	0.1609±0.1580	0.2736±0.1459	0.1809±0.1689	**0.1334**±0.1285
Car	0.6410±0.0381	-	-	-	**0.6278**±0.0283
Zoo	0.2998±0.1024	**0.2529**±0.0884	0.2885±0.1053	0.2624±0.0979	0.2563±0.0898

DM3 metric generally improves as the value of β increases. This is because, when the threshold β is too small, many useless relationships between attributes are considered, which will degrade the accuracy of obtained object distances. By contrast, when β is large to a certain degree (e.g., $\beta > 0.3$), the performance of DM3 metric often degrades obviously as β increases. Here, Car is a special data set, in which the attributes are totally independent of each other. Therefore, changing the value of β has no influence on the performance of DM3 on it. Moreover, the performance of DM3 with β equal to β_0 has also been indicated in Fig. 1. Overall, β_0 is a good choice for β as it can get satisfying practical effectiveness without spending useless computations.

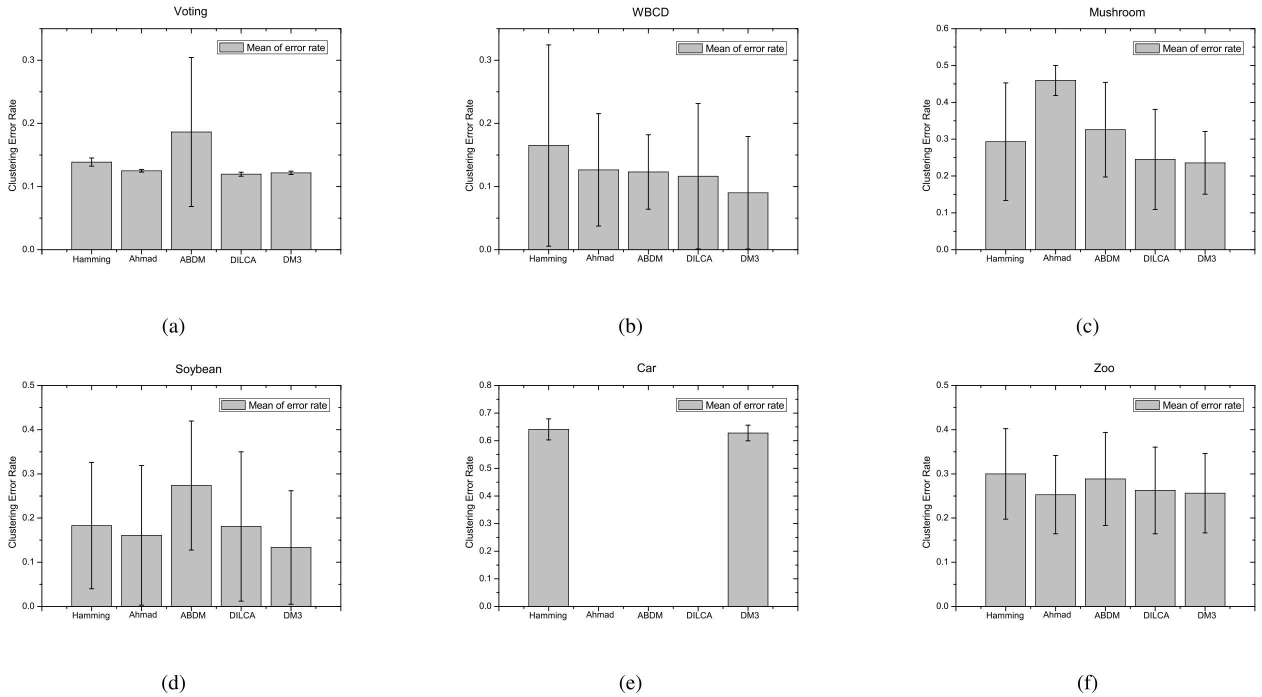
C. Clustering Analysis

In general, clustering analysis based on distance measure is to partition the given objects into several clusters such that the distances between objects in the same cluster are small while the distances between objects in different clusters are large. That is, the distance metric plays a key role in clustering accuracy. Therefore, in this experiment, we further investigated the effectiveness of the proposed distance metric by embedding it into the framework of the k-modes algorithm [45], which is the most popular distance-based clustering method for purely categorical data, and comparing its clustering result with the original k-modes method (i.e., the k-modes algorithm with the Hamming distance metric) and the k-modes algorithm with Ahmad's distance, ABDM, and DILCA metrics. According to [46], the clustering accuracy is a direct criterion to evaluate clustering result, which is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(c_i, \mathrm{map}(l_i))}{n}$$

where c_i stands for the provided label, $\mathrm{map}(l_i)$ is a mapping function that maps the obtained cluster label l_i to the equivalent label from the data corpus, and the delta function $\delta(c_i, \mathrm{map}(l_i)) = 1$ only if $c_i = \mathrm{map}(l_i)$, otherwise 0. Correspondingly, the clustering error rate is computed as $e = 1 - \mathrm{ACC}$.

In our experiments, the clustering analysis was conducted on the six categorical data sets: 1) Voting; 2) WBCD; 3) Mushroom; 4) Soybean; 5) Car; and 6) Zoo. Each algorithm has been executed 50 times on every data set, and the average clustering error rate as well as the standard deviation in error has been recorded in Table X. Moreover, the graphical representation of the clustering results for the five methods



(a) (b) (c)

(d) (e) (f)

Fig. 2. Graphical representation of clustering error rate and standard deviation for different methods on (a) Voting data set, (b) WBCD data set, (c) Mushroom data set, (d) Soybean data set, (e) Car data set, and (f) Zoo data set.

is shown in Fig. 2. It can be observed that, for distance-based clustering on categorical data, the k-modes algorithm with the proposed distance metric has a competitive advantage in terms of clustering accuracy compared with the other four methods. DM3 has obtained the best result on four data sets. The average improvement in clustering accuracy on these six data sets obtained by DM3 metric is over 20% in comparison with the Hamming distance. In addition, although the k-modes algorithm with Ahmad's distance metric is superior to the original k-modes method on Voting, WBCD, and Soybean data sets and gets the best result on Zoo data set, its performance degrades significantly on the other two data sets. In particular, for the Car data set, since the distance between each pair of data objects has been estimated as 0 according to Ahmad's, ABDM, and DILCA metrics, the k-modes algorithm based on them could not get a reasonable result, as they had classified all the objects into a single cluster.

In addition, to make comprehensive evaluation for these clustering algorithms' performance, three more popular validity indices, namely, rand index (RI), normalized mutual information (NMI), and Davies–Bouldin Index (DBI), were further adopted in this paper. Among these indices, RI and NMI are external criteria, whereas DBI belongs to internal criteria. The definitions of these three indices are as follows.

1) *Rand Index:*

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

2) *Normalized Mutual Information:*

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{c} n_{i,j} \log\left(\frac{n \cdot n_{i,j}}{n_i \cdot n_j}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i \log \frac{n_i}{n}\right)\left(\sum_{j=1}^{c} n_j \log \frac{n_j}{n}\right)}}$$

where c stands for the true number of classes, k is the number of clusters obtained by the algorithm, $n_{i,j}$ denotes the number of agreements between cluster i and class j, n_i is the number of data objects in cluster i, n_j is the number of objects in class j, and n is the number of objects in the whole data set.

3) *Davies–Bouldin Index:*

$$\mathrm{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\Delta_i + \Delta_j}{D(\mu_i, \mu_j)} \right)$$

where μ_i is the mode of cluster i, Δ_i is the average distance of all objects in cluster i to cluster mode μ_i, and $D(\mu_i, \mu_j)$ denotes the distance between cluster modes μ_i and μ_j.

In general, both RI and NMI have values from interval $[0, 1]$ and larger values of them indicate better clustering performance. In contrast, a smaller value of DBI is usually preferred. Nevertheless, similar to other internal criteria, DBI has the potential drawback that a good value does not necessarily imply a better clustering result. The evaluations of clustering outcomes obtained by the k-modes algorithm with different distance metrics have been listed in Tables XI–XIII in the form of the means and standard deviations of RI, NMI, and DBI, respectively. From the statistical result, we can find that the performance of the k-modes algorithm with DM3 is also superior to the other algorithms according to these

TABLE XI

CLUSTERING PERFORMANCE IN TERMS OF RI OF *k*-MODES ALGORITHM WITH THE DIFFERENT DISTANCE METRICS

Data sets	Hamming Distance	Ahmad's Distance	ABDM	DILCA	DM3
Voting	0.7599±0.0098	0.7808±0.0033	0.6856±0.1324	**0.7890**±0.0022	0.7823±0.0016
WBCD	0.7768±0.1525	0.7877±0.1024	0.7894±0.0787	0.8353±0.1298	**0.8827**±0.0752
Mushroom	0.6209±0.1133	0.5048±0.0075	0.6025±0.1134	0.6661±0.1092	**0.6732**±0.0880
Soybean	0.8863±0.0744	0.9029±0.1084	0.7900±0.1213	0.9059±0.0801	**0.9314**±0.0758
Car	0.4905±0.0174	-	-	-	**0.5059**±0.0123
Zoo	0.8822±0.0492	0.9029±0.0440	0.8535±0.0786	0.8991±0.0490	**0.9064**±0.0450

TABLE XII

CLUSTERING PERFORMANCE IN TERMS OF NMI OF *k*-MODES ALGORITHM WITH THE DIFFERENT DISTANCE METRICS

Data sets	Hamming Distance	Ahmad's Distance	ABDM	DILCA	DM3
Voting	0.4483±0.0211	0.4908±0.0098	0.3048±0.2475	**0.5161**±0.0034	0.4987±0.0078
WBCD	0.4788±0.2435	0.5116±0.1715	0.4854±0.1157	0.6208±0.1818	**0.6917**±0.1304
Mushroom	0.2556±0.1877	0.0027±0.0026	0.2221±0.2154	0.3081±0.1714	**0.3182**±0.1372
Soybean	0.8183±0.1175	0.8749±0.1335	0.7188±0.1527	0.8743±0.1075	**0.8991**±0.1089
Car	0.0467±0.0241	-	-	-	**0.0725**±0.0253
Zoo	0.7615±0.0702	**0.8042**±0.0567	0.7432±0.0854	0.7917±0.0744	0.7927±0.0630

TABLE XIII

CLUSTERING PERFORMANCE IN TERMS OF DBI OF *k*-MODES ALGORITHM WITH THE DIFFERENT DISTANCE METRICS

Data sets	Hamming Distance	Ahmad's Distance	ABDM	DILCA	DM3
Voting	0.5576±0.0066	0.4039±0.0020	2.4576±2.3897	0.4579±0.0028	**0.3903**±0.0011
WBCD	1.5912±0.9618	1.4773±0.6859	3.4899±1.2946	0.5602±0.1284	**0.4058**±0.0310
Mushroom	1.2357±0.4301	**0.8043**±0.1546	15.9025±8.7200	0.9352±0.4968	0.8773±0.3868
Soybean	0.9733±0.1862	0.7665±0.2280	2.3589±1.6608	0.8491±0.3817	**0.7576**±0.1821
Car	2.0481±0.6170	-	-	-	**1.6078**±0.6015
Zoo	**1.0528**±0.2703	1.1216±0.4934	6.3431±6.2965	1.3701±0.7805	1.1417±0.6611

TABLE XIV

CLUSTERING ERRORS OBTAINED BY *k*-MODES ALGORITHM WITH PROPOSED DISTANCE METRICS

Data sets	DM1	DM2	DM3
Voting	0.1321±0.0045	0.1307±0.0046	0.1217±0.0026
WBCD	0.0987±0.0907	0.0981±0.0972	0.0902±0.0891
Mushroom	0.2649±0.0986	0.2483±0.1187	0.2357±0.0851
Soybean	0.1800±0.1604	0.1487±0.1321	0.1334±0.1285
Car	0.6329±0.0361	0.6279±0.0281	0.6278±0.0283
Zoo	0.2798±0.0936	0.2656±0.0864	0.2563±0.0898

three evaluation criteria, as it has obtained the best result in most cases.

Taken together, the results of this experiment indicate that the proposed distance metric is more appropriate for the unsupervised data analysis as it can better reveal the true relationship between categorical data objects. Moreover, to further investigate the effect of the different strategies in the proposed metric, we have compared the clustering results of DM1, DM2, and DM3 in Table XIV. We can find that both dynamic attribute weights and relevant attribute analysis have improved the performance of the proposed distance metric. In addition, the simple metric DM1 still had much better performance than the Hamming distance in the clustering analysis. This has validated the effectiveness of the proposed distance definition.

V. CONCLUSION

In this paper, we have presented a new distance metric, which measures the distance between categorical data based on the frequency probability of each attribute value in the whole data set. Dynamic weight has been designed to adjust the contribution of each attribute distance to the whole object distance. Moreover, the interdependence redundancy measure has been utilized to evaluate the dependency degree between each pair of attributes. Subsequently, the distance between two values from one attribute is not only measured by their own frequency probabilities, but also determined by the values of other attributes that have high interdependence with the calculated one. Experiments on benchmark data sets have shown the effectiveness of the proposed metric in comparison with the existing counterparts.

REFERENCES

[1] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proc. 8th SIAM Int. Conf. Data Mining*, Atlanta, GA, USA, Apr. 2008, pp. 243–254.

[2] B. Liu, M. Wang, R. Hong, Z. Zha, and X.-S. Hua, "Joint learning of labels and distance metric," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 3, pp. 973–978, Jun. 2010.

[3] C. Shen, J. Kim, F. Liu, L. Wang, and A. van den Hengel, "Efficient dual approach to distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 394–406, Feb. 2014.

[4] F. Esposito, D. Malerba, V. Tamma, and H.-H. Bock, "Classical resemblance measures," in *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information From Complex Data*, H.-H. Bock and E. Diday, Eds. Berlin, Germany: Springer-Verlag, 2002, pp. 139–152.

[5] S. Q. Le and T. B. Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2549–2557, 2005.

[6] Z. Hubálek, "Coefficients of association and similarity, based on binary (presence–absence) data: An evaluation," *Biol. Rev.*, vol. 57, no. 4, pp. 669–689, 1982.

[7] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *J. Classification*, vol. 3, no. 1, pp. 5–48, 1986.

[8] V. Batagelj and M. Bren, "Comparing resemblance measures," *J. Classification*, vol. 12, no. 1, pp. 73–90, 1995.

[9] C.-C. Hsu and S.-H. Wang, "An integrated framework for visualized and exploratory pattern discovery in mixed data," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 161–173, Feb. 2006.

[10] C.-C. Hsu, "Generalizing self-organizing map for categorical data," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 294–304, Mar. 2006.

[11] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Mach. Learn.*, vol. 10, no. 1, pp. 57–78, 1993.

[12] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognit. Lett.*, vol. 28, no. 1, pp. 110–118, 2007.

[13] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in *Proc. 8th Int. Symp. Intell. Data Anal.*, Lyon, France, Aug. 2009, pp. 83–94.

[14] D. Ienco, R. G. Pensa, and R. Meo, "From context to distance: Learning dissimilarity for categorical data clustering," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, 2012, Art. ID 1.

[15] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, no. 4, pp. 882–907, 1966.

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

[17] K. C. Gowda and E. Diday, "Unsupervised learning through symbolic clustering," *Pattern Recognit. Lett.*, vol. 12, no. 5, pp. 259–264, 1991.

[18] K. C. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 2, pp. 368–378, Mar./Apr. 1992.

[19] M. Ichino and H. Yaguchi, "Generalized Minkowski metrics for mixed feature-type data analysis," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 4, pp. 698–708, Apr. 1994.

[20] F. de A. T. de Carvalho, "Proximity coefficients between Boolean symbolic objects," in *New Approaches in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization*, vol. 5, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Eds. Berlin, Germany: Springer-Verlag, 1994, pp. 387–394.

[21] F. de A. T. de Carvalho, "Extension based proximities between constrained Boolean symbolic objects," in *Data Science, Classification, and Related Methods*, E. C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, and Y. Baba, Eds. Tokyo, Japan: Springer-Verlag, 1998, pp. 370–378.

[22] Y.-M. Cheung and H. Jia, "A unified metric for categorical and numerical attributes in data clustering," in *Proc. 17th Pacific-Asia Conf. Knowl. Discovery Data Mining*, Gold Coast, QLD, Australia, Apr. 2013, pp. 135–146.

[23] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognit.*, vol. 46, no. 8, pp. 2228–2238, 2013.

[24] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15. Vancouver, BC, Canada, Dec. 2002, pp. 505–512.

[25] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2012.

[26] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data Mining Knowl. Discovery*, vol. 29, no. 2, pp. 534–564, 2015.

[27] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. Vancouver, BC, Canada, Dec. 2004, pp. 513–520.

[28] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18. Vancouver, BC, Canada, Dec. 2006, pp. 1473–1480.

[29] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 209–216.

[30] G. Niu, B. Dai, M. Yamada, and M. Sugiyama, "Information-theoretic semi-supervised metric learning via entropy regularization," in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, U.K., Jun. 2012, pp. 89–96.

[31] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "A new kernelization framework for Mahalanobis distance learning algorithms," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1570–1579, 2010.

[32] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet, "Multi-class object localization by combining local contextual interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 113–120.

[33] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and kernel learning using a linear transformation," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 519–547, 2012.

[34] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 539–546.

[35] D. Kedem, S. Tyree, K. Q. Weinberger, and F. Sha, "Non-linear metric learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25. Lake Tahoe, NV, USA, Dec. 2012, pp. 2582–2590.

[36] Y. He, W. Chen, Y. Chen, and Y. Mao, "Kernel density metric learning," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 271–280.

[37] W.-H. Au, K. C. C. Chan, A. K. Wong, and W. Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 2, no. 2, pp. 83–101, Apr./Jun. 2005.

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

[39] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Commun. ACM*, vol. 29, no. 12, pp. 1213–1228, 1986.

[40] A. Ahmad and L. Dey, "A feature selection technique for classificatory analysis," *Pattern Recognit. Lett.*, vol. 26, no. 1, pp. 43–56, 2005.

[41] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[42] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[43] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.

[44] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS–clustering categorical data using summaries," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, Aug. 1999, pp. 73–83.

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

[46] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17. Vancouver, BC, Canada, Dec. 2005, pp. 507–514.

Hong Jia received the M.Sc. degree from the School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2013.

She is currently a Post-Doctoral Research Fellow with the Department of Computer Science, Hong Kong Baptist University. Her current research interests include machine learning, data mining, and pattern recognition.

Yiu-ming Cheung (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung is a Senior Member of the Association for Computing Machinery. He is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He serves as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Knowledge and Information Systems*, and the *International Journal of Pattern Recognition and Artificial Intelligence*.

Jiming Liu (F'11) received the M.Eng. and Ph.D. degrees from McGill University, Montréal, QC, Canada.

He is currently the Chair Professor of Computer Science and the Associate Dean of the Faculty of Science (Research) with Hong Kong Baptist University, Hong Kong. His current research interests include data mining and data analytics, health informatics, computational epidemiology, complex systems, and collective intelligence.

Prof. Liu has served as the Editor-in-Chief of *Brain Informatics* (Springer) and *Web Intelligence* (IOS), and an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CYBERNETICS, *Big Data and Information Analytics* (AIMS), *Computational Intelligence* (Wiley), and *Neuroscience and Biomedical Engineering* (Bentham).