2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A New Distance Metric for Unsupervised Learning of Categorical Data

Hong Jia[a] and Yiu-ming Cheung[a,b]
[a]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[b]The United International College, BNU-HKBU, Zhuhai, China
Email: {hjia, ymc}@comp.hkbu.edu.hk

***Abstract*—Distance metric is the basis of many learning algorithms and its effectiveness usually has significant influence on the learning results. Generally, measuring distance for numerical data is a tractable task, but for categorical data sets, it could be a nontrivial problem. This paper therefore presents a new distance metric for categorical data based on the characteristics of categorical values. Specifically, the distance between two values from one attribute measured by this metric is determined by both of the frequency probabilities of these two values and the values of other attributes which have high interdependency with the calculated one. Promising experimental results on different real data sets have shown the effectiveness of proposed distance metric.**

I. Introduction

Measuring the distance between two data objects plays an important role in many data mining and machine learning tasks, such as clustering, classification, feature selection, outlier detection, and so on. Generally, 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 significant influence on the performance of the whole learning method [4], [5]. 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 [1]. By contrast, measuring distance for categorical data is a more challenging problem as there is no explicit ordering information in categorical values and the only numerical operation that can be straightforwardly applied is the identical comparison operation [2]. Under the circumstances, a simplest way to overcome this problem is to transform the categorical values into numerical ones, e.g. the binary strings [3], [6], [7], 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 [8], [9]. Therefore, it is desirable to solve this problem by proposing new distance metric for categorical data based on the characteristics of categorical values.

Among the existing work, the most straightforward and widely used distance metric for categorical data is the Hamming distance [1], in which the distance between different categorical values is set at 1 while a distance of 0 is assigned to identical values. Then, for a pair of categorical data objects with multiple attributes, the Hamming distance between them will be equal to the number of attributes in which they mismatch. 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 considered equally and the statistical properties of different values have not been distinguished [10]. For this reason, more researchers attempt to measure the distance for categorical data with the distribution characteristics of categorical values. For example, Cost and Salzberg [11] had proposed a distance metric namely Modified Value Distance Matrix (MVDM) for supervised learning task, in which the distance between two categorical values is calculated with respect to the class label of the data set. Additionally, for unsupervised distance measure of categorical data, Le and Ho [2] presented an indirect method which 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 by [12], in which the distance of two values from one attribute is quantified with respect to the co-occurrence probabilities of the values from all the other attributes with these two values.

Besides of the aforementioned methods which directly propose special distance metric for categorical data sets, some similarity measures [13], [14], [15], [16], [17], [18], [19] 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 [13] 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. Subsequently, the similarities between pairs of values are integrated with Lancaster's method [20] to estimate the similarity between data objects. Moreover, Gowda and Diday [14], [15], [16] have proposed an algebraic method to measure the similarity between categorical data. In this method, the similarity between two attribute values is defined based on three components: position, span, and content. Here, the component "position" works only when the attribute type is quantitative, the "span" indicates the relative sizes of the attribute values without referring to common parts between them, and the "content" is to measure the common parts between attribute values. Finally, the summation of these three similarity components is the estimate of the similarity between given attribute values.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

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 unsupervised learning environment. This distance metric is presented based on the characteristics of categorical values and the core idea is to measure the distance with frequency probability of each attribute value in the whole data set. Moreover, in order to well utilize the useful relationship information accompanying with each pair of attributes, the interdependence redundancy measure [24] has been introduced to evaluate the dependent 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 which are highly correlated with this one. The effectiveness of the proposed metric has been experimentally investigated on different real data sets in terms of cluster discrimination and clustering analysis. Competitive results indicate that the proposed distance metric is appropriate for unsupervised learning on categorical data as it can well reveal the true relationship between categorical objects.

II. Proposed Distance Metric for Categorical Data

This section will propose a metric to well quantify the distance between categorical data for unsupervised clustering analysis. In this new distance metric, not only the characteristics of categorical value but also the relationship between different attributes will be taken into account.

A. Frequency Probability based Distance Metric

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 $dom(A_r)$ ($r = 1, 2, \ldots, d$), which contains all the possible values that can be chosen by this attribute. Since the value domains of categorical attributes are finite and unordered, the domain of A_r with m_r elements can be expressed as $dom(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rm_r}\}$ and for any $a, b \in dom(A_r)$, either $a = b$ or $a \neq b$ [21]. Subsequently, each object $\mathbf{x}_i$ can be denoted by a vector $(x_{i1}, x_{i2}, \ldots, x_{id})^T$, where $x_{ir} \in dom(A_r)$ and T is the transpose operator of a matrix.

Generally, the distance between two categorical data objects $\mathbf{x}_i$ and $\mathbf{x}_j$ can be calculated by

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

Therefore, the key point is to define the distance between two categorical values. To this end, we first consider the characteristic of clustering analysis on categorical data, which is the learning task we mainly focus on in this paper. Generally, good prediction on new arriving data is an important goal of clustering analysis. For purely categorical data, the lower uncertainty the cluster structure has, the higher predictive accuracy can usually be achieved. Therefore, to reduce the uncertainty of the samples in each cluster, along one attribute, the objects with the different dominative attribute values tend to be divided into different clusters. This implies that two different categorical values both with high frequency from one attribute should have larger distance, while the distance between an infrequent value and others should be smaller. Consequently, the distance between categorical values can be defined based on frequency probability as follows:

$$D(x_{ir}, x_{jr}) = \begin{cases} p(A_r = x_{ir}|X) + p(A_r = x_{jr}|X), & \text{if } x_{ir} \neq x_{jr}, \\ 0, & \text{if } x_{ir} = x_{jr}, \end{cases} \tag{2}$$

where $i, j \in \{1, 2, \ldots, n\}$, $r \in \{1, 2, \ldots, d\}$, and the frequency probability $p(A_r = x_{ir}|X)$ is calculated by

$$p(A_r = x_{ir}|X) = \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)}. \tag{3}$$

Here, the operation $\sigma_{A_r = x_{ir}}(X)$ counts the number of objects in data set X that have the value x_{ir} for attribute A_r and the symbol NULL refers to the empty. Subsequently, the expression of distance between categorical data $\mathbf{x}_i$ and $\mathbf{x}_j$ can be written as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{d} [\delta(x_{ir}, x_{jr})(p(A_r = x_{ir}|X) + p(A_r = x_{jr}|X))], \tag{4}$$

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

$$\delta(x_{ir}, x_{jr}) = \begin{cases} 1, & \text{if } x_{ir} \neq x_{jr}, \\ 0, & \text{if } x_{ir} = x_{jr}. \end{cases} \tag{5}$$

It can be easily derived that the distance metric defined by Eq. (4) has the following properties:

(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. 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 which are highly dependent on each other. Under the circumstances, the computation of similarity or dissimilarity for categorical attributes in unsupervised learning task should be considered based on frequently co-occurring items [22]. That is, the distance between two values from one attribute should be calculated by taking into account the other attributes which are highly correlated with this one. Specifically, given the data set X, the dependent 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 [23] between them, which is defined as

$$I(A_i; A_j) = \sum_{r=1}^{m_i} \sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log \left(\frac{p(a_{ir}, a_{jl})}{p(a_{ir}) p(a_{jl})} \right). \tag{6}$$

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

$$p(a_{ir}) = p(A_i = a_{ir}|X) = \frac{\sigma_{A_i = a_{ir}}(X)}{\sigma_{A_i \neq NULL}(X)} \tag{7}$$

$$p(a_{jl}) = p(A_j = a_{jl}|X) = \frac{\sigma_{A_j=a_{jl}}(X)}{\sigma_{A_j \neq NULL}(X)}. \quad (8)$$

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

$$\begin{aligned} p(a_{ir}, a_{jl}) =& p(A_i = a_{ir} \wedge A_j = a_{jl}|X) \\ =& \frac{\sigma_{A_i=a_{ir} \wedge A_j=a_{jl}}(X)}{\sigma_{A_i \neq NULL \wedge A_j \neq NULL}(X)}. \end{aligned} \quad (9)$$

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 [23]. 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. [24] proposed to normalize the mutual information with joint entropy, which yields the interdependence redundancy measure denoted as

$$R(A_i; A_j) = \frac{I(A_i; A_j)}{H(A_i, A_j)}, \quad (10)$$

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

$$H(A_i, A_j) = -\sum_{r=1}^{m_i} \sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log[p(a_{ir}, a_{jl})]. \quad (11)$$

This interdependence redundancy measure evaluates the degree of deviation from independence between two attributes [24]. Specifically, $R(A_i; A_j) = 1$ 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 dependent degree between different categorical attributes.

Utilizing the interdependence measure, we can maintain a $d \times d$ relationship matrix $\mathcal{R}$ to store the dependent 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 take into account the interdependent attributes simultaneously in distance measure, for each attribute A_i we find out all the attributes that have obvious interdependency with it and store them in a set denoted as S_i. Specifically, the set S_i is constructed by

$$S_i = \{A_r | R(A_i; A_r) > \beta, 1 \leq r \leq d\}, \quad (12)$$

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

$$D(x_{ir}, x_{jr}) = \begin{cases} \sum\limits_{A_l \in S_r} \mathcal{R}(r, l)[p(x_{ir}, x_{il}) + p(x_{jr}, x_{jl})], \\ \qquad\qquad \text{if } x_{ir} \neq x_{jr}, \\ \sum\limits_{A_l \in S_r} \mathcal{R}(r, l)\delta(x_{il}, x_{jl})[p(x_{ir}, x_{il}) + p(x_{jr}, x_{jl})], \\ \qquad\qquad \text{if } x_{ir} = x_{jr}, \end{cases} \quad (13)$$

where $\delta(x_{il}, x_{jl})$ is defined by Eq. (5), and the joint probability $p(x_{ir}, x_{il})$ and $p(x_{jr}, x_{jl})$ are calculated by

$$p(x_{ir}, x_{il}) = p(A_r = x_{ir} \wedge A_l = x_{il}|X) \quad (14)$$

$$p(x_{jr}, x_{jl}) = p(A_r = x_{jr} \wedge A_l = x_{jl}|X). \quad (15)$$

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 attributes are totally independent with each other, $\mathcal{R}$ will become an identity matrix and the set S_i will only contain one item A_i for all $i \in \{1, 2, \ldots, d\}$. Under the circumstances, Eq. (13) will degenerate to Eq. (2). That is, the distance metric defined by Eq. (2) is actually a special case of the one given by Eq. (13).

C. Algorithm for Distance Computation

According to the newly defined distance measure, for the given categorical data set X, the algorithm to calculate the distance between each pair of objects can be summarized as Algorithm 1. Moreover, it can be observed that this algorithm has a threshold parameter β to be set in advance. Generally, the value of β has effect on the number of attributes that should be jointly considered in the distance calculation. Specifically, a too small β will result in many attributes with insignificant interdependence relationship being jointly considered. The dependency information between these attributes actually has negligible contribution to the distance measure and will lead an unnecessarily increase in computation load. By contrast, a too large value of β will lead to the loss of useful dependency information and degrade the contribution of correlated attributes to the distance measure. By a rule of thumb, we find that a value from $[0.1, 0.4]$ is more appropriate for the parameter β in practice.

Algorithm 1 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 $i, j \in \{1, 2, \ldots, n\}$
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. (10).
Construct the relationship matrix $\mathcal{R}$.
Get the index set S_r for each attribute A_r by $S_r = \{l | \mathcal{R}(r, l) > \beta, 1 \leq l \leq d\}$.
Choose two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ from X.
Let $D(\mathbf{x}_i, \mathbf{x}_j) = 0$.
for $r = 1$ **to** d **do**
  if $x_{ir} \neq x_{jr}$ **then**
    $D(x_{ir}, x_{jr}) = \sum\limits_{l \in S_r} \mathcal{R}(r, l)[p(x_{ir}, x_{il}) + p(x_{jr}, x_{jl})]$
  else
    $D(x_{ir}, x_{jr}) = \sum\limits_{l \in S_r} R(r, l)\delta(x_{il}, x_{jl})$
      $[p(x_{ir}, x_{il}) + p(x_{jr}, x_{jl})]$
  end if
  $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_i, \mathbf{x}_j) + D(x_{ir}, x_{jr})$
end for

III. EXPERIMENTS

To investigate the effectiveness of the unsupervised distance metric for categorical data proposed in this paper, two different kinds of experiments have been conducted on four real data sets in comparison with existing distance metric. The first experiment is to validate the ability of the proposed distance metric in discriminating different clusters and the other one is to investigate its effectiveness in 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 cluster discrimination is a significant criterion to evaluate its effectiveness in data analysis. That is, given a data set with true class labels, a good distance metric should make the intra-cluster distances as small as possible and the inter-cluster distances as large as possible. Therefore, to investigate the cluster-discrimination ability of proposed distance metric, we utilized it to calculate the average intra-cluster and inter-cluster 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 intra-cluster distance is calculated by

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

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

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

Additionally, 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. In our experiments, the value of β was set at 0.2 and the information of the data sets we utilized is as follows:

- *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: *democrat* (267 votes) and *republican* (168 votes).
- *Wisconsin Breast Cancer Database (WBCD)*: This data set has 699 instances described by 9 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).
- *Small Soybean Database*: There are 47 instances characterized by 35 multi-valued categorical attributes. According to the different kind of diseases, all the instances should be divided into four groups.
- *Zoo Data Set*: This data set consists of 101 instances represented by 16 attributes, in which each instance belongs to one of the 7 animal categories.

The average intra-cluster distance of each cluster and the average inter-cluster distance between each pair of clusters obtained by the proposed distance metric on the four data sets have been presented in Tables I–IV. For 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 the average intra-cluster distances calculated based on the proposed distance metric have a significant decrease in comparison with that obtained by Hamming distance while the inter-cluster distances obtained by these two metrics are comparable. Moreover, although the inter-cluster distances obtained by Hamming distance are slightly larger than that obtained by proposed metric on Voting and WBCD data sets as shown in Table I and Table II, the difference between intra-cluster and inter-cluster distances in the result of the proposed metric is larger than that of Hamming distance. This indicates that the proposed distance metric can better distinguish the different clusters in these two data sets.

TABLE I. AVERAGE INTRA/INTER-CLUSTER 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	0.3806	0.6630
C_2	0.6757	0.3125	C_2	0.6630	0.2531

TABLE II. AVERAGE INTRA/INTER-CLUSTER 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	0.3059	0.7001
C_2	0.8716	0.8128	C_2	0.7001	0.3106

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

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

That is, the value of CDI is determined by the average ratio of intra-cluster distance to the inter-cluster distance. Generally, a smaller value of CDI indicates a better discrimination on the cluster structure of the data set. Table V records the CDI values obtained by different distance metrics on each data set. In this table, DM1 means the distance metric defined by Eq. (2) without considering the relationship between attributes while DM2 stands for the complete distance metric given by Eq. (13). It can be found from the table that the DM2 metric has obtained the best result on every tested data set and the average improvement is over 26% in comparison with the Hamming distance metric. Both without considering the attribute interdependency, the average performance of DM1 metric is still over 10% better than the Hamming distance. This result indicates that quantifying distance between categorical

TABLE III. AVERAGE INTRA/INTER-CLUSTER 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	0.1515	0.6445	0.6047	0.6723
C_2	0.6632	0.2379	0.8237	0.7616	C_2	0.6445	0.1090	0.8966	0.8314
C_3	0.6011	0.8237	0.2463	0.4985	C_3	0.6047	0.8966	0.1351	0.4758
C_4	0.6421	0.7616	0.4985	0.2968	C_4	0.6723	0.8314	0.4758	0.2517

TABLE IV. AVERAGE INTRA/INTER-CLUSTER 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	0.16	0.72	0.52	0.66	0.59	0.72	0.77
C_2	0.60	0.14	0.42	0.55	0.46	0.46	0.52	C_2	0.72	0.11	0.41	0.52	0.48	0.42	0.47
C_3	0.44	0.42	0.21	0.33	0.27	0.51	0.42	C_3	0.52	0.41	0.21	0.31	0.28	0.51	0.43
C_4	0.59	0.55	0.33	0.08	0.34	0.70	0.45	C_4	0.66	0.52	0.31	0.06	0.30	0.67	0.44
C_5	0.47	0.46	0.27	0.34	0.08	0.48	0.12	C_5	0.59	0.48	0.28	0.30	0.06	0.51	0.37
C_6	0.65	0.46	0.51	0.69	0.48	0.12	0.35	C_6	0.72	0.42	0.51	0.67	0.51	0.12	0.31
C_7	0.68	0.52	0.42	0.45	0.37	0.35	0.21	C_7	0.77	0.47	0.43	0.44	0.37	0.31	0.16

values with frequency probability rather than constant is more reasonable for the analysis of relationship between categorical objects. Moreover, comparing the performance of DM1 and DM2 we can find that the information of interdependency between attributes is important for distance measurement. Making a good use of this information can significantly improve the effectiveness of learning method on categorical data. Additionally, it can be observed that the DM1 and DM2 metrics have very similar results on the WBCD data set. This is because that the dependent degree between attributes in this data set is very low and there is only one pair of attributes whose value of the interdependence redundancy measure has exceeded the threshold β.

TABLE V. CLUSTER-DISCRIMINATION INDEX OBTAINED BY THE DIFFERENT METRICS ON FOUR REAL DATA SETS

Data sets	Hamming Distance	DM1	DM2
Voting	0.5517	0.5232	**0.4778**
WBCD	0.6840	0.4457	**0.4403**
Soybean	0.3856	0.3478	**0.2402**
Zoo	0.3045	0.3012	**0.2678**

B. Study of the Threshold Parameter

To investigate the impact of the threshold parameter β on the effectiveness of proposed distance metric, we have utilized the DM2 metric with the different values of β to calculate the intra-cluster and inter-cluster distances for Soybean and Zoo data sets. The curves which depict the changing trend of obtained CDI values with increasing β have been shown in Fig. 1 and Fig. 2. From the figures, we can find that, when β is set at a very small value (i.e. $\beta < 0.1$), the performance of DM2 metric improves as the value of β increases. This is because, when the threshold β is too small, many useless relationships between attributes are taken into account, which will degrade the accuracy of obtained object distances. By contrast, when β is larger than 0.1, the performance of DM2 metric degrades as β increases. Overall, the effectiveness of DM2 metric can keep at a satisfactory level with $\beta \leq 0.4$ and when the value of β exceeds 0.4, the performance of this distance metric degrades obviously. Therefore, a value from $[0.1, 0.4]$ for β can get a good balance between computational load and practical effectiveness for the proposed distance metric.

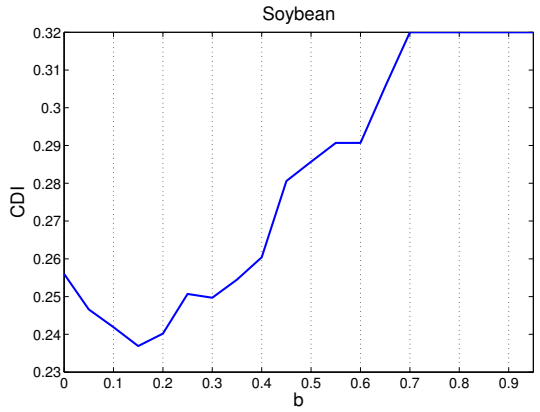

Fig. 1. Cluster-discrimination index obtained by the proposed metric with the different values of β on the Soybean data set.

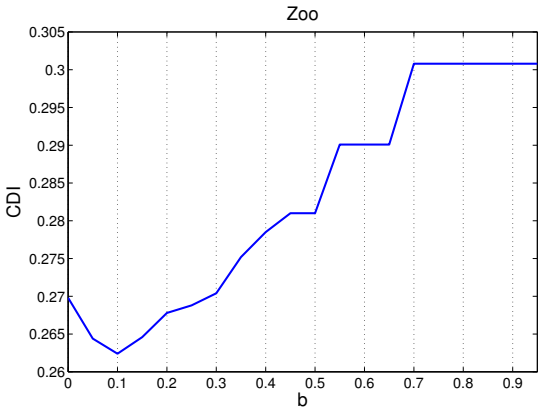

Fig. 2. Cluster-discrimination index obtained by the proposed metric with the different values of β on the Zoo data set.

TABLE VI. CLUSTERING ERRORS OBTAINED BY K-MODES ALGORITHM WITH THE DIFFERENT DISTANCE METRICS

Data sets	k-modes with Hamming Distance	k-modes with DM1	k-modes with DM2
Voting	0.1391±0.0070	0.1307±0.0047	**0.1216**±0.0021
WBCD	0.1612±0.1574	0.1008±0.1136	**0.0809**±0.0904
Soybean	0.1631±0.1719	0.1589±0.1617	**0.1107**±0.1256
Zoo	0.2884±0.0953	0.2518±0.1018	**0.2347**±0.0915

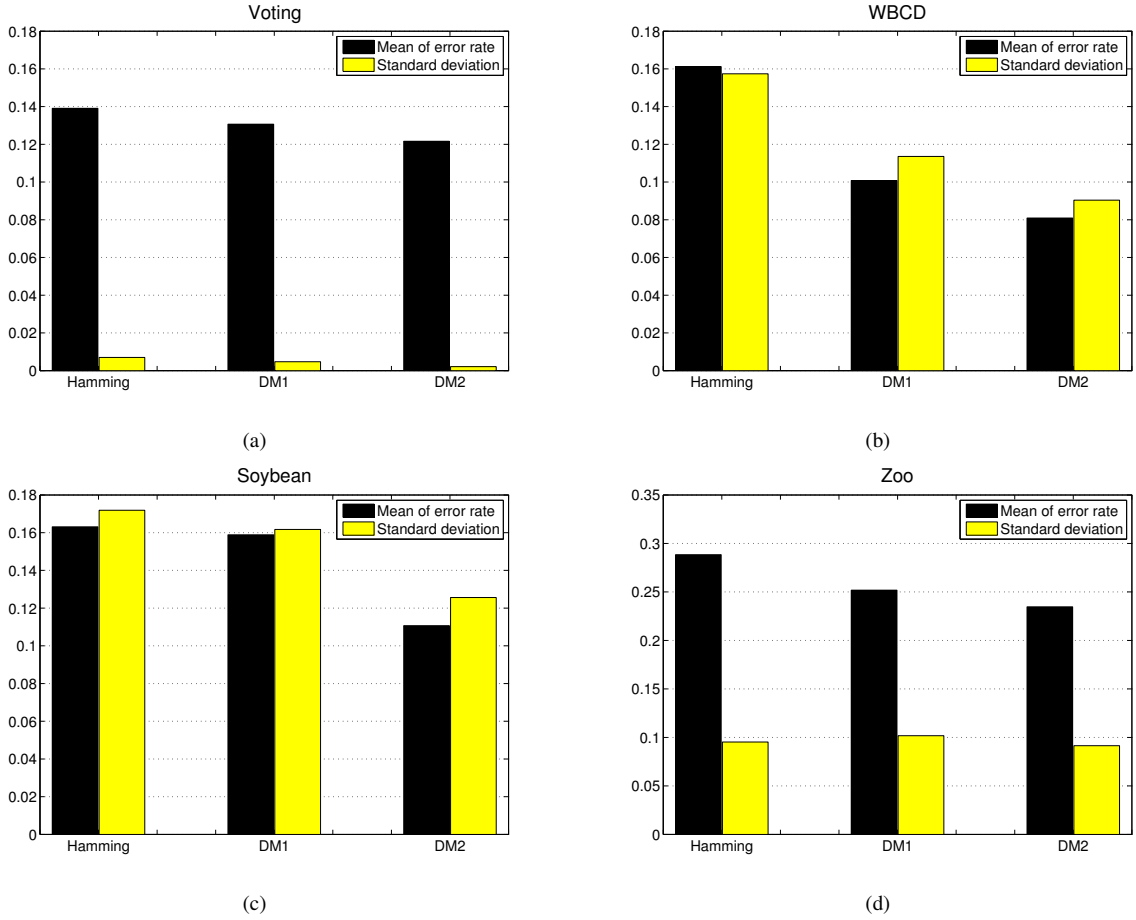

Fig. 3. Graphical representation of clustering error rate and standard deviation for different methods on (a) Voting data set, (b) WBCD data set, (c) Soybean data set, and (d) Zoo data set.

C. Clustering Analysis

Generally, 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, 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 k-modes algorithm [25], 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., k-modes algorithm with Hamming distance metric). According to [26], the clustering accuracy has been estimated by

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

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

Clustering analysis was conducted on the four categorical data sets: Voting, WBCD, Soybean, and 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 VI. Moreover, the graphical representation of the clustering results for the three methods is shown in Fig. 3. It can be seen that, for distance based clustering on categorical data, k-modes algorithm with the proposed distance metric has competitive advantage in terms of clustering accuracy compared to the other two methods. The average improvement in clustering accuracy on these four data sets obtained by DM2 metric is over 27% in comparison with the Hamming distance. It means that the proposed distance metric is more appropriate for unsupervised data analysis as

it can better reveal the true relationship between categorical objects.

IV. Conclusion

In this paper, we have presented a new distance metric, which measures the distance between categorical data with the frequency probability of each attribute value in the whole data set. Moreover, the interdependence redundancy measure was utilized to evaluate the dependent 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 which have high interdependency with the calculated one. Different experiments on benchmark data sets have shown the effectiveness of the proposed metric.

Moreover, the basic assumption of the proposed metric in this paper has paid more attention to the major clusters in a given data set, i.e., the clusters with a noticeable number of objects. Therefore, under some special situations, such as the existence of noise or cluster with few objects, clustering algorithm based on this distance metric may not so applicable as it tends to merge these minority objects together with the other clusters. Our future work will further investigate this problem elsewhere.

Acknowledgment

This work was supported by the Faculty Research Grant of Hong Kong Baptist University (HKBU) under Project FRG2/12-13/082, the National Science Foundation of China under Grant 61272366, and the Strategic Development Fund of HKBU: 03-17-033. Yiu-ming Cheung is the corresponding author.

References

[1] F. Esposito, D. Malerba, V. Tamma, and H. H. Bock, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, 2002, ch. Classical Resemblance Measures, pp. 139–152.

[2] S. Q. Le and T. B. Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognition Letters*, vol. 26, pp. 2549–2557, 2005.

[3] Z. Hubálek, "Coefficients of association and similarity, based on binary (presencecabsence) data: an evaluation," *Biological Reviews*, vol. 57, no. 4, pp. 669–689, 1982.

[4] Y. M. Cheung and H. Jia, "A unified metric for categorical and numerical attributes in data clustering," in *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013, pp. 135–146.

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

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

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

[8] C. C. Hsu and S. H. Wang, "An integrated framework for visualized and exploratory pattern discovery in mixed data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 2, pp. 161–173, February 2005.

[9] C. C. Hsu, "Generalizing self-organizing map for categorical data," *IEEE Transactions on Neural Networks*, vol. 17, no. 2, pp. 294–304, March 2006.

[10] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: a comparative evaluation," in *Proceedings of the eighth SIAM International Conference on Data Mining*, 2008, pp. 243–254.

[11] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Machine Learning*, 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 Recognition Letters*, vol. 28, pp. 110–118, 2007.

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

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

[15] K. Gowda and E. Diday, "Unsupervised learning thought symbolic clustering," *Pattern Recognition Letters*, vol. 12, pp. 259–264, 1991.

[16] K. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 2, pp. 368–378, 1992.

[17] M. Ichino and H. Yaguchi, "Generalized minkowski metrics for mixed feature-type data analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 4, pp. 698–708, 1994.

[18] F. de Carvalho, "Proximity coefficients between boolean symbolic objects," in *New Approaches in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organisation*, E. Diday, Ed., vol. 5. Springer-Verlag, 1994, pp. 387–394.

[19] F. de Carvalho, "Extension based proximities between constrained boolean symbolic objects," in *Classification and Related Methods*, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, and Y. Baba, Eds. Springer-Verlag, 1998, pp. 370–378.

[20] H. Lancaster, "The combining of probabilities arising from data in discrete distributions." *Biometrika*, vol. 36, pp. 370–382, 1949.

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

[22] V. Ganti, J. Gehrke, and R. Ramakrishan, "Cactuslclustering categorical data using summaries," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 73–83.

[23] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[24] W.-H. Au, K. C. C. Chan, and A. K. C. W. Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 83–101, 2005.

[25] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Know ledge Discovery*, 1997, pp. 1–8.

[26] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems*, 2005.