A New Distance Metric Exploiting Heterogeneous Interattribute Relationship for Ordinal-and-Nominal-Attribute Data Clustering

Yiqun Zhang, *Member, IEEE*, and Yiu-Ming Cheung, *Fellow, IEEE*

***Abstract*—Ordinal attribute has all the common characteristics of a nominal one but it differs from the nominal one by having naturally ordered possible values (also called categories interchangeably). In clustering analysis tasks, categorical data composed of both ordinal and nominal attributes (also called mixed-categorical data interchangeably) are common. Under this circumstance, existing distance and similarity measures suffer from at least one of the following two drawbacks: 1) directly treat ordinal attributes as nominal ones, and thus ignore the order information from them and 2) suppose all the attributes are independent of each other, measure the distance between two categories from a target attribute without considering the valuable information provided by the other attributes that correlate with the target one. These two drawbacks may twist the natural distances of attributes and further lead to unsatisfactory clustering results. This article, therefore, presents an entropy-based distance metric that quantifies the distance between categories by exploiting the information provided by different attributes that correlate with the target one. It also preserves the order relationship among ordinal categories during the distance measurement. Since attributes are usually correlated in different degrees, we also define the interdependence between different types of attributes to weight their contributions in forming distances. The proposed metric overcomes the two above-mentioned drawbacks for mixed-categorical data clustering. More important, it conceptually unifies the distances of ordinal and nominal attributes to avoid information loss during clustering. Moreover, it is parameter free, and will not bring extra computational cost compared to the existing state-of-the-art counterparts. Extensive experiments show the superiority of the proposed distance metric.**

***Index Terms*—Clustering analysis, interdependence, ordinal-and-nominal-attribute data, unified distance metric (UDM).**

I. Introduction

CATEGORICAL data are common in clustering analysis tasks, and attributes composing categorical data can be divided into two classes, that is: 1) nominal attribute and 2) ordinal attribute [1], [2], as shown in Fig. 1. Ordinal attribute has all of the characteristics of the nominal one, but it differs from the nominal one by having naturally ordered possible values (also called categories interchangeably) [3], [4]. In clustering analysis tasks, it is common that a categorical data consists of both nominal and ordinal attributes [5]. Table I demonstrates a fragment of such a categorical dataset extracted from a teaching assistant (TA) evaluation dataset, where the two ordinal attributes, "Attribute 1" and "Attribute 2," record the helpfulness and professional level of each TA, respectively; the two nominal attributes, "Attribute 3" and "Attribute 4," record the corresponding course and course type, respectively; and the "Label" attribute records whether the TAs have been awarded. In this mixed-categorical dataset, it is obvious that the values of Label are relevant to the ordering information of the ordinal categories. For example, the values that are closer to "Agree" in Attribute 1 and Attribute 2 tend to indicate the label "Yes." If we treat them as nominal ones, such information will be ignored and may lead to unsatisfactory clustering results. Therefore, ordinal and nominal attributes should be treated differently in the clustering analysis of such data.

Unfortunately, most existing categorical data similarity/distance measures are proposed provided that the categorical dataset consists of nominal attributes only [6]. Among these measures, the simplest and most popular one is the Hamming distance [7], which directly assigns distance "0" to identical categories and distance "1" to any pair of unequal categories. Goodall's similarity measure (GSM) [8] directly assigns similarity 0 to any pair of unequal categories, and attempts to measure similarity for identical categories by exploiting their statistical information (i.e., occurrence frequencies). Later, association-based [9], Ahmad's [10], and context-based [11], [12] distance metrics have been proposed. They adopt a similar basic idea to exploit interattribute relationship information for more reasonable distance measurement. Nevertheless, since they rely on the information offered by the related attributes, they are incompetent for the datasets that comprise independent attributes. To solve this problem, Jia's distance metric (JDM) [13] is proposed, which

Manuscript received July 1, 2019; revised November 11, 2019 and January 31, 2020; accepted March 16, 2020. Date of publication April 27, 2020; date of current version February 16, 2022. This work was supported in part by NSFC under Grant 61672444, in part by the HKBU Initiation Grant—Faculty Niche Research Areas 2018/19 under Grant RC-FNRA-IG/18-19/SCI/03, in part by the ITF of Innovation and Technology Commission of the Government of the Hong Kong under Project ITS/339/18, and in part by SZSTI under Grant JCYJ20160531194006833. This article was recommended by Associate Editor A. F. Skarmeta Gomez. *(Corresponding author: Yiu-Ming Cheung.)*

Yiqun Zhang is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: yqzhang@comp.hkbu.edu.hk).

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University (HKBU), Hong Kong, and also with the HKBU Institute of Research and Continuing Education, Shenzhen 518057, China (e-mail: ymc@comp.hkbu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.2983073.

Digital Object Identifier 10.1109/TCYB.2020.2983073

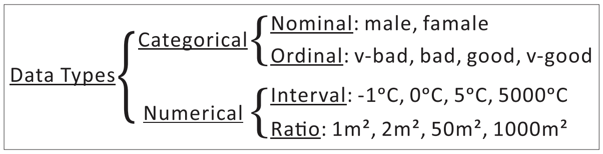
Data Types
Categorical
Nominal: male, famale
Ordinal: v-bad, bad, good, v-good
Numerical
Interval: -1°C, 0°C, 5°C, 5000°C
Ratio: 1m², 2m², 50m², 1000m²

Fig. 1. Relationship among different attribute types. “v-bad” and “v-good” indicate very-bad and very-good, respectively.

TABLE I
FRAGMENT OF THE TA EVALUATION DATASET

No.	Attribute 1 (Helpfulness)	Attribute 2 (Professional)	Attribute 3 (Course)	Attribute4 (Type)	Label (Award)
1	Agree	Agree	Culture	Lecture	Yes
2	Agree	Marginal	Finance	Lecture	Yes
3	Marginal	Marginal	Culture	Lecture	No
4	Disagree	Disagree	Oral	Lab	No

further exploits the occurrence frequency of the target categories. However, all the above-mentioned measures/metrics ignore the order relationship among categories when processing ordinal attributes, and are thus inappropriate for the clustering of datasets with ordinal attributes.

In the literature, Lin’s similarity measure (LSM) [14] and ordinal distance metric [15] take into account the order relationship from the perspective of information theory. However, they have not addressed the distance measurement problem in the nominal case, and are thus inappropriate for the datasets with nominal attributes. A more straightforward solution is numerical coding (NC), which codes ordinal categories into consecutive integers and nominal categories into binary attributes, and then treat the coded data as a numerical one in clustering analysis. Although the order relationship is preserved in this way, the statistical information of categories is ignored and there is an awkward gap between ordinal and nominal attributes, which may cause misinterpretation of the distances.

Most recently, an entropy-based distance metric (EBDM) [16] proposes unifying the distance concept for ordinal and nominal attributes to avoid information loss caused by the awkward gap. However, it only exploits the occurrence frequency of the target categories, and does not consider the valuable information provided by the other attributes. Many existing works (see [9]–[13]) have shown that information extracted from the correlated attributes is very useful for defining distances in categorical data clustering. Taking the data fragment shown in Table I as an example, suppose ordinal Attribute 1 and nominal Attribute 4 are interdependent, the distance between “Agree” and “Marginal” of Attribute 1 should be very short from the perspective of Attribute 4, because the corresponding values of these two categories are all the same in Attribute 4. Evidently, distances defined without sufficiently exploiting such information will be somewhat unreasonable, and will therefore affect clustering performance. Therefore, properly exploiting the information offered by heterogeneous ordinal and nominal attributes and reasonably quantifying the interattribute dependence are both key factors for correctly defining distances in mixed-categorical data clustering.

Ideally, a comprehensive distance metric should take into account the following five aspects: 1) intraattribute statistical information; 2) interattribute correlation information; 3) attribute weighting; 4) order relationship among ordinal categories; and 5) unification of the distance definitions. By taking into account the former two aspects, a measure like JDM [13] will be robust to the interdependence degrees among attributes. The third aspect is important because the information offered by different attributes will have different contributions in forming distances. The fourth aspect ensures that a metric does not roughly treat ordinal attributes as nominal ones, while the last aspect guarantees that the distances of ordinal and nominal attributes are defined in a unified manner to eliminate the information loss caused by awkwardly combining different types of distances. Aspects that are taken into account by different metrics are summarized in Table II, where “Intra.,” “Inter.,” “Weight,” “Order,” and “Unif.” indicate the above-mentioned five aspects, respectively. Obviously, all existing metrics have at least two defects, and may thus yield unsatisfactory results in mixed-categorical data clustering [16]–[19]. Therefore, it is necessary to propose a distance metric that comprehensively takes into account all five aspects for mixed-categorical data clustering.

TABLE II
ASPECTS CONSIDERED BY DIFFERENT MEASURES/METRICS

Measure/Metric	Intra.	Inter.	Weight	Order	Unif.
Hamming [7]					
Goodall’s [8]	✓				
Association-based [9]		✓			
Ahmad’s [10]		✓			
Context-based [11], [12]		✓	✓		
Jia’s [13]	✓	✓	✓		
Numerical Coding				✓	
Lin’s [14]	✓			✓	
Ordinal [15]	✓			✓	
Entropy-based [16]	✓			✓	✓
Proposed	✓	✓	✓	✓	✓

In this article, we propose such a distance metric. The main difficulty lies in how to reasonably define a distance by quantifying the information extracted from the different types of interattribute relationships (i.e., those between ordinal attributes, those between nominal attributes, and those between ordinal and nominal attributes) caused by the natural differences between ordinal and nominal attributes. The basic idea is to conceptually unify the distance indicated in different situations from the perspective of information theory [20], [21]. More specifically, we use information amount (i.e., entropy) of categories quantified according to different attributes to indicate their distance, because dissimilar categories usually have corresponding dissimilar values on the other attributes, which can be properly quantified in a unified way by entropy. Through our design, the distances indicated by the information extracted under the different types of interattribute relationships can be conceptually unified, and the order relationship among ordinal categories can also be preserved. In addition, since attributes are usually interdependent in different degrees, the information offered by them may have different contributions in distance measurement. Hence, we also propose a measure to quantify the interdependence under the different types of interattribute relationship in a unified way provided that two attributes are interdependent if two data objects show their consistency on these two attributes (e.g., having identical values on each of the two attributes). In practice, we use the percentage of inspected object pairs that show the consistency to quantify the interdependence. Consequently, a unified mixed-categorical data distance metric is formed by weighting the distances measured on different attributes according to the interdependence. Experiments conducted in this article show that the proposed distance metric can reasonably quantify distances and outperforms the existing counterparts in clustering analysis. The main contributions of this article are four-fold.

1) A series of rules has been formed to offer guidance about how to design a mixed-categorical data distance metric that can comprehensively take into account all five aspects shown in Table II for distance measurement.
2) Distance of ordinal and nominal attributes is defined in a unified way from the perspective of information theory. This definition takes into account both intra- and inter-attribute statistical information while preserving the order relationship among ordinal categories.
3) Interdependence is quantified in a unified way for the three types of interattribute relationship: a) those between ordinal attributes; b) those between nominal attributes; and c) those between ordinal and nominal attributes.
4) A unified and parameter-free distance metric is formed by integrating the proposed distance definition and interdependence measure. In comparison with the counterparts, the proposed metric is more comprehensive and effective, and is competent for the distance measurement in the clustering analysis of mixed-categorical data.

The remainder of this article is organized as follows. Section II reviews the existing related works. Section III presents the details of the proposed distance metric. The experimental results are demonstrated and discussed in Section IV. Finally, we draw a conclusion in Section V.

II. Overview of Existing Related Works

A. Existing Distance and Similarity Measures

The traditional Hamming distance metric (HDM) [7] is commonly used in categorical data clustering analysis. It assigns binary distances 0 and 1 to each pair of identical and unequal categories, respectively. Similarly, Goodall's measure [8] assigns 0 similarity to each pair of unequal categories, but attempts to distinguish the similarity between different pairs of identical categories by exploiting the occurrence frequency of them. Association-based [9], Ahmad's [10], and context-based [11], [12] distance metrics adopt a similar basic idea that similar probability distributions of the corresponding values on the other attributes indicates a shorter distance between two categories. Since the association-based metric and Ahmad's metric treats each attribute equally, which is obviously unreasonable [22], [23], context-based metric further selects a set of more dependent attributes as the context for distance measurement. However, since they all rely on the interdependence of attributes, they may fail when all of the attributes are independent of each other. Although JDM [13] solves this problem by further considering the intraattribute statistical information, all the above-mentioned six metrics are still inappropriate for mixed-categorical data, as they are designed for nominal data only.

In the literature, LSM [14] and the ordinal distance metric [15] have been proposed to especially exploit the order information for ordinal data distance measurement. They both quantify similarities/distances from the perspective of information theory and preserve the order relationship among ordinal categories. However, they did not exploit the valuable interattribute correlation information, and cannot properly address the distance measurement problem of ordinal and nominal attributes. NC that simply assigns consecutive integers to the ordinal categories and converts each nominal category into a binary attribute is a feasible solution. Nevertheless, since it ignores the statistical information of categories, the natural intercategory distances will be twisted. It also creates more attributes due to the coding of nominal categories. These defects may surely influence the efficiency and accuracy of clustering analysis. EBDM [16] extends the metric proposed in [15] by unifying the distance concept of ordinal and nominal attributes, and is thus suitable for the distance measurement of mixed-categorical data. However, it still has not exploited the valuable information that can be extracted from the interattribute relationship, which makes it somewhat unreasonable.

B. Existing Interdependence Measures

Symmetric uncertainty [24] and interdependence redundancy [13] are two interdependence measures adopted by categorical data distance metrics proposed in [12] and [13], respectively. Both of these two interdependence measures are symmetrical, and they calculate dependence degrees between categorical attributes from the perspective of information theory. Symmetric uncertainty is based on information gain [25], and interdependence redundancy is based on mutual information [26]. Actually, the concepts of information gain and mutual information are equivalent to each other in the scenario of interattribute dependence measurement [27]. These two measures differ from each other in how they compensate for the bias of information gain and mutual information toward attributes with more values. Symmetric uncertainty divides the information gain of two attributes by their total entropy, while interdependence redundancy divides the mutual information of two attributes by their joint entropy. However, they are both inappropriate for the interdependence measurement of ordinal attributes, because they did not take into account the order relationship among the ordinal categories.

Spearman's rank correlation [28], Kendall's tau coefficient (KTC) [29], and rank mutual information [30] are three interdependence measures designed for ordinal data. The former two adopt a similar basic idea that the dependence degree between two attributes will be high when their corresponding values reflect a higher degree of agreement in terms of the order [31]–[34]. The difference is that Spearman's rank correlation measures interdependence based on the "rank difference" between two attributes, while KTC quantifies interdependence based on the "concordance" between the corresponding value pairs of two attributes. The latter one (i.e., rank mutual information) takes into account the order relationship of ordinal attributes by computing mutual information based on dominance rough set [35]–[39]. However, these three measures do not apply to nominal attributes because they all measure interdependence based on the order of categories.

III. Proposed Metric

We first formulate the problem of mixed-categorical data distance measurement, discuss the challenging issues, and make an overview of the proposed metric in Section III-A. Then, we give the technical details and discussions in the remaining part of this section.

A. Preliminaries

Given a categorical dataset with N data objects $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ represented by d attributes $A_1, A_2, \ldots, A_d$, which have $v_1, v_2, \ldots, v_d$ categories, respectively (the vectors and matrices are indicated by boldface hereinafter). It is assumed that the former d^{ord} and the latter d^{nom} attributes are ordinal and nominal, respectively, where $d = d^{\text{ord}} + d^{\text{nom}}$. A category of an attribute is denoted in the form of $o_{r,t}$, where $r \in \{1, 2, \ldots, d\}$ is the sequential number of attribute A_r that $o_{r,t}$ belongs to, and $t \in \{1, 2, \ldots, v_r\}$ is the sequential number of $o_{r,t}$. If A_r is an ordinal attribute (i.e., $r \leq d^{\text{ord}}$), its v_r categories are naturally ordered as $o_{r,1} \succ o_{r,2} \succ \cdots \succ o_{r,v_r}$, where the symbol "$\succ$" indicates that the categories on its left rank higher than the categories on its right. A data object $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$, $i \in \{1, 2, \ldots, N\}$, is represented by d categories, each of which belongs to an attribute. For two objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, their distance $\text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ consists of d subdistances measured between their 1st, 2nd, …, dth values. Measure the distance between their rth values is equivalent to measure the distance between the two corresponding categories of A_r. Suppose A_r and A_s are interdependent, the distance between $o_{r,t}$ and $o_{r,h}$ measured according to A_s is denoted as $\varphi_{A_s}(o_{r,t}, o_{r,h}) \cdot R(A_r, A_s)$, where $\varphi_{A_s}(o_{r,t}, o_{r,h})$ is the distance between $o_{r,t}$ and $o_{r,h}$ indicated by A_s, and $R(A_r, A_s)$ is the interdependence degree of A_r and A_s that controls the contribution of A_s in forming the overall distance between $o_{r,t}$ and $o_{r,h}$, which is denoted as $\Phi(o_{r,t}, o_{r,h})$.

TABLE III
Example of Mixed-Categorical Data Set. A_1 Is an Ordinal Attribute With $o_{1,1}$ = "A," $o_{1,2}$ = "B," $o_{1,3}$ = "C," and $o_{1,1} \succ o_{1,2} \succ o_{1,3}$. A_2 Is an Ordinal Attribute With $o_{2,1}$ = "↑," $o_{2,2}$ = "∼," $o_{2,3}$ = "↓," and $o_{2,1} \succ o_{2,2} \succ o_{2,3}$. A_3 Is a Nominal Attribute with $o_{3,1}$ = "■," $o_{3,2}$ = "★," and $o_{3,3}$ = "▲"

Object	A_1 (Ordinal)	A_2 (Ordinal)	A_3 (Nominal)
$\boldsymbol{x}_1$	A	↑	■
$\boldsymbol{x}_2$	A	∼	★
$\boldsymbol{x}_3$	B	∼	▲
$\boldsymbol{x}_4$	C	↓	★
$\boldsymbol{x}_5$	C	↓	■

It is a challenging task to reasonably define a distance of mixed-categorical data because the relationship among categories of ordinal and nominal attributes exists in different ways, which yields different types of intercategory distance. Let us take the dataset shown in Table III as an example. Intuitively, the dissimilarity of "↑" and "↓" is lower than that of "↑" and "∼" indicated by "A_3," because "↑" and "↓" have common corresponding value (i.e., "■"), but "↑" and "∼" do not. Also, the dissimilarity of "↑" and "↓" is higher than that of "↑" and "∼" indicated by "A_1" because the corresponding values of "↑" and "↓" (i.e., $\{A\}$ of "↑" and $\{C, C\}$ of "↓") are with the larger order difference than that of "↑" and "∼" (i.e., $\{A\}$ of "↑" and $\{A, B\}$ of "∼"). It is obvious that the ordinal "A_1" and the nominal "A_3" indicate the distance in different ways. Moreover, by considering the ordinal nature of "A_2," the dissimilarity of "↑" and "↓" could not be lower than that of "↑" and "∼," because the former two have larger order difference. Apparently, there exists awkward gap among the intercategory distance $\varphi_{A_s}(o_{r,t}, o_{r,h})$ indicated in the following four situations: 1) both A_r and A_s are ordinal; 2) A_r is ordinal and A_s is nominal; 3) A_r is nominal and A_s is ordinal; and 4) both A_r and A_s are nominal. Combining these types of distances that are not defined in a unified way to form $\Phi(o_{r,t}, o_{r,h})$ will surely cause information loss in clustering analysis. In this article, we circumvent this issue by adopting an information theory-based idea that two categories containing more different information are usually more dissimilar to each other. By quantifying the information of categories indicated by different attributes using entropy, the distances indicated in the above-mentioned four situations are unified into homogeneous concepts and can thus be directly combined for clustering analysis without causing information loss. We will present the details of the unified distance definition in Section III-B.

In practice, attributes are usually interdependent to a certain degree. It turns out that the information from the different attributes is of different importance in forming a distance. Therefore, how to reasonably measure the interdependence $R(A_r, A_s)$ in the following three situations: 1) both A_r and A_s are ordinal; 2) both A_r and A_s are nominal; and 3) A_r and A_s are of different types, which are caused by the differences between ordinal and nominal attributes, is also crucial to the success of mixed-categorical data clustering. Although some interdependence measures [13], [24], [28]–[30] have been defined, they are only applicable to one of the first two situations. How to measure the interdependence in the third situation and how to eliminate the concept gap among the three types of interdependence are both nonstraightforward tasks that have yet to be studied in the literature. Thus, we study the interdependence in all three situations, and propose a measure that quantifies the interdependence in a unified way. The proposed measure uses the number of object pairs that indicate the dependence between two attributes in a nonconflicting way to indicate the interdependence degree. The details of the proposed unified interdependence measure will be presented in Section III-C.

B. Entropy-Based Distance Definition

Distance $\varphi_{A_s}(o_{r,t}, o_{r,h})$ can be defined according to rule 1.

Rule 1: Given two interdependent attributes A_r and A_s, if $o_{r,t}$ and $o_{r,h}$ of A_r have no common corresponding values on A_s, distance $\varphi_{A_s}(o_{r,t}, o_{r,h})$ can be quantified by summing up the entropy values of the two joint probability distributions of $o_{r,t}$'s and $o_{r,h}$'s corresponding values on A_s.

Remark 1: Given joint probability distributions $P(o_{r,t}, A_s) = \{p(o_{r,t}, o_{s,1}), p(o_{r,t}, o_{s,2}), \ldots, p(o_{r,t}, o_{s,v_s})\}$ and $P(o_{r,h}, A_s) = \{p(o_{r,h}, o_{s,1}), p(o_{r,h}, o_{s,2}), \ldots, p(o_{r,h}, o_{s,v_s})\}$, we note that the entropy $E(o_{r,t}, A_s) = -\sum_{u=1}^{v_s} p(o_{r,t}, o_{s,u}) \log_2 p(o_{r,t}, o_{s,u})$ of $P(o_{r,t}, A_s)$ and the entropy $E(o_{r,h}, A_s) = -\sum_{u=1}^{v_s} p(o_{r,h}, o_{s,u}) \log_2 p(o_{r,h}, o_{s,u})$ of $P(o_{r,h}, A_s)$ quantify the information amount of $o_{r,t}$ and $o_{r,h}$, respectively, according to A_s. Here, $p(o_{r,t}, o_{s,u})$ is a joint probability defined as $p(o_{r,t}, o_{s,u}) = [(\sigma_{o_{r,t} \wedge o_{s,u}}(\boldsymbol{X}))/N]$, where $\sigma_{o_{r,t} \wedge o_{s,u}}(\boldsymbol{X})$ is a function counting the number of objects in $\boldsymbol{X}$ with their rth and sth values equal to $o_{r,t}$ and $o_{s,u}$, respectively. It is intuitive that larger amount of different information contained by two different categories indicates that they are more dissimilar. Therefore, if $o_{r,t}$ and $o_{r,h}$ have no common corresponding values on A_s, they contain totally different information indicated by A_s, and the distance $\varphi_{A_s}(o_{r,t}, o_{r,h})$ can be quantified by

$$\varphi_{A_s}\big(o_{r,t}, o_{r,h}\big) = \begin{cases} \sum_{g=\{t,h\}} E\big(o_{r,g}, A_s\big), & \text{if } t \neq h \\ 0, & \text{if } t = h \end{cases}. \tag{1}$$

The distance in the case $t = h$ is 0 because the distance between a category to itself is always zero.

The distance described by (1) is defined under the hypothesis that $o_{r,t}$ and $o_{r,h}$ do not have common information from the perspective of A_s. However, in practice, the common part of their information, if any, is counted twice by (1), which is unreasonable. Thus, rule 2 is yielded.

Rule 2: $o_{r,t}$ and $o_{r,h}$ should be treated as a whole to avoid the double counting of their common information.

Remark 2: Given two joint probability distributions $P(o_{r,t}, A_s)$ and $P(o_{r,h}, A_s)$, if $o_{r,t}$ and $o_{r,h}$ are teated as a whole, a new joint probability distribution $P(o_{r,th}, A_s) = \{p(o_{r,th}, o_{s,1}), p(o_{r,th}, o_{s,2}), \ldots, p(o_{r,th}, o_{s,v_s})\}$ is yielded, where $p(o_{r,th}, o_{s,1}) = p(o_{r,t}, o_{s,1}) + p(o_{r,h}, o_{s,1})$, $p(o_{r,th}, o_{s,2}) = p(o_{r,t}, o_{s,2}) + p(o_{r,h}, o_{s,2}), \ldots, p(o_{r,th}, o_{s,v_s}) = p(o_{r,t}, o_{s,v_s}) + p(o_{r,h}, o_{s,v_s})\}$. Based on $P(o_{r,th}, A_s)$, the distance is redefined as

$$\varphi_{A_s}\big(o_{r,t}, o_{r,h}\big) = \begin{cases} E\big(o_{r,\mathrm{th}}, A_s\big), & \text{if } t \neq h \\ 0, & \text{if } t = h \end{cases} \tag{2}$$

where $E(o_{r,\mathrm{th}}, A_s) = -\sum_{u=1}^{v_s} p(o_{r,\mathrm{th}}, o_{s,u}) \log_2 p(o_{r,\mathrm{th}}, o_{s,u})$. If there is no common information contained by $o_{r,t}$ and $o_{r,h}$, $E(o_{r,\mathrm{th}}, A_s) = \sum_{g=\{t,h\}} E(o_{r,g}, A_s)$ is consistent with rule 1; if there exists common information, $E(o_{r,th}, A_s) < \sum_{g=\{t,h\}} E(o_{r,g}, A_s)$ is consistent with rule 2.

According to the definition of entropy [25], a larger v_s may result in a larger $\varphi_{A_s}(o_{r,t}, o_{r,h})$. But this effect does not correctly reveal the true contribution of A_s. This is called v_s-effect hereinafter, and rule 3 is yielded accordingly.

Rule 3: v_s-effect should be eliminated when computing $\varphi_{A_s}(o_{r,t}, o_{r,h})$.

Remark 3: Standard information $S_{A_s} = -\log_2(1/v_s)$ that calculates the maximum entropy of an attribute (i.e., entropy of an attribute when the occurrence frequency of its categories are identical) is presented in [16] for eliminating the v_s-effect. Hence, we adopt it to redefine the distance as

$$\varphi_{A_s}\big(o_{r,t}, o_{r,h}\big) = \begin{cases} \frac{E(o_{r,\mathrm{th}}, A_s)}{S_{A_s}}, & \text{if } t \neq h \\ 0, & \text{if } t = h \end{cases}. \tag{3}$$

If A_r is an ordinal attribute, the order relationship among its categories should not be ignored. Then, we have rule 4.

Rule 4: If A_r is ordinal, any pairs of distances of A_r measured according to A_s should satisfy: if $o_{r,t} \succeq o_{r,p} \succeq o_{r,u} \succeq o_{r,h}$ or $o_{r,t} \preceq o_{r,p} \preceq o_{r,u} \preceq o_{r,h}$, then $\varphi_{A_s}(o_{r,t}, o_{r,h}) \geq \varphi_{A_s}(o_{r,p}, o_{r,u})$, where $t, p, u, h \in \{1, 2, \ldots, v_r\}$. Here, the symbol "$\succeq$" ("$\preceq$") indicates that the categories on its left rank not lower (higher) than the categories on its right.

Remark 4: rule 4 is to ensure that the distance between two categories (e.g., $o_{r,t}$ and $o_{r,h}$) is not smaller than the distance between another two categories that are ordered between them (e.g., $o_{r,p}$ and $o_{r,u}$). In other words, it guarantees that the distances defined for ordinal categories do not violate their order relationship. Thus, the distance is redefined as

$$\varphi_{A_s}\big(o_{r,t}, o_{r,h}\big) = \begin{cases} \frac{\sum_{g=\min(t,h)}^{\max(t,h)-1} E(o_{r,gw}, A_s)}{S_{A_s}}, & \text{if } t \neq h,\ r \leq d^{\mathrm{ord}} \\ \frac{E(o_{r,th}, A_s)}{S_{A_s}}, & \text{if } t \neq h,\ r > d^{\mathrm{ord}} \\ 0, & \text{if } t = h \end{cases} \tag{4}$$

where $w = g + 1$. In this way, order relationship among ordinal categories is preserved and the measured distances are consistent with rule 4.

Consequently, we have intercategory distance

$$\Phi\big(o_{r,t}, o_{r,h}\big) = \frac{1}{d} \sum_{s=1}^{d} \varphi_{A_s}\big(o_{r,t}, o_{r,h}\big) \tag{5}$$

and the distance between two data objects can be written as

$$\mathrm{Dist}\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \sqrt{\sum_{r=1}^{d} \Phi\big(x_{ir}, x_{jr}\big)^2} \tag{6}$$

where x_{ir} and x_{jr} are the r^{th} values of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively.

Distance defined by (5) assumes that each attribute contributes equally in forming $\Phi(o_{r,t}, o_{r,h})$. However, in real categorical datasets, attributes are usually interdependent in different degrees, and thus have different contributions. In Section III-C, we further discuss how to quantify the interdependence for the attributes of mixed-categorical data.

C. Concordant-Based Interdependence Measure

If the interdependence degree $R(A_r, A_s)$ between A_r and A_s is low, $\varphi_{A_s}(o_{r,t}, o_{r,h})$ will be "unreliable," which means that the distances measured according to the information offered by A_s should contribute less in forming $\Phi(o_{r,t}, o_{r,h})$. This is because that a low $R(A_r, A_s)$ indicates that A_s does not contain much information about A_r. Hence, the contributions of different attributes should be weighted according to the interdependence. For mixed-categorical data, interattribute relationship exists in the following three cases.

1) *Case 1:* Both A_r and A_s are nominal.
2) *Case 2:* A_r and A_s are of the different attribute types.
3) *Case 3:* Both A_r and A_s are ordinal.

Subsequently, we have rule 5 as follows.

Rule 5: In cases 1 and 2, A_r and A_s should be treated as nominal ones for computing $R(A_r, A_s)$. In case 3, A_r and A_s should be treated as ordinal ones for computing $R(A_r, A_s)$.

Remark 5: case 1 is very common, and has been widely studied by the existing works, (see [9]–[13]). In case 2, an ordinal attribute can be viewed as a special case of the nominal attribute that the categories are with order constraints. Since the nominal attribute does not have such order information, the interdependence in case 2 can just exist in a nominal level. So A_r and A_s should be treated as nominal ones to measure their interdependence degree. In case 3, since A_r and A_s are both ordinal, the interdependence surely exists in an ordinal level. To sum up, there are two types of interdependence.

1) *Type-1:* Ordinal-level interdependence in case 3.
2) *Type-2:* Nominal-level interdependence in cases 1 and 2.

The difficulty lies in how to make the definitions of these two types of interdependence unified. Then, we have rule 6 as follows.

Rule 6: The two types of interdependence can be conceptually unified by counting the number of data object pairs that indicate the interdependence between two attributes in a nonconflicting way.

Remark 6: The existing nominal-level [13], [24] and ordinal-level [28]–[30] interdependence measures quantify the interdependence from different perspectives. Combining them for attributes weighting will surely cause information loss. The heterogeneous interdependence is caused by the two different attribute types. So we unify the interdependence as the number of inspected data object pairs that indicate the interdependence between two attributes in a nonconflicting way. To find out such "nonconflicting" object pairs, a series of concepts about "concordant" is defined as follows. The concept concordant is derived from the definition of KTC [29]. Since KTC is only feasible for the type-1 interdependence, we refine the concept of concordant, and define its variations to unify the two types of interdependence.

Definition 1: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are equal-concordant on A_r and A_s, if $x_{ir} = x_{jr}$ and $x_{is} = x_{js}$.

Remark 7: Definition 1 applies to both the two types of interdependence, because $x_{ir} = x_{jr}$ and $x_{is} = x_{js}$ indicate that A_r and A_s have consistent behavior indicated by $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. KTC ignores the situation of equal-concordant, and subtracts the number of "concordance" by the number of "discordance," which may result in negative interdependence that is unsuitable for the attributes weighting in this article.

Then, we define three concordant-related concepts that only apply to the type-1 interdependence.

Definition 2: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are positive-concordant on A_r and A_s, if $x_{ir} \prec x_{jr}$ and $x_{is} \prec x_{js}$, or $x_{ir} \succ x_{jr}$ and $x_{is} \succ x_{js}$.

Definition 3: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are negative-concordant on A_r and A_s, if $x_{ir} \prec x_{jr}$ and $x_{is} \succ x_{js}$, or $x_{ir} \succ x_{jr}$ and $x_{is} \prec x_{js}$.

Definition 4: For A_r and A_s, all the object pairs that are positive-concordant, and all the object pairs that are negative-concordant, are concordant-conflict to each other.

Remark 8: Concordant-conflict is defined because positive-concordant and negative-concordant are opposite to each other in indicating the consistency of two attributes.

We use the percentage of object pairs that are nonconcordant-conflict to quantify the type-1 interdependence

$$R^{\text{ord}}(A_r, A_s) = \frac{C^{=}_{r,s} + \left(\max\left(C^{\text{ord+}}_{r,s}, C^{\text{ord-}}_{r,s}\right) - \min\left(C^{\text{ord+}}_{r,s}, C^{\text{ord-}}_{r,s}\right)\right)}{N(N-1)/2} \tag{7}$$

where $C^{=}_{r,s}$, $C^{\text{ord}+}_{r,s}$, and $C^{\text{ord}-}_{r,s}$ are the total number of object pairs that are equal-concordant, positive-concordant, and negative-concordant, respectively.

Then, we define four concordant-related concepts that only apply to type-2 interdependence.

Definition 5: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are unequal-concordant on A_r and A_s, if $x_{ir} \neq x_{jr}$ and $x_{is} \neq x_{js}$.

Remark 9: Since there is no order relationship among nominal categories, unequal-concordant is also a kind of consistency of A_r and A_s indicated by $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, because by changing the perspective from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$, it can be found that the object values on A_r and A_s have changed.

Definition 6: Two object sets $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$ are unequal-concordant on A_r and A_s, if $o_{r,t} \neq o_{r,h}$ and $o_{s,g} \neq o_{s,u}$.

Remark 10: In Definition 6, the notation $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ represents an object set containing all the objects in $\boldsymbol{X}$ with their rth values equal to $o_{r,t}$ and s^{th} values equal to $o_{s,g}$. In this definition, $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$ are unequal-concordant because any pairs of objects between them are unequal-concordant according to Definition 5. For $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$, there are $|\boldsymbol{X}_{o_{r,t},o_{s,g}}| \times |\boldsymbol{X}_{o_{r,h},o_{s,u}}|$ unequal-concordant object pairs in total, where $|\boldsymbol{X}_{o_{r,t},o_{s,g}}|$ and $|\boldsymbol{X}_{o_{r,h},o_{s,u}}|$ are the numbers of objects in $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$, respectively.

Definition 7: A pair of unequal-concordant object sets $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$ and another pair of unequal-concordant object sets $\boldsymbol{X}_{o_{r,t},o_{s,u}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,g}}$ are concordant-conflict to each other, if $o_{s,g} \neq o_{s,u}$.

Remark 11: In Definition 7, the sth values of the objects are completely reversed in the two pairs of object sets. Hence, they are conflicting in indicating consistency of A_r and A_s.

Definition 8: If a pair of sets $\boldsymbol{X}_{o_{r,t},o_{s,g}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,u}}$, and another pair of sets $\boldsymbol{X}_{o_{r,t},o_{s,u}}$ and $\boldsymbol{X}_{o_{r,h},o_{s,g}}$, are concordant-conflict to each other, the pair with larger number of object pairs are judged to be positive-concordant, and the remaining pair are judged to be negative-concordant.

Remark 12: The role of Definitions 7 and 8 for type-2 interdependence measurement is equivalent to the role of Definitions 2–4 for type-1 interdependence measurement.

Similar to (7), we have the type-2 interdependence

$$R^{\text{nom}}(A_r, A_s) = \frac{C^{=}_{r,s} + \sum_{t=1}^{v_r-1} \sum_{h=t+1}^{v_r} \left(\boldsymbol{C}^{\text{nom+}}_{r,s}(t,h) - \boldsymbol{C}^{\text{nom-}}_{r,s}(t,h)\right)}{N(N-1)/2} \tag{8}$$

where $\boldsymbol{C}^{\text{nom}+}_{r,s}$ and $\boldsymbol{C}^{\text{nom}-}_{r,s}$ are two $v_r \times v_r$ upper triangular matrices containing the numbers of object pairs that are positive-concordant and negative-concordant, respectively. $\boldsymbol{C}^{\text{nom}+}_{r,s}(t,h) = \sum_{g=1}^{v_s-1} \sum_{u=g+1}^{v_s} \max(|\boldsymbol{X}_{o_{r,t},o_{s,g}}| \times |\boldsymbol{X}_{o_{r,h},o_{s,u}}|, |\boldsymbol{X}_{o_{r,t},o_{s,u}}| \times |\boldsymbol{X}_{o_{r,h},o_{s,g}}|)$ is an element of $\boldsymbol{C}^{\text{nom+}}_{r,s}$ and $\boldsymbol{C}^{\text{nom}-}_{r,s}(t,h) = \sum_{g=1}^{v_s-1} \sum_{u=g+1}^{v_s} \min(|\boldsymbol{X}_{o_{r,t},o_{s,g}}| \times |\boldsymbol{X}_{o_{r,h},o_{s,u}}|, |\boldsymbol{X}_{o_{r,t},o_{s,u}}| \times |\boldsymbol{X}_{o_{r,h},o_{s,g}}|)$ is an element of $\boldsymbol{C}^{\text{nom-}}_{r,s}$.

The proposed interdependence measure described by (7) and (8) is summarized as follows:

$$R(A_r, A_s) = \begin{cases} R^{\text{ord}}(A_r, A_s), & \text{if } r, s \leq d^{\text{ord}} \\ R^{\text{nom}}(A_r, A_s), & \text{else.} \end{cases} \quad (9)$$

$R(A_r, A_s)$ satisfies: 1) $0 \leq R(A_r, A_s) \leq 1$; 2) $R(A_r, A_s) = R(A_s, A_r)$; and 3) $R(A_r, A_s) = 1$ if $r = s$.

Remark 13: Although (7) and (8) are a little different in the form, they have the homogeneous concept, that is, nonconcordant-conflict object pairs as a percentage of the total number of object pairs. Hence, the two types of interdependence measured by (9) are unified in concept and comparable in magnitude.

D. Unified Distance Metric

With $R(A_r, A_s)$ defined in Section III-C, contributions of different attributes are weighted in forming $\Phi(o_{r,t}, o_{r,h})$ by

$$\Phi\big(o_{r,t}, o_{r,h}\big) = \frac{1}{d} \sum_{s=1}^{d} R(A_r, A_s) \cdot \varphi_{A_s}\big(o_{r,t}, o_{r,h}\big). \quad (10)$$

A unified distance metric (UDM) is thus formed, which is described by (4), (6), (9), and (10). Here, we demonstrate the computation process of $\Phi(o_{1,1}, o_{1,3})$ for the dataset shown in Table III to further explain the details of UDM.

Example 1: Interdependence degrees are computed as follows. $R(A_1, A_1) = [(2 + (8 - 0))/(5 \times 4 \div 2)] = 1$ because all the inspected data object pairs are not concordant-conflict; $R(A_1, A_2) = R^{\text{ord}}(A_1, A_2) = [(1 + (7 - 0))/(5 \times 4 \div 2)] = 0.8$, where $C^{=}_{1,2} = 1$ is obtained by the inspecting object pair $\{\boldsymbol{x}_4, \boldsymbol{x}_5\}$, $C^{\text{ord}+}_{1,2} = 7$ is obtained by inspecting $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_4\}$, $\{\boldsymbol{x}_1, \boldsymbol{x}_5\}$, $\{\boldsymbol{x}_2, \boldsymbol{x}_4\}$, $\{\boldsymbol{x}_2, \boldsymbol{x}_5\}$, $\{\boldsymbol{x}_3, \boldsymbol{x}_4\}$, and $\{\boldsymbol{x}_3, \boldsymbol{x}_5\}$, and $C^{\text{ord}-}_{1,2} = 0$; $R(A_1, A_3) = R^{\text{nom}}(A_1, A_3) = [(0 + (2 - 0) + (1 - 1) + (2 - 0))/(5 \times 4 \div 2)] = 0.4$, where $C^{=}_{1,3} = 0$, $\boldsymbol{C}^{\text{nom}-}_{1,3}(1, 2) = 0$, and $\boldsymbol{C}^{\text{nom}-}_{1,3}(2, 3) = 0$. $\boldsymbol{C}^{\text{nom}+}_{1,3}(1, 2) = 2$ is obtained by inspecting $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\boldsymbol{C}^{\text{nom}+}_{1,3}(1, 3) - \boldsymbol{C}^{\text{nom}-}_{1,3}(1, 3) = 1 - 1$ is obtained by inspecting $\{\boldsymbol{x}_1, \boldsymbol{x}_4\}$ and $\{\boldsymbol{x}_2, \boldsymbol{x}_5\}$, and $\boldsymbol{C}^{\text{nom}+}_{1,3}(2, 3) = 2$ is obtained by inspecting $\{\boldsymbol{x}_3, \boldsymbol{x}_4\}$ and $\{\boldsymbol{x}_3, \boldsymbol{x}_5\}$; with $R(A_1, A_1)$, $R(A_1, A_2)$, and $R(A_1, A_3)$, $\Phi(o_{1,1}, o_{1,3})$ is computed as follows. $\varphi_{A_1}(o_{1,1}, o_{1,3}) = [(-[2/5]\log_2[2/5] - [1/5]\log_2[1/5]) + (-[1/5]\log_2[1/5] - [2/5]\log_2[2/5])]/(-\log_2[1/3]) = 1.25$, where $E(o_{1,12}, A_1) = -(2/5)\log_2(2/5) - (1/5)\log_2(1/5)$, $E(o_{1,23}, A_1) = -(1/5)\log_2(1/5) - (2/5)\log_2(2/5)$, and $S_{A_1} = -\log_2(1/3)$. In the same way, we have $\varphi_{A_2}(o_{1,1}, o_{1,3}) = 1.25$, and $\varphi_{A_3}(o_{1,1}, o_{1,3}) = 1.76$. Then we have $\Phi(o_{1,1}, o_{1,3}) = (1/3) \times (1 \times 1.25 + 0.8 \times 1.25 + 0.4 \times 1.76) = 0.98$.

As a distance metric, UDM satisfies the following four conditions for all $o_{r,t}, o_{r,h}, o_{r,p}$ with $r \in \{1, 2, \ldots, d\}$ and $t, h, p \in \{1, 2, \ldots, v_r\}$:

1) $0 \leq \Phi(o_{r,t}, o_{r,h})$;
2) $\Phi(o_{r,t}, o_{r,h}) = 0 \Leftrightarrow o_{r,t} = o_{r,h}$;
3) $\Phi(o_{r,t}, o_{r,h}) = \Phi(o_{r,h}, o_{r,t})$;
4) $\Phi(o_{r,t}, o_{r,h}) \leq \Phi(o_{r,t}, o_{r,p}) + \Phi(o_{r,p}, o_{r,h})$

and also satisfies the following four conditions for all $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_l \in \boldsymbol{X}$:

1) $0 \leq \text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_j)$;
2) $\text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \Leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_j$;
3) $\text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \text{Dist}(\boldsymbol{x}_j, \boldsymbol{x}_i)$;
4) $\text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \text{Dist}(\boldsymbol{x}_i, \boldsymbol{x}_l) + \text{Dist}(\boldsymbol{x}_l, \boldsymbol{x}_j)$.

Algorithm 1 Distance Measurement Using UDM

```
Input: Data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$
Output: $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $i, j \in \{1, 2, \ldots, N\}$
for $r = 1$ to $d$ do
    for $s = 1$ to $d$ do
        Calculate $R(A_r, A_s)$ according to Eq. (9);
    end for
end for
for $r = 1$ to $d$ do
    for $s = 1$ to $d$ do
        Calculate the distance between $x_{ir}$ and $x_{jr}$
        according to $A_s$ by using Eq. (4);
    end for
    Calculate the overall distance between $x_{ir}$ and $x_{jr}$
    by using Eq. (10);
end for
Calculate $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$ by using Eq. (6).
```

The algorithm of distance measurement using UDM is summarized as Algorithm 1. To save computation cost, d $v_r \times v_r$ distance matrices $\boldsymbol{M}_r$, $r \in \{1, 2, \ldots, d\}$, that contain intercategory distances of the d attributes calculated by (10), is maintained during distance measurement. With these distance matrices, the distance between any two of the data objects can be directly read off.

Time complexity for the distance measurement using UDM is $O(d^2N)$, which is the same as the distance metrics proposed in [9]–[13]. The time complexity consists of three parts.

1) $O(d^2N)$ for calculating interdependence degrees.
2) $O(d^2N)$ for generating distance matrices.
3) $O(d)$ for reading off distance between two data objects.

We prove them in the following.

Theorem 1: Time complexity for calculating the interdependence degree between each pair of d attributes is $O(d^2N)$.

Proof: Before calculating $R(A_r, A_s)$, the rth and sth values of the N data objects are scanned once to form a $v_r \times v_s$ matrix $\boldsymbol{Q}_{r,s}$. An element $\boldsymbol{Q}_{r,s}(t, g)$ records the occurrence frequency of data objects in $\boldsymbol{X}$ with their rth and sth values equal to $o_{r,t}$ and $o_{s,g}$, respectively. Time complexity for obtaining $\boldsymbol{Q}_{r,s}$ is $O(N)$. Then, we select two different rows from $\boldsymbol{Q}_{r,s}$ [each row is a $1 \times v_s$ vector, and there are $v_r(v_r - 1)/2$ pairs of different vectors], and multiply each pair of the intervector values to form a $v_s \times v_s$ matrix. In this way, $v_r(v_r - 1)/2$ matrices $\boldsymbol{P}_{1,2}, \boldsymbol{P}_{1,3}, \ldots, \boldsymbol{P}_{v_r-1,v_r}$ are produced, where $\boldsymbol{P}_{t,h}(g, u) = \boldsymbol{Q}_{r,s}(t, g) \times \boldsymbol{Q}_{r,s}(h, u)$. Time complexity for producing a matrix $\boldsymbol{P}_{t,h}$ is $O(v_s^2)$, and for producing $v_r(v_r - 1)/2$ matrices is $O(v_r^2 v_s^2)$. With $\boldsymbol{Q}_{r,s}$ and $\boldsymbol{P}_{1,2}, \boldsymbol{P}_{1,3}, \ldots, \boldsymbol{P}_{v_r-1,v_r}$, $R(A_r, A_s)$ is measured as follows. First, $C^{=}_{r,s} = \sum_{t=1}^{v_r} \sum_{g=1}^{v_s} \boldsymbol{Q}_{r,s}(t, g)(\boldsymbol{Q}_{r,s}(t, g) - 1)/2$ is calculated with time complexity $O(v_r v_s)$. For type-1 interdependence, $C^{\text{ord}+}_{r,s} = \sum_{t=1}^{v_r-1} \sum_{h=t+1}^{v_r} \sum_{g=1}^{v_s-1} \sum_{u=g+1}^{v_s} \boldsymbol{P}_{t,h}(g, u)$ and $C^{\text{ord}-}_{r,s} = \sum_{h=1}^{v_r-1} \sum_{t=h+1}^{v_r} \sum_{u=1}^{v_s-1} \sum_{g=u+1}^{v_s} \boldsymbol{P}_{t,h}(g, u)$ are computed to obtain $R(A_r, A_s)$ according to (7), which has time complexity $O(v_r^2 v_s^2)$. For type-2 interdependence,

$\boldsymbol{C}^{\text{nom}+}_{r,s}(t,h) = \sum_{g=1}^{v_s-1}\sum_{u=g+1}^{v_s} \max(\boldsymbol{P}_{t,h}(g,u), \boldsymbol{P}_{t,h}(u,g))$ and $\boldsymbol{C}^{\text{nom}-}_{r,s}(t,h) = \sum_{g=1}^{v_s-1}\sum_{u=g+1}^{v_s} \min(\boldsymbol{P}_{t,h}(g,u), \boldsymbol{P}_{t,h}(u,g))$ are computed with time complexity $O(v_s^2)$. Then we calculate $R(A_r, A_s)$ according to (8). In this process, $\boldsymbol{C}^{\text{nom}+}_{r,s}(t,h)$ and $\boldsymbol{C}^{\text{nom}-}_{r,s}(t,h)$ with different t and h ($t, h \in \{1, 2, \ldots, v_r\}$) should be generated $v_r(v_r-1)/2$ times. Therefore, time complexity for calculating the type-2 $R(A_r, A_s)$ is still $O(v_r^2 v_s^2)$. Consequently, overall time complexity for obtaining $R(A_r, A_s)$ is $O(N+v_r^2 v_s^2)$. For $d(d-1)/2$ pairs of attributes in total, the time complexity is $O(d^2N + d^2V^4)$, where $V = \max(v_1, v_2, \ldots, v_d)$. Since V is a small constant from the practical viewpoint, the overall time complexity for calculating interdependence degrees between each pair of the d attributes is $O(d^2N)$. ■

Theorem 2: Time complexity for calculating the d distance matrices is $O(d^2N)$.

Proof: Before calculating $\varphi_{A_s}(o_{r,t}, o_{r,h})$, $\boldsymbol{Q}_{r,s}$ is produced in the same way as the proof of Theorem 1, and the time complexity is $O(N)$. If A_r is a nominal attribute, the tth and hth rows of $\boldsymbol{Q}_{r,s}$ should be added up and divided by N to form a $1 \times v_s$ vector, which is the joint probability distribution of the corresponding values of $o_{r,t}$ and $o_{r,h}$. Then, we calculate $\varphi_{A_s}(o_{r,t}, o_{r,h})$ by (4) with time complexity $O(v_s)$. Since there are $v_r(v_r-1)/2$ distances between the categories of A_r, time complexity for obtaining these distances according to A_s is $O(N + v_r^2 v_s)$. If A_r is an ordinal attribute, the worst case time complexity for calculating $\varphi_{A_s}(o_{r,t}, o_{r,h})$ using (4) is $O(v_r v_s)$. Therefore, time complexity for obtaining the $v_r(v_r-1)/2$ distances of A_r according to A_s is $O(N + v_r^3 v_s)$. By adopting $V = \max(v_1, v_2, \ldots, v_d)$, time complexity for obtaining the distances of A_r is $O(N + V^4)$. Since all the d attributes and corresponding interdependence degrees should be considered for calculating the distances of A_r, time complexity for obtaining the distance matrix $\boldsymbol{M}_r$ of A_r is $O(dN + dV^4)$, and for obtaining the d distance matrices is $O(d^2N + d^2V^4)$. Since V is a small constant, the overall time complexity is $O(d^2N)$. ■

Theorem 3: The time complexity for reading off the distance between any two of the data objects is $O(d)$.

Proof: For a pair of data objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the d distances between their corresponding values are directly read off from the d distance matrices $\boldsymbol{M}_r$, $r \in \{1, 2, \ldots, d\}$. Therefore, the time complexity for reading off a distance is $O(d)$. ■

IV. Experiments

To evaluate the performance of UDM, we compare it with the existing counterparts on 16 real and benchmark datasets. Four validity indices are utilized to evaluate the performance, and five experiments are designed to prove the effectiveness of UDM from different aspects.

A. Experimental Settings

Six distance/similarity measures, including HDM [7], GSM [8], LSM [14], context-based distance metric (CBDM) [12], JDM [13], and EBDM [16], are selected as counterparts of the proposed UDM. Among them, HDM, GSM, and LSM are conventional categorical data metrics. CBDM, which is the improved version of association-based [9] and Ahmad's [10] metrics, is a representative metric that exploits interattribute relationship for categorical data clustering. JDM and EBDM are both state-of-the-art categorical data metrics. Since the intercategory distances measured by EBDM and the proposed UDM can be utilized for coding ordinal attributes into numerical ones, EBDM coding (EBDMC) and UDM coding (UDMC) are also treated as two counterparts. For completeness, the simple NC is also treated as a counterpart. Since the distances of nominal attributes measured by EBDM and UDM cannot be utilized for coding nominal attribute, both of them adopt the binary coding strategy of NC for nominal category coding.

All the selected distance/similarity measures are embedded into a clustering algorithm, and their clustering performance on different datasets are compared in Section IV-B. To better evaluate the performance of different distance/similarity measures, we embed NC, EBDMC, and UDMC into the simplest **k**-MeanS (KMS) [40], and embed the remainders into the simplest **k**-MoDes (KMD) [41].

Six representative categorical data clustering algorithms, that is, the conventional KMD [41], entropy-based categorical data clustering (ECC) [42], the representative attribute weighting **k**-modes (WKM) [43], mixed attribute WKM (MWKM) [44], and the state-of-the-art subspace clustering of categories (SCC) [45] and attribute weighting object-cluster similarity-based clustering (WOC) [46], are chosen for comparison in Section IV-C. Since the original distance/similarity measures of KMD, WKM, and WOC can be easily replaced by UDM without influencing their optimization process, we also embed UDM into them to form another three counterparts. According to the suggestion in [44], the parameter of WKM and MWKM (i.e., β) is set at 2, and the two parameters of MWKM (i.e., T_v and T_s) are both set at 1. The parameter θ of SCC is determined according to [45].

To verify the effectiveness of rules 1–6 presented in Section III, we compare UDM with its five versions (denoted as DM1–5, respectively) in Section IV-D. The former four are formed according to (1)–(4), and are compared with UDM for the justification of rules 1–4. DM5 is the version of UDM that only adopts (8) for attributes weighting. Compare UDM with DM5 can justify rules 5 and 6.

We collect 16 real and benchmark datasets, including six mixed-categorical datasets, five ordinal datasets, and five nominal datasets. Primary Tumor (abbreviated as Primary), Hayes Roth (abbreviated as Hayes), Lymphography (abbreviated as Lym), Mammographic Mass (abbreviated as Mass), and Nursery are benchmark datasets collected from the UCI Machine Learning Repository[1] [5]. Fruit is a real dataset collected from a business survey of an advertising company. Employee Rejection/Acceptance (abbreviated as Employee), Lecturer Evaluation (abbreviated as Lecturer), and Social Works (abbreviated as Social) are benchmark datasets collected from the Weka website[2] [47]. Photo Evaluation (abbreviated as Photo) and Internship Questionnaire (abbreviated as Internship) are two real ordinal datasets collected from the student questionnaires of the College of International Exchange

[1] http://archive.ics.uci.edu/ml/datasets.html

[2] https://www.cs.waikato.ac.nz/ml/weka/datasets.html

TABLE IV
STATISTICS OF THE 16 DATA SETS

Data type	Data set	# Ins.	# Att.(O)	# Att.(N)	# Class
Categorical	Primary	123	1	15	12
	Hayes	132	2	2	3
	Lym	148	3	15	4
	Mass	824	2	2	2
	Nursery	12,960	7	1	4
	Fruit	100	3	2	5
Ordinal	Employee	1,000	4	0	9
	Lecturer	1,000	4	0	5
	Social	1,000	10	0	4
	Photo	66	4	0	3
	Internship	90	3	0	2
Nominal	Lenses	24	0	4	3
	Soybean	47	0	21	4
	Zoo	101	0	16	7
	Voting	435	0	16	2
	Tictac	958	0	9	2

of Shenzhen University and the Education University of Hong Kong, respectively. All the five nominal datasets, that is, Lenses, Soybean, Zoo, Voting Records (abbreviated as Voting), and Tictac, are benchmark datasets collected from the UCI Machine Learning Repository[1] [5]. All the instances with missing values are omitted. Both Primary and Mass datasets have one numerical attribute, which has been omitted. Since we focus on mixed-categorical data clustering, to ensure a meaningful evaluation, we have made sure that the orders among the categories of each ordinal attributes in the collected datasets are related to the label classes. Statistics of the datasets are shown in Table IV. "# Ins.," "# Att.(O)," "# Att.(N)," and "# Class" indicate the number of instances, ordinal attributes, nominal attributes, and classes, respectively. In the experiments, the sought number of clusters k is set at the "true" number of label classes of each dataset.

Three validity indices, that is, clustering accuracy (CA)[3] [48], adjusted rand index (ARI) [49]–[51], and normalized mutual information (NMI) [13], [52] are adopted for the performance evaluation. CA is a popular and conventional index, which has values in the interval [0, 1]. ARI is a popular and powerful index, which has values in the interval $[-1, 1]$. NMI evaluates the clustering performance from the perspective of information theory and the values of NMI is in the interval [0, 1]. For all the three above-mentioned indices, a larger value indicates a better clustering performance.

The other validity index, that is, label-order consistency (LOC), is defined in this article to evaluate if a metric correctly preserves the natural-order information for ordinal data distance measurement. LOC is defined as

$$\mathrm{LOC} = \frac{\sum_{i,j,I,J}\sum_{h=1}^{|\boldsymbol{c}_I|\cdot|\boldsymbol{c}_J|}\xi(\boldsymbol{H}_{i,j}, \boldsymbol{H}_{I,J}(h))}{\sum_{I,J}|\boldsymbol{c}_I|\cdot|\boldsymbol{c}_J|} \tag{11}$$

where $\boldsymbol{c}_i$, $\boldsymbol{c}_j$, $\boldsymbol{c}_I$, and $\boldsymbol{c}_J$ are four object sets containing the objects with ith, jth, Ith, and Jth benchmark class labels in $\boldsymbol{X}$, respectively, and the ordinal class labels satisfy $I \leq i < j < J$ or $I < i < j \leq J$. $|\boldsymbol{c}_i|$, $|\boldsymbol{c}_j|$, $|\boldsymbol{c}_I|$, and $|\boldsymbol{c}_J|$ are the number of objects of $\boldsymbol{c}_i$, $\boldsymbol{c}_j$, $\boldsymbol{c}_I$, and $\boldsymbol{c}_J$, respectively. $\boldsymbol{H}_{i,j}$ is a vector containing the $|\boldsymbol{c}_i| \times |\boldsymbol{c}_j|$ inter-object distances/similarities between $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$. $\boldsymbol{H}_{I,J}$ is a vector containing the $|\boldsymbol{c}_I| \times |\boldsymbol{c}_J|$ inter-object distances/similarities between $\boldsymbol{c}_I$ and $\boldsymbol{c}_J$, and $\boldsymbol{H}_{I,J}(h)$ is the hth value of $\boldsymbol{H}_{I,J}$. For a distance measure, $\xi(\boldsymbol{H}_{i,j}, \boldsymbol{H}_{I,J}(h)) = 1$ if mean $(\boldsymbol{H}_{i,j}) + \mathrm{std}(\boldsymbol{H}_{i,j}) < \boldsymbol{H}_{I,J}(h)$, otherwise, $\xi(\boldsymbol{H}_{i,j}, \boldsymbol{H}_{I,J}(h)) = 0$, while for a similarity measure, the opposite is true. The LOC values are in the interval [0, 1], and a larger value indicates better performance. In general, LOC measures the percentage of the inspected distance pairs that are consistent with the order of labels. Therefore, we choose the three monotonic datasets (i.e., Employee, Lecturer, and Social datasets with both ordinal attributes and ordinal labels [30], [53], [54]) for the LOC evaluation in Section IV-E.

To verify the time complexity analysis in Section III-D, we also report the execution time of UDM and all the compared distance and similarity measures in Sections VI-F. The results of all the experiments involving the randomization process are obtained by averaging the results of ten runs of the experiments. All the experiments are coded by MATLAB R2019b and implemented by a PC (Intel Xeon 3.30 GHz, 16-GB RAM).

B. Evaluation of UDM Distance Metric

Clustering performance on the six mixed-categorical datasets is shown in Table V. It can be observed that UDM and UDMC outperform all the other counterparts on all the six mixed-categorical datasets. Three more detailed observations are discussed as follows.

1) Superiority of UDM is not that obvious on the Lym dataset. In this dataset, the number of nominal attributes is obviously larger than that of ordinal attributes, which may thus weaken the superiority of UDM, because one advantage of UDM is that it exploits heterogeneous correlation information extracted from different types attributes, and another advantage of UDM is that it preserves order information of ordinal attributes for distance measurement. In contrast, UDM obviously outperforms the others on the three datasets with a similar number of ordinal and nominal attributes (i.e., Hayes, Mass, and Fruit) and the dataset with more ordinal attributes (i.e., Nursery).
2) The performance of NC, EBDMC, and UDMC is not competitive on most datasets because they adopt the same binary coding criteria for nominal attributes, which may ignore the interattribute relationship information and frequency of categories. Another reason is that the KMS algorithm adopts them to compute the "mean" for the coded categorical values, but categorical values are naturally unsuitable for the mean arithmetic operation.
3) Performance of CBDM is not reported for Nursery dataset because CBDM fails to measure distance for the datasets with independent attributes like Nursery.

Clustering performance on the five ordinal datasets and five nominal datasets is shown in Figs. 2 and 3. It can be observed that UDM and UDMC perform well and are very competitive on all the ordinal and nominal datasets. Six more detailed observations are discussed as follows.

1) UDM, UDMC, EBDM, and EBDMC outperform the other counterparts on all the ordinal datasets in general,

[3] The CA here computes the matching rate based on the "best permutation mapping" between the obtained clusters and true classes [48].

TABLE V

Clustering Performance of the Ten Distance and Similarity Measures on the Six Mixed-Categorical Data Sets. The Best and Second-Best Results Are Highlighted Using Boldface and Underline, Respectively

Index	Data Set	KMD+HDM	KMD+GSM	KMD+LSM	KMD+CBDM	KMD+JDM	KMS+NC	KMS+EBDMC	KMD+EBDM	**KMS+UDMC**	**KMD+UDM**
CA	Primary	0.410±0.04	0.393±0.04	0.396±0.04	0.399±0.03	0.391±0.03	0.401±0.04	0.399±0.04	0.361±0.04	0.398±0.04	**0.412±0.03**
	Hayes	0.389±0.02	0.398±0.04	0.392±0.04	0.383±0.04	0.386±0.03	0.442±0.07	0.439±0.07	0.407±0.03	0.440±0.07	**0.446±0.04**
	Lym	0.459±0.05	0.451±0.04	0.459±0.05	0.489±0.05	0.476±0.07	0.563±0.07	0.556±0.08	0.450±0.03	**0.564±0.07**	0.494±0.03
	Mass	0.818±0.00	0.820±0.00	0.831±0.00	0.828±0.00	0.769±0.14	0.743±0.12	0.745±0.12	0.807±0.00	0.728±0.13	**0.837±0.00**
	Nursery	0.375±0.04	0.356±0.03	0.375±0.03	-	0.345±0.03	0.323±0.03	0.326±0.03	0.400±0.02	0.325±0.03	**0.411±0.03**
	Fruit	0.477±0.05	0.478±0.04	0.526±0.05	0.494±0.04	0.440±0.05	0.429±0.04	0.425±0.04	0.541±0.04	0.416±0.03	**0.546±0.03**
ARI	Primary	0.193±0.03	0.173±0.04	0.184±0.03	0.185±0.03	0.175±0.02	0.193±0.04	0.192±0.04	0.140±0.05	0.192±0.04	**0.196±0.04**
	Hayes	-0.003±0.01	-0.000±0.01	-0.001±0.02	0.000±0.02	-0.003±0.01	**0.032±0.04**	0.030±0.03	0.008±0.02	0.030±0.03	**0.032±0.02**
	Lym	0.105±0.03	0.106±0.04	0.114±0.04	0.135±0.04	0.095±0.04	0.183±0.07	0.192±0.09	0.151±0.04	**0.195±0.09**	0.157±0.05
	Mass	0.404±0.00	0.409±0.00	0.438±0.00	0.429±0.00	0.356±0.19	0.290±0.17	0.295±0.18	0.376±0.00	0.263±0.19	**0.455±0.00**
	Nursery	0.048±0.02	0.045±0.01	0.058±0.02	-	0.036±0.02	0.022±0.02	0.023±0.02	0.072±0.02	0.024±0.02	**0.089±0.02**
	Fruit	0.214±0.05	0.188±0.04	0.282±0.04	0.257±0.04	0.184±0.07	0.151±0.06	0.136±0.06	0.314±0.04	0.128±0.05	**0.321±0.04**
NMI	Primary	0.442±0.03	0.443±0.04	0.435±0.03	0.438±0.04	0.425±0.03	**0.455±0.04**	0.453±0.04	0.413±0.04	**0.455±0.04**	0.447±0.03
	Hayes	0.016±0.01	0.016±0.02	0.014±0.02	0.014±0.02	0.013±0.01	0.042±0.03	0.040±0.03	0.024±0.02	0.039±0.03	**0.061±0.04**
	Lym	0.171±0.04	0.178±0.06	0.172±0.04	0.185±0.04	0.169±0.05	0.226±0.06	0.231±0.07	0.210±0.04	**0.236±0.07**	0.218±0.03
	Mass	0.325±0.01	0.323±0.00	0.345±0.00	0.341±0.00	0.282±0.14	0.241±0.14	0.245±0.14	0.305±0.00	0.219±0.15	**0.359±0.00**
	Nursery	0.051±0.02	0.047±0.02	0.065±0.02	-	0.044±0.02	0.039±0.03	0.041±0.03	0.083±0.02	0.041±0.03	**0.099±0.03**
	Fruit	0.369±0.06	0.352±0.05	0.444±0.04	0.430±0.03	0.322±0.08	0.287±0.09	0.268±0.08	0.479±0.04	0.257±0.08	**0.510±0.03**

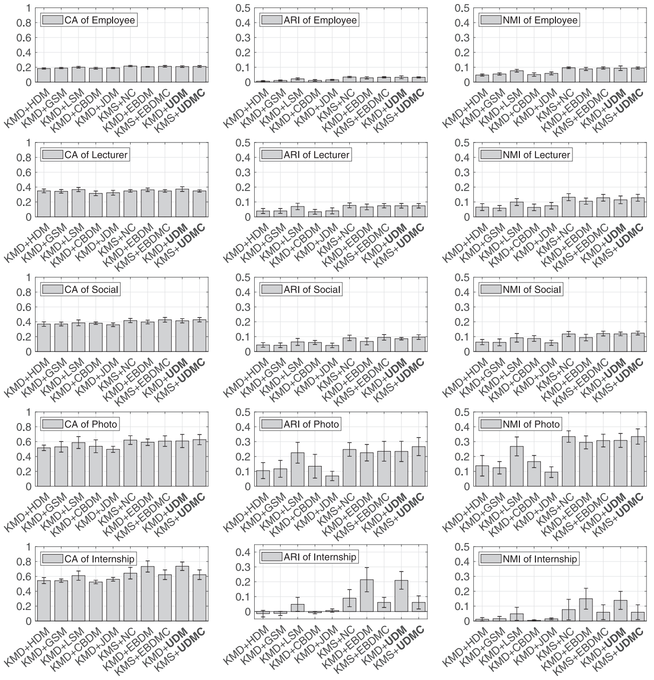

Fig. 2. Clustering performance of the ten distance and similarity measures on the five ordinal datasets.

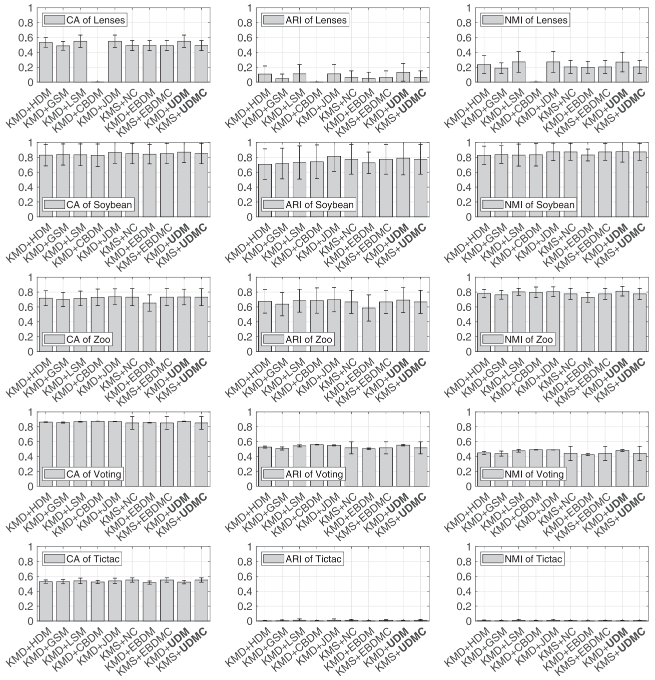

Fig. 3. Clustering performance of the ten distance and similarity measures on the five nominal datasets.

because they preserve order relationship among ordinal categories for more reasonable distance measurement.

2) UDM and UDMC perform better than EBDM and EBDMC in general, because UDM and UDMC take into account the intra- and inter-attribute statistical information for distance measurement while EBDM and EBDMC only consider the intraattribute statistical information.
3) Superiority of UDM on the five nominal datasets is not as significant as that on the five ordinal datasets, because UDM is designed for mixed-categorical data, and the unified distance and interdependence measures of it will not have significant impacts on pure nominal datasets.
4) UDMC and EBDMC have identical performance as NC on all the nominal datasets because they adopt the same binary coding strategy for nominal attributes.
5) JDM is very competitive on nominal datasets, because it is a state-of-the-art distance metric that simultaneously exploits the intra- and inter-attribute information for more accurate distance measurement.
6) CBDM fails to measure distances for Lenses dataset because the Lenses dataset comprises independent attributes.

In general, the results of this experiment indicate the superiority of UDM in comparison with the existing distance and similarity measures for categorical data clustering.

C. Comparison With Categorical Data Clustering Algorithms

The clustering performance of the six clustering algorithms and the three UDM-based algorithms on the six mixed-categorical datasets is demonstrated in Table VI. Clustering

TABLE VI
CLUSTERING PERFORMANCE OF THE NINE CLUSTERING ALGORITHMS ON THE SIX MIXED-CATEGORICAL DATA SETS. THE SYMBOL "+" INDICATES THAT THE PERFORMANCE OF KMD, WKM, AND WOC IS BOOSTED BY ADOPTING UDM

Index	Data Set	ECC	MWKM	SCC	KMD	**KMD+UDM**	WKM	**WKM+UDM**	WOC	**WOC+UDM**
CA	Primary	0.415±0.03	0.367±0.03	0.400±0.03	0.410±0.04	0.412±0.03 +	0.418±0.03	**0.427±0.03** +	0.405±0.02	0.402±0.05
	Hayes	0.418±0.05	0.398±0.03	0.346±0.01	0.389±0.02	0.446±0.04 +	0.443±0.04	**0.483±0.07** +	0.425±0.09	0.431±0.05 +
	Lym	0.520±0.04	0.484±0.05	0.422±0.05	0.459±0.05	0.494±0.03 +	0.386±0.05	0.422±0.04 +	**0.550±0.03**	0.482±0.03
	Mass	0.765±0.09	0.759±0.10	0.823±0.00	0.818±0.00	**0.837±0.00** +	0.824±0.00	0.830±0.00 +	0.829±0.00	0.831±0.00 +
	Nursery	0.335±0.00	0.355±0.03	-	0.375±0.04	0.411±0.03 +	0.391±0.07	**0.429±0.11** +	0.294±0.01	0.351±0.05 +
	Fruit	0.503±0.05	0.444±0.05	0.542±0.04	0.477±0.05	**0.546±0.03** +	0.449±0.02	0.504±0.01 +	0.466±0.07	0.537±0.04 +
ARI	Primary	**0.207±0.04**	0.153±0.03	0.179±0.03	0.193±0.03	0.196±0.04 +	0.191±0.04	0.188±0.03	0.195±0.02	0.196±0.04 +
	Hayes	0.013±0.02	0.008±0.02	-0.013±0.00	-0.003±0.01	0.032±0.02 +	0.022±0.02	**0.057±0.03** +	0.024±0.04	0.032±0.02 +
	Lym	**0.210±0.04**	0.106±0.05	0.113±0.04	0.105±0.03	0.157±0.05 +	0.041±0.03	0.060±0.03 +	0.169±0.06	0.148±0.03
	Mass	0.313±0.16	0.303±0.15	0.416±0.00	0.404±0.00	**0.455±0.00** +	0.419±0.00	0.435±0.00 +	0.432±0.00	0.438±0.00 +
	Nursery	0.003±0.00	0.047±0.02	-	0.048±0.02	0.089±0.02 +	0.067±0.06	**0.129±0.16** +	0.002±0.00	0.041±0.05 +
	Fruit	0.229±0.05	0.179±0.07	0.270±0.06	0.214±0.05	**0.321±0.04** +	0.191±0.02	0.275±0.02 +	0.194±0.09	0.284±0.01 +
NMI	Primary	**0.484±0.04**	0.413±0.03	0.365±0.04	0.442±0.03	0.447±0.03 +	0.440±0.03	0.434±0.03	0.461±0.03	0.449±0.04
	Hayes	0.027±0.02	0.019±0.01	0.001±0.00	0.016±0.01	0.061±0.04 +	0.027±0.01	**0.062±0.02** +	0.037±0.04	0.042±0.02 +
	Lym	**0.261±0.04**	0.163±0.04	0.184±0.02	0.171±0.04	0.218±0.03 +	0.096±0.03	0.120±0.04 +	0.239±0.05	0.232±0.04
	Mass	0.261±0.12	0.243±0.12	0.334±0.00	0.325±0.01	0.359±0.00 +	0.344±0.00	**0.364±0.00** +	0.344±0.00	0.349±0.00 +
	Nursery	0.034±0.01	0.051±0.03	-	0.051±0.02	0.099±0.03 +	0.089±0.09	**0.155±0.18** +	0.005±0.00	0.070±0.07 +
	Fruit	0.401±0.05	0.322±0.09	0.446±0.05	0.369±0.06	**0.510±0.03** +	0.371±0.02	0.466±0.03 +	0.352±0.11	0.472±0.03 +

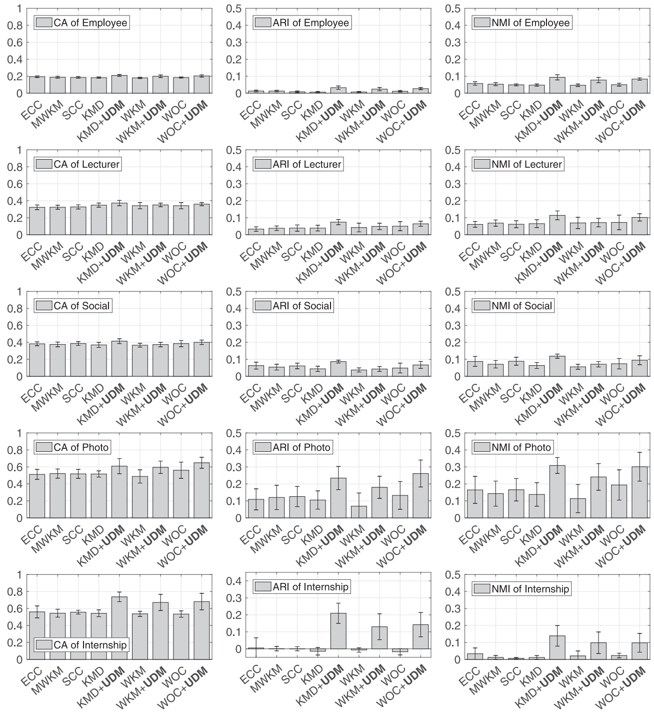

Fig. 4. Clustering performance of the nine clustering algorithms on the five ordinal datasets.

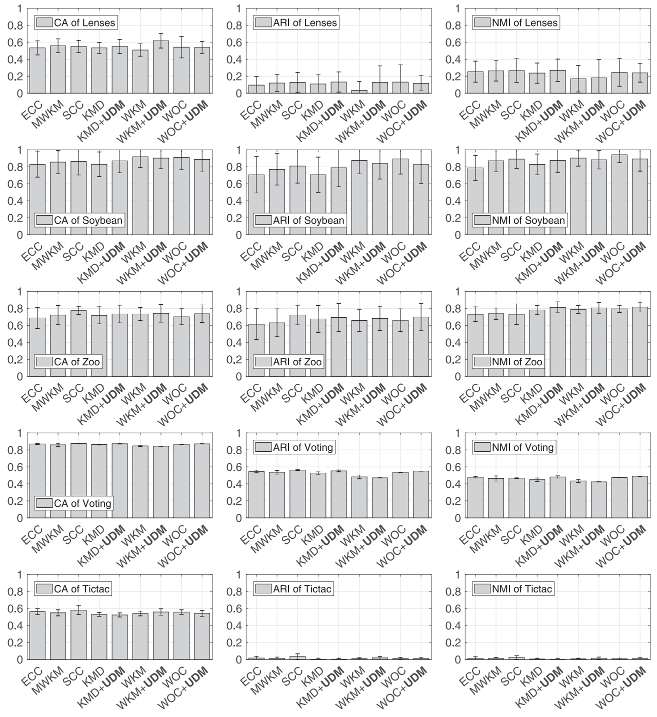

Fig. 5. Clustering performance of the nine clustering algorithms on the five nominal datasets.

performance on the five ordinal and five nominal datasets is shown in Figs. 4 and 5. It can be observed that out of a total of 48 comparisons (18, 15, and 15 for mixed-categorical, ordinal, and nominal datasets, respectively), almost all the best performing algorithms are UDM-based. Four more detailed observations are discussed as follows.

1) UDM boosts the clustering performance of KMD, WKM, and WOC on almost all the 16 datasets, which illustrates the effectiveness of UDM in the clustering analysis of any type of categorical data.
2) For the five nominal datasets, UDM does not boost the performance of KMD, WKM, and WOC a lot, because the merits of UDM (i.e., preserving order relationship and quantifying interdependence in a unified way) will not have an obvious impact on nominal data sets.
3) The performance of MWKM and SCC is competitive on nominal datasets, because they are originally designed for nominal data. Since they do not specifically exploit the order information of ordinal attributes, their clustering performance is not that good on mixed-categorical datasets and ordinal datasets.
4) Clustering performance of ECC and KMD is generally not good, because they are conventional methods that neither weight the attribute contribution nor exploit the order information of ordinal attributes.

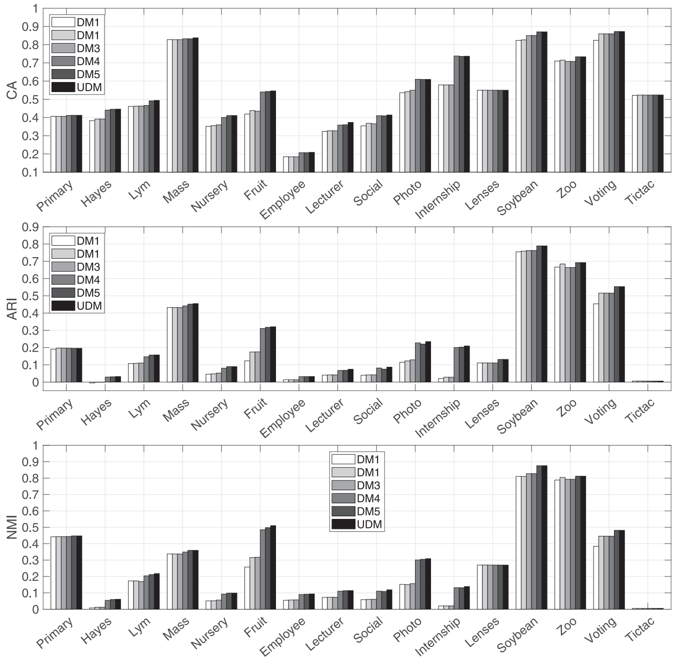

Fig. 6. Clustering performance of DM1–5 and UDM on the six mixed-categorical datasets (left six), five ordinal datasets (middle five), and five nominal datasets (right five).

In general, UDM obviously boosts the performance of existing clustering algorithms, and UDM-based algorithms perform very well in the clustering analysis of categorical data.

D. Effectiveness Verification of Rules 1–6

UDM is compared with DM1–5 in Fig. 6. According to the results, UDM performs the best on all 16 datasets in general, which proves the effectiveness of the rules presented in Section III. Furthermore, it can be observed as follows.

1) DM2 outperforms DM1 on almost all the datasets, which indicates that rule 2 can effectively avoid the "double counting" problem caused by rule 1.
2) DM3 cannot outperform DM2 on most of the 16 datasets, because rule 3 (i.e., eliminating the "v_s-effect") that guides the forming of DM3 is not intended to boost the performance of DM2. rule 3 ensures that the contribution of an attribute can be reasonably weighted by using the interdependence measure proposed in Section III-C.
3) DM4 outperforms DM1–3 on the 11 datasets with ordinal attributes, which indicates the effectiveness of rule 4. Since DM4 is equivalent to DM3 when processing nominal data, their performance is the same on the five nominal datasets.
4) DM5 evidently outperforms DM1–4 on the five nominal datasets only, because DM5 is only suitable for nominal-level interdependence measurement, which is consistent with rule 5.
5) UDM outperforms DM5 on the 11 datasets with ordinal attributes, which demonstrates the reasonableness of rules 5 and 6. Since UDM is equivalent to DM5 when processing nominal data, their performance is the same on the five nominal datasets.

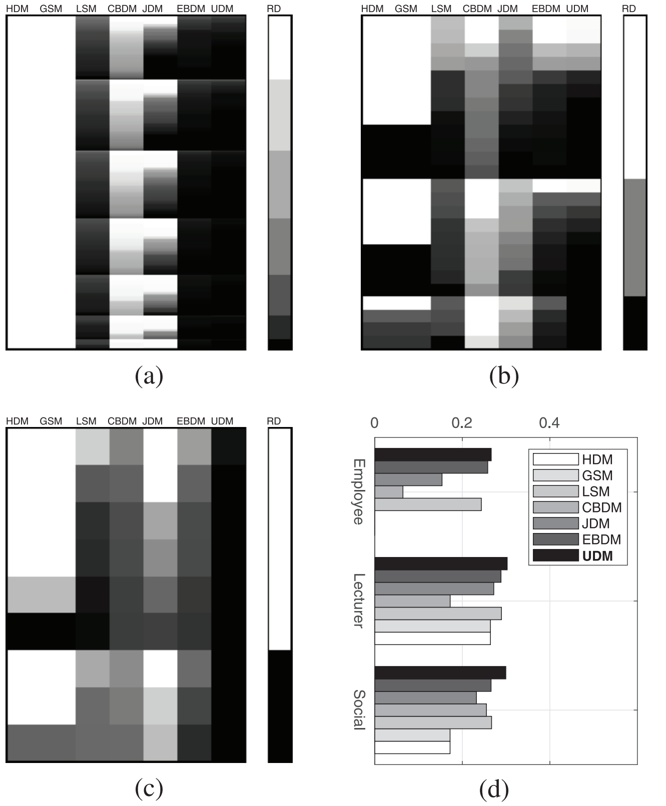

Fig. 7. LOC performance of the seven distance/similarity metrics: (a) sub-LOC on Employee; (b) sub-LOC on Lecturer; (c) sub-LOC on Social; and (d) overall LOC performance.

6) For Lenses and Tictac datasets, the performance of DM1–5 and UDM is very close to each other, because in these two datasets, most categories of the same attribute have identical corresponding values on the other attributes, which makes the distances measured by DM1–5 and UDM very similar to each other.

E. Label-Order Consistency Evaluation of UDM

LOC performance of the compared metrics on the three monotonic ordinal datasets (i.e., Employee, Lecturer, and Social) is demonstrated in Fig. 7. This experiment evaluates whether the existing distance/similarity measures violate the natural-order relationship during distance measurement. Sub-LOC values computed by $\sum_{h=1}^{|\boldsymbol{c}_I|\cdot|\boldsymbol{c}_J|} \xi(\boldsymbol{H}_{i,j}, \boldsymbol{H}_{I,J}(h))/(|\boldsymbol{c}_I|\cdot|\boldsymbol{c}_J|)$ defined in (11) are preprocessed using min–max scaling and demonstrated using grayscale mapping in Fig. 7(a)–(c). In these three subfigures, the larger the sub-LOC values are, the darker the corresponding grayscale blocks will be. According to the definition of LOC, there are 294, 25, and 9 sub-LOCs in total for Employee, Lecturer, and Social datasets, respectively. Accordingly, there are 294×7, 25×7, and 9×7 grayscale blocks indicating the sub-LOCs of the seven measures. In each of Fig. 7(a)–(c), the corresponding range difference [see the RD columns on the right of Fig. 7(a)–(c)] between two pairs of class labels is also demonstrated for reference. RD is computed by RD=$|I - J| - |i - j|$. According to the definition of LOC, there are 7, 3, and 2 possible RD values for

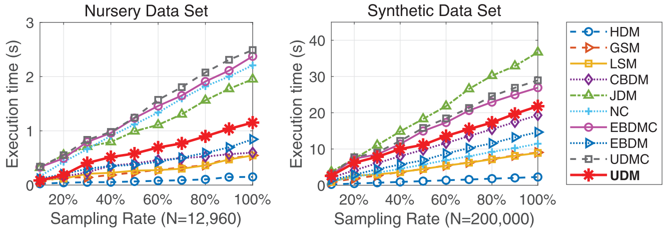

Fig. 8. Execution time on Nursery and Synthetic datasets.

Employee, Lecturer, and Social datasets, respectively. Overall LOC computed by (11) is also demonstrated in Fig. 7(d).

According to the results, it can be observed that the UDM columns shown in Fig. 7(a)–(c) are the darkest in general, and the overall LOC of UDM shown in Fig. 7(d) is higher than all the counterparts. The comparison results intuitively show that the distances between data objects measured by UDM are more consistent with the order of their labels. Three more detailed observations are discussed as follows.

1) LSM, EBDM, and UDM obviously outperform the other counterparts, because they are capable to preserve the order relationship among ordinal categories.
2) Compared to EBDM, the superiority of UDM is not significant on Employee and Lecturer datasets. To understand this, we should first know that the average interdependence degree of Employee, Lecturer, and Social datasets calculated using the proposed interdependence measure are 0.0695, 0.0807, and 0.1989, respectively. Obviously, Employee and Lecturer have relatively low interdependence degrees, and thus UDM cannot adequately extract valuable interattribute information for distance measurement. This observation also indirectly illustrates the effectiveness of the proposed interdependence measure.
3) Sub-LOC values of LSM, EBDM, and UDM are proportional to the corresponding RD values. Intuitively, according to the definition of LOC, it is hard for a measure to obtain a large LOC value when the two pairs of class labels have small RD value.

F. Efficiency Evaluation of UDM

To evaluate the efficiency of UDM, the execution time of UDM and all its counterparts is reported on the clustering of the Nursery dataset and a generated synthetic dataset with 200 000 objects represented by five ordinal and five nominal attributes, respectively, as illustrated in Fig. 8. For the synthetic dataset, the number of clusters is set at 5. Nursery dataset is chosen for the evaluation because its size is relatively large. To evaluate the changing trend of the execution time, we perform clustering on the datasets that are sampled using the different sampling rates. It can be observed that UDM will not bring extra computation cost in comparison with the state-of-the-art measures, and the computation cost of UDM is almost linear with data size, which is consistent with the time complexity analysis in Section III-D.

V. Conclusion

This article has studied the distance measurement problems under the circumstance of ordinal-and-nominal-attribute data clustering. Several rules about how to exploit valuable but heterogeneous information extracted from ordinal and nominal attributes for distance measurement are formed according to our studies. Based on the rules, a distance metric, which quantifies the distances for ordinal and nominal attributes according to extracted intra- and inter-attribute information in a unified way, has been proposed for the distance measurement of ordinal-and-nominal-attribute data. The proposed metric is parameter-free and will not bring extra computation cost compared to the existing state-of-the-art categorical data measures/metrics. More important, it is suitable for processing categorical data consisting of any-type attributes. Extensive experiments have shown the superiority of the proposed distance metric in categorical data clustering analysis.

References

[1] A. Agresti and M. Kateri, *Categorical Data Analysis*. Heidelberg, Germany: Springer, 2011.

[2] A. Agresti, *Analysis of Ordinal Categorical Data*. Hoboken, NJ, USA: Wiley, 2010.

[3] V. E. Johnson and J. H. Albert, *Ordinal Data Modeling*. New York, NY, USA: Springer, 1999.

[4] F. Fernandez-Navarro, P. Campoy-Munoz, C. Hervás-Martínez, and X. Yao, "Addressing the EU sovereign ratings using an ordinal regression approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2228–2240, Dec. 2013.

[5] D. Dua and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[7] F. Esposito, D. Malerba, V. Tamma, and H. Bock, "Classical resemblance measures," in *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information From Complex Data*, vol. 15. New York, USA: Wiley, 2000, pp. 139–152.

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

[9] 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.

[10] 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.

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

[12] D. Ienco, R. G. Pensa, and R. Meo, "From context to distance: Learning dissimilarity for categorical data clustering," *ACM Trans. Knowl. Disc. Data*, vol. 6, no. 1, pp. 1–25, 2012.

[13] H. Jia, Y.-M. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1065–1079, May 2016.

[14] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.

[15] Y. Zhang and Y.-M. Cheung, "Exploiting order information embedded in ordered categories for ordinal data clustering," in *Proc. Int. Symp. Methodol. Intell. Syst.*, 2018, pp. 247–257.

[16] Y. Zhang, Y.-M. Cheung, and K. C. Tan, "A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 39–52, Jan. 2020.

[17] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proc. SIAM Int. Conf. Data Min.*, 2008, pp. 243–254.

[18] M. Alamuri, B. R. Surampudi, and A. Negi, "A survey of distance/similarity measures for categorical data," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2014, pp. 1907–1914.

[19] T. R. dos Santos and L. E. Zárate, "Categorical data clustering: What similarity measure to recommend?" *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1247–1260, 2015.

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

[21] S. Kullback, *Information Theory and Statistics*. Chelmsford, MA, USA: Courier Corp., 1997.

[22] D. S. Yeung and X. Z. Wang, "Improving performance of similarity-based clustering by feature weight learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 556–561, Apr. 2002.

[23] W.-H. Au, K. C. C. Chan, A. K. C. Wong, and 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.

[24] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 856–863.

[25] T. Mitchell, *Machine Learning*. New York, NY, USA: McGraw Hill, 1997.

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

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[28] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.

[29] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1, pp. 81–93, 1938.

[30] Q. Hu, X. Che, L. Zhang, D. Zhang, M. Guo, and D. Yu, "Rank entropy-based decision trees for monotonic classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 2052–2064, Nov. 2012.

[31] J. H. Zar, "Significance testing of the spearman rank correlation coefficient," *J. Amer. Stat. Assoc.*, vol. 67, no. 339, pp. 578–580, 1972.

[32] T. D. Gautheir, "Detecting trends using spearman's rank correlation coefficient," *Environ. Forensics*, vol. 2, no. 4, pp. 359–362, 2001.

[33] M. G. Kendall, "Rank correlation," in *Van Nostrand's Scientific Encyclopedia*. Milton Keynes, U.K.: Open Univ., 2005.

[34] H. Abdi, "The Kendall rank correlation coefficient," in *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA, USA: Sage, 2007, pp. 508–510.

[35] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation of a preference relation by dominance relations," *Eur. J. Oper. Res.*, vol. 117, no. 1, pp. 63–83, 1999.

[36] D. Sen and S. K. Pal, "Generalized rough sets, entropy, and image ambiguity measures," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 117–128, Feb. 2009.

[37] P. Maji and P. Garai, "Fuzzy–rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1166–1177, Aug. 2013.

[38] P. Maji and P. Garai, "IT2 fuzzy-rough sets and max relevance-max significance criterion for attribute selection," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1657–1668, Aug. 2015.

[39] J. Dai, Q. Hu, J. Zhang, H. Hu, and N. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2460–2471, Sep. 2017.

[40] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behav. Sci.*, vol. 12, no. 2, pp. 153–155, 1967.

[41] Z. Huang, "Extensions to the *k*-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Disc.*, vol. 2, no. 3, pp. 283–304, 1998.

[42] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 536–543.

[43] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern Recognit.*, vol. 37, no. 5, pp. 943–952, 2004.

[44] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel attribute weighting algorithm for clustering high-dimensional categorical data," *Pattern Recognit.*, vol. 44, no. 12, pp. 2843–2861, 2011.

[45] L. Chen, S. Wang, K. Wang, and J. Zhu, "Soft subspace clustering of categorical data with probabilistic distance," *Pattern Recognit.*, vol. 51, pp. 322–332, Mar. 2016.

[46] H. Jia and Y.-M. Cheung, "Subspace clustering of categorical and numerical data with an unknown number of clusters," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3308–3325, Aug. 2018.

[47] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[48] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 507–514.

[49] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2009, pp. 175–184.

[50] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[51] A. J. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3049–3076, 2017.

[52] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[53] C.-W. Seah, I. W. Tsang, and Y.-S. Ong, "Transfer ordinal label learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 11, pp. 1863–1876, Nov. 2013.

[54] Y. Qian, H. Xu, J. Liang, B. Liu, and J. Wang, "Fusing monotonic decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2717–2728, Oct. 2015.

Yiqun Zhang (Member, IEEE) received the B.Eng. degree from the School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, in 2013, and the M.S. and Ph.D. degrees from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2014 and 2019, respectively.

He is currently a Postdoctoral Research Fellow with the Department of Computer Science, Hong Kong Baptist University. His current research interests include machine learning, data mining, and pattern recognition.

Yiu-Ming Cheung (Fellow, IEEE) 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 (HKBU), Hong Kong, and also with the HKBU Institute of Research and Continuing Education, Shenzhen, China. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, and *Neurocomputing*.