A Unified Entropy-Based Distance Metric for Ordinal-and-Nominal-Attribute Data Clustering

Yiqun Zhang, Yiu-ming Cheung, *Fellow, IEEE*, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*— Ordinal data are common in many data mining and machine learning tasks. Compared to nominal data, the possible values (also called categories interchangeably) of an ordinal attribute are naturally ordered. Nevertheless, since the data values are not quantitative, the distance between two categories of an ordinal attribute is generally not well defined, which surely has a serious impact on the result of the quantitative analysis if an inappropriate distance metric is utilized. From the practical perspective, ordinal-and-nominal-attribute categorical data, i.e., categorical data associated with a mixture of nominal and ordinal attributes, is common, but the distance metric for such data has yet to be well explored in the literature. In this paper, within the framework of clustering analysis, we therefore first propose an entropy-based distance metric for ordinal attributes, which exploits the underlying order information among categories of an ordinal attribute for the distance measurement. Then, we generalize this distance metric and propose a unified one accordingly, which is applicable to ordinal-and-nominal-attribute categorical data. Compared with the existing metrics proposed for categorical data, the proposed metric is simple to use and nonparametric. More importantly, it reasonably exploits the underlying order information of ordinal attributes and statistical information of nominal attributes for distance measurement. Extensive experiments show that the proposed metric outperforms the existing counterparts on both the real and benchmark data sets.**

***Index Terms*— Categorical data, clustering algorithms, data analysis, distance metric, entropy, order information, ordinal attribute.**

I. Introduction

In GENERAL, the types of data attributes are composed of two major classes, i.e., categorical and numerical attributes, as illustrated in Fig. 1. Under the categorical class, there are still two subclasses, i.e., nominal attributes and ordinal attributes, where ordinal attributes inherit some properties of nominal attributes [1], [2]. On the one hand, like nominal attributes, the categories (i.e., the possible values) of attributes in ordinal data, i.e., the data associated with the ordinal attributes only, are all qualitative and unsuitable for arithmetic operations such as mean, division, and summation [3]. On the other hand, the categories of an ordinal attribute are naturally ordered and comparable. Hereinafter, for simplicity, nominal (ordinal) data refer to either *nominal (ordinal) attribute* or *the data associated with the nominal (ordinal) attributes only*, which are easy to be distinguished based on the context.

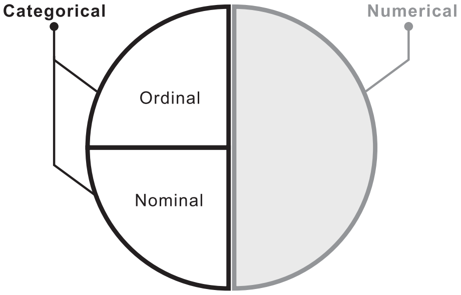


Fig. 1. Relationships among different data types.

TABLE I
Fragment of Assistant Data Set

No.	Helpful	Professional	Course	Type
1	Agree	Agree	Culture	Lecture
2	Disagree	Agree	Oral	Tutorial
3	Weak-agree	Agree	Culture	Practice
4	Agree	Weak-agree	Finance	Lab

In many real categorical data analysis tasks, both nominal and ordinal attributes exist in data sets, e.g., the data obtained through questionnaires, evaluation system, and so on. Table I shows a fragment of a teaching assistant evaluation data set, where each object, i.e., each row record in the table, is an evaluation result in terms of four attributes: Helpful, Professional, Course, and Type. If we treat ordinal attributes, i.e., Helpful and Professional, as nominal ones, the preservation of their natural order relationship may not be guaranteed. For example, the distance between Agree and Weak-agree should be smaller than that between Agree and Disagree. However, this order relationship will be ignored if we treat the categories as nominal ones. Therefore, it is more reasonable to treat ordinal attributes differently from the nominal ones to take their order information into account for data analysis.

In the literature, several works have been proposed for ordinal data regression [4]–[6], ordinal data classification [7]–[9], and ordinal data ranking [10]–[12]. Nevertheless, all of them focus on ordinal data only. In fact, from the practical perspective, the ordinal-and-nominal-attribute categorical data (i.e., the categorical data with a mixture of nominal and ordinal

Manuscript received June 27, 2018; revised December 22, 2018; accepted February 11, 2019. Date of publication March 19, 2019; date of current version January 3, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61672444 and Grant 61272366 and in part by the Hong Kong Baptist University through the Faculty Research Grant under Project FRG2/17-18/082. *(Corresponding author: Yiu-ming Cheung.)*

Y. Zhang and Y.-m. Cheung are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: yqzhang@comp.hkbu.edu.hk; ymc@comp.hkbu.edu.hk).

K. C. Tan is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: kaytan@cityu.edu.hk).

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

Digital Object Identifier 10.1109/TNNLS.2019.2899381

attributes) are common as shown in Table I. Unfortunately, as far as we know, the distance metric for such categorical data has yet to be well explored in the literature. Therefore, this paper will study the metric for such data within the framework of clustering analysis, which is generally a nontrivial task because the heterogeneous information offered by ordinal and nominal attributes should be simultaneously taken into account when assigning the data objects into the proper groups (also called *clusters* interchangeably).

Over the past two decades, a number of clustering algorithms have been proposed for the categorical data, which are essentially applicable to nominal data only. A typical example is k-modes (KM) algorithm [13], which seeks for a partition by iteratively assigning objects into their closest modes. Later, Huang *et al.* [14] propose extended KM, in which the contributions of different attributes are weighted during the data clustering process. Also, another improved version of KM has been presented in [15]. Instead of weighting the attributes for a whole data set, this version weights the contributions of each attribute for different clusters. Furthermore, another example is the entropy-based clustering (EBC) algorithms [16], [17] provided that the information entropy of a cluster will not increase a lot after adopting an object if this object is similar to the cluster. In addition, Jia and Cheung [18] have proposed a clustering algorithm, namely attribute-weighted object-cluster similarity metric-based iterative clustering learning (weighted-OCIL), which weights the attributes on each cluster by simultaneously considering their contributions in terms of intracluster similarity and intercluster difference. Although all the above-mentioned algorithms can be applied to ordinal-and-nominal-attribute categorical data, their clustering results will be degraded to a certain degree because the metrics adopted by these algorithms have not taken into account the order information of the ordinal attributes.

In fact, as far as we know, most of the existing metrics on categorical data are essentially for nominal data only. For example, the commonly used Hamming distance metric (HDM) [19]) does not consider the inherent relationships among attributes. Also, association-based distance metric (ABDM) [20] and Ahmad's distance metric (ADM) [21], [22] have been proposed, provided that, for two intraattribute categories, the distance between two categories will be shorter if the distributions of their corresponding values from the other attributes are similar to each other. However, both the metrics treat each attribute equally, which is usually unreasonable. To address this problem, a context-based distance metric (CBDM) [23], [24] has been proposed to measure the distance between two intraattribute categories according to the selected relevant attributes. Furthermore, categorical data distance metric (CDDM) proposed by Jia *et al.* [25] not only measures the distance according to the relevant attributes but also considers the occurrence probabilities of them. In this way, even all the attributes are independent of each other, this metric still works. Nevertheless, all the metrics mentioned earlier are actually proposed for nominal data, which are surely unsuitable for exploiting order information of ordinal attributes.

In the literature, some other measures have been presented to measure the similarity between two value lists according to the order of the values. For example, Kendall's rank correlation [26] and Spearman's rank correlation [27]–[29] measure the correlation degree between two variables according to the matching degree of their ranking values. However, most of the ordinal attributes in categorical data sets have a small number of possible values, which cannot provide valid ranking values for the computation of these two measures. Another measure, called rank mutual information (RMI), has been presented in [30] for monotonic classification. Similar to Kendall's and Spearman's rank correlations, RMI is designed to measure the monotonic level between attributes and is also unsuitable for the distance measurement of data objects. Recently, Hu *et al.* [30] have shown that entropy-based measures are proper for quantifying the order information of ordinal attributes. In fact, entropy-based metrics have been successfully used for nominal data clustering (see [15]–[17], [23], [24]). Therefore, along this line, proposing an entropy-based distance metric (EBDM) that can simultaneously exploit valuable information of ordinal and nominal attributes will be a feasible choice for ordinal-and-nominal-attribute categorical data clustering.

In this paper, we will propose a new categorical data distance metric that can exploit the order information of ordinal attributes and unify the heterogeneous information offered by ordinal and nominal attributes for categorical data clustering. To exploit such order information, we compute the distance between two ordinal categories according to the entropy values of all categories ordered between them (including themselves). This idea is analogous to answering a questionnaire. For example, given a multiple-choice question with the ordered choices: {very-good, good, neutral, bad, and very-bad}. When we are comparing two choices, i.e., "neutral" and "very-good," to make a final decision for this question, both of these two choices should be considered together with another choice "good" because "good" is an intermediate choice and cannot be skipped.

We also generalize the proposed metric by unifying the distance concepts of both ordinal and nominal attributes. Since the choices of a question are unordered in a nominal case, it is unnecessary to consider other choices when we are deciding the final choice from the two choices. According to this, the concept of distance has a uniform definition, which is the so-called "thinking cost" for all the choices that should be considered for choosing a choice from two choices, no matter whether the choices are ordered or not. Therefore, the information offered by ordinal and nominal attributes can be quantified and combined for indicating the distance between two data objects of an ordinal-and-nominal-attribute categorical data set. Furthermore, by taking into account the different contributions of attributes in the clustering task, we also present a unified attributes weighting scheme to adjust the contributions of different attributes.

Experimental results on different real and benchmark data sets have shown the effectiveness of the proposed distance metric for ordinal-and-nominal-attribute categorical data clustering. The main contributions of this paper are summarized into threefold.

1) An EBDM is proposed to quantify the distance between ordinal categories from the perspective of information theory. Analogous to the thinking procedure of a human being in choosing a choice from two choices, the distance between two ordinal categories is measured by calculating the entropy values of all the categories ordered between them (including themselves).
2) A unified metric featuring parameter-free, robust, and easy to use is developed by further taking the nominal case into account. It unifies the concept of distance for both ordinal and nominal categories and can be applied for the distance measurement of all types of categorical data, i.e., ordinal data, nominal data, and ordinal-and-nominal-attribute categorical data.
3) An attribute weighting scheme is designed to weight the contributions of different ordinal attributes for the distance measurement. It not only assigns a larger weight to the attributes offering more information for the distance measurement but also unifies the distance scales of different attributes. This weighting scheme is also generalized for both ordinal and nominal attributes.

The rest of this paper is organized as follows. In Section II, we will make an overview of some related metrics and measures on categorical data analysis. Section III proposes a generalized distance metric for both ordinal and nominal attributes, in which its time complexity and limitations are also discussed. Section IV gives the experimental results on both real and benchmark data sets. Finally, we draw a conclusion in Section V.

II. Overview of Related Work

This section reviews the existing related works on: 1) distance metrics proposed for categorical data clustering and 2) measures designed for ordinal data analysis.

A. Distance Metrics for Categorical Data Clustering

Five distance metrics for categorical data clustering, including HDM [13], ABDM [20], ADM [21], CBDM [24], and CDDM [25], are reviewed in this section.

HDM [13] is simple and popular for categorical data analysis. It uniformly assigns distance “1” to a pair of different categories while assigns distance “0” to a pair of identical categories. Therefore, HDM is incapable to distinguish the distances between different pairs of categories, and will thus completely ignore the order relationships among ordinal categories. Moreover, it treats each attribute equally and will also ignore the relationship among attributes.

To extract valuable information from correlated attributes for more accurate distance measurement, ABDM is proposed in [20]. It adopts the idea that if the probability distributions of the corresponding values from another attribute of two categories are dissimilar to each other, the distance between the two categories will be larger. In practice, Kullback–Leibler divergence method [31], [32] is utilized to compute the distance between two probability distributions. Later, ADM proposed in [21] and [22] adopts similar idea as the ABDM. The difference is, ADM calculates the distance between two categories according to their separating power, which is defined in [33].

All the above-mentioned metrics treat each attribute equally, which is not always reasonable. Therefore, CBDM is proposed in [23] and [24] to calculate the distance between two categories from a target attribute according to the selected relevant attributes, which are called context. For each attribute, the relevant but not redundant attributes are determined according to the symmetrical uncertainty defined in [34] as the context. Then, the distance between the two categories from the target attribute is calculated according to it.

ABDM, ADM, CBDM are all indirect metrics that measure the distance between categories according to the other attributes. Therefore, when the attributes are independent of each other, their performance will be significantly influenced. To solve this problem, and to further improve the distance measurement of categorical data, CDDM is proposed in [25]. It measures the distance by simultaneously considering the occurrence probabilities of two given categories, and their conditional probabilities of their cooccurred values from the other relevant attributes. This metric also weights the contributions of different attributes according to the idea that uncommon categories offer more valuable information for the distance measurement. Moreover, it selects the attributes according to the normalized version of mutual information [35] and calculates the distance between two categories by simultaneously considering the target attribute itself and the selected attributes.

B. Measures for Ordinal Data Analysis

Three measures, i.e., Kendall’s rank correlation [26], Spearman’s rank correlation [28], and RMI [30], have been proposed for data analysis in the literature. We discuss them in the following because they can be applied to ordinal data analysis.

Kendall’s rank correlation [26] is presented to measure the association degree between two value lists in terms of the orders of the values. It counts the number of concordant pairs and discordant pairs of observations. For the concordant case, if the ith value of list A_1 is larger (or smaller) than the jth value of A_1, and the ith value of list A_2 is also larger (or smaller) than the jth value of list A_2, then the pair of the ith and jth observations are judged to be concordant. For the discordant case, if the ith value of list A_1 is larger (or smaller) than the jth value of A_1, and the ith value of list A_2 is smaller (or larger) than the jth value of list A_2, then the pair of the ith and the jth observations are judged to be discordant. If two observations are completely the same, they are judged as neither concordant nor discordant. More concordant pairs and less discordant pairs indicate a higher order correlation level between the two ordinal value lists.

Spearman’s rank correlation [28] measures the dependence degree between two value lists in terms of their orders. It first sorts the two value lists according to the order values of one of the lists. Then, the differences between the order values of the two lists are utilized to calculate their rank correlation. An extremely high or low difference usually indicates that the orders of the two value lists are completely inverse or identical to each other. However, both the Kendall's and Spearman's rank correlations need sufficient unique order values for the rank correlation calculation. Since there are usually a small number of categories in an ordinal attributes, the two rank correlation measures are unsuitable for ordinal data analysis. Moreover, since the two measures are designed to calculate the rank correlation degree between two value lists, they are more likely to be utilized to measure the relevance degree between attributes, but not the distance between data objects.

RMI [30] is proposed for monotonic classification. It can be viewed as the extended version of mutual information. RMI exploits the order information of values to measure the monotonicity relevance between two value lists. Originally, mutual information cannot reflect the dependence in terms of orders because it is calculated by summing up the subentropies and subconditional entropies of different categories. RMI extends it by summing up the subentropies and subconditional entropies of different dominance rough sets, and therefore, competent for indicating the dependence degree between two value lists in terms of their orders. More details about rough sets theory can be found in [36]–[39]. Similar to the two above-mentioned measures, RMI is also designed to measure the relevance between two value lists, which cannot be utilized to measure the distance between two possible values of an attribute.

III. Proposed Metric

In this section, we first propose the EBDM to exploit the embedded order information for ordinal data distance measurement. Then, this metric is generalized to an ordinal-and-nominal-attribute version. Finally, we discuss the time complexity and limitations of the proposed metric.

A. Preliminaries

In this paper, for a data set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ with N data objects represented by d attributes, it is assumed that the former d_{ord} attributes are ordinal attributes: $\{A_1, A_2, \ldots, A_{d_{\mathrm{ord}}}\}$ and the latter d_{nom} attributes are nominal attributes: $\{A_{d_{\mathrm{ord}}+1}, A_{d_{ord}+2}, \ldots, A_d\}$, where $d = d_{\mathrm{ord}} + d_{\mathrm{nom}}$. Ordinal data set can be viewed as a special case that $d_{\mathrm{ord}} = d$ and $d_{\mathrm{nom}} = 0$, while nominal data set is another special case that $d_{\mathrm{ord}} = 0$ and $d_{\mathrm{nom}} = d$.

Possible values of an attribute A_r can be represented by a category set $O_r = \{O_r(1), O_r(2), \ldots, O_r(\upsilon_r)\}$, where υ_r is the number of categories of A_r. In this paper, each category is represented in the form of $O_{sa}(sc)$. Here, sa ($sa \in \{1, 2, \ldots, d\}$) stands for the sequence number of an attribute A_{sa}, and sc ($sc \in \{1, 2, \ldots, \upsilon_{sa}\}$) stands for the sequence number of a category belonging to A_{sa}. The difference between ordinal and nominal categories is that the ordinal categories from one attribute are naturally ordered, and their sequence numbers are also the order values of them. Specifically, for an ordinal attribute A_r, its categories satisfy $O_r(1) \prec O_r(2) \prec \ldots \prec O_r(\upsilon_r)$ where the symbol "$\prec$" means that the categories on its left ranked higher than the categories on its right. For a nominal attribute, the sequence numbers of its categories do not indicate the order relationships among the categories.

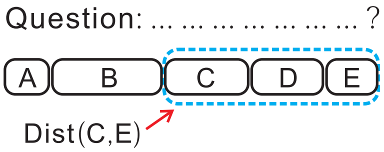


Fig. 2. Example of a multiple-choice question with ordered choices.

From the viewpoint of data objects, O_r is the value space of the rth value of a data object. Specifically, a data object $\boldsymbol{x}_i$ can be expressed by $\boldsymbol{x}_i = \{O_1(i_1), O_2(i_2), \ldots, O_{d_{ord}}(i_{d_{\mathrm{ord}}}), O_{d_{\mathrm{ord}}+1}(i_{d_{\mathrm{ord}}+1}), O_{d_{ord}+2}(i_{d_{ord}+2}), \ldots, O_d(i_d)\}$, where the former d_{ord} take values from the category sets of each of the d_{ord} ordinal attributes, while the latter d_{nom} take values from the category sets of each of the d_{nom} nominal attributes. For the ordinal part of $\boldsymbol{x}_i$, the sequence numbers $i_1, i_2, \ldots, i_{d_{\mathrm{ord}}}$ indicate that the categories ranked i_1th, i_2th,..., $i_{d_{\mathrm{ord}}}$th in category sets $O_1, O_2, \ldots, O_{d_{\mathrm{ord}}}$ have been taken by the 1st, 2nd,..., d_{ord}th values of object $\boldsymbol{x}_i$, respectively. For the nominal part of $\boldsymbol{x}_i$, the sequence numbers $i_{d_{\mathrm{ord}}+1}, i_{d_{\mathrm{ord}}+2}, \ldots, i_d$ indicate that the $i_{d_{\mathrm{ord}}+1}$th, $i_{d_{\mathrm{ord}}+2}$th,..., i_dth categories in category sets $O_{d_{\mathrm{ord}}+1}, O_{d_{\mathrm{ord}}+2}, \ldots, O_d$ have been taken by the $d_{\mathrm{ord}} + 1$th, $d_{\mathrm{ord}} + 2$th,..., dth values of object $\boldsymbol{x}_i$, respectively.

For a reasonable ordinal data distance metric, the distances produced by it should consistent with the order relationships among the ordered categories of an ordinal attribute. More specifically, the produced distance values should satisfy $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \le Dist(\boldsymbol{x}_i, \boldsymbol{x}_l)$, if $O_r(i_r) \preceq O_r(j_r) \preceq O_r(l_r)$ or $O_r(l_r) \preceq O_r(j_r) \preceq O_r(i_r)$ where $i, j, l \in \{1, 2, \ldots, N\}$, $i_r, j_r, l_r \in \{1, 2, \ldots, \upsilon_r\}$ and $r \in \{1, 2, \ldots, d_{\mathrm{ord}}\}$. Here, the symbol "$\preceq$" means that the order values of the categories on its left are not lower than that of the categories on its right.

Frequently used symbols in this paper are sorted out in Table II.

B. Entropy-Based Distance Metric

For an ordinal data set collected from questionnaires, each attribute is a multiple-choice question with all its choices naturally ordered. Ordered categories of an attribute are the ordered choices of a question. A data object is a sample containing answers to each of the questions given by a participant. If a participant is trying to select a choice from C and E for a question, all the choices between C and E including themselves, i.e., C, D, and E, have been considered by him/her as shown in Fig. 2. In other words, C, D, and E cost thinking for a participant to choose a choice from C and E. It is obvious that the thinking cost for choosing the choice from two choices is not only related to the two choices themselves but also related to the choices ordered between them, that is, if choice D costs more thinking, it will be more difficult for a participant to decide a final answer from C and E. Moreover, choosing a choice from two choices by considering more choices will cost more thinking. For instance, choosing a choice from C and E will cost more thinking than that from C and D. In addition, since all the choices are different from each other, each of them costs different thinking cost.

TABLE II
FREQUENTLY USED SYMBOLS

Symbol	Meaning
X	An $N \times d$ matrix. Values of each row represent a data object and values of each column represent an attribute. In this paper, we assume that the former d_{ord} attributes are ordinal and the latter d_{nom} ones are nominal.
N	Number of data objects in X.
d	Number of attributes in X, $d = d_{ord} + d_{nom}$.
d_{ord}	Number of ordinal attributes in X.
d_{nom}	Number of nominal attributes in X.
$\boldsymbol{x}_i$	The ith data object of X, $1 \leq i \leq N$.
$\boldsymbol{x}_i(r)$	The rth value of data object $\boldsymbol{x}_i$, $1 \leq r \leq d$.
A_r	The rth attribute of X, $1 \leq r \leq d$.
v_r	Number of categories (possible values) of A_r.
O_r	A set containing the v_r categories of A_r. If A_r is an ordinal attribute, the categories are ordered from the top (smallest order value) to the bottom (largest order value) in O_r. If A_r is a nominal attribute, there is no order relationships among the categories.
$O_r(s)$	The value of the sth category in O_r.
$\prec$	The categories on its left are ranked higher than the categories on its right.
$\preceq$	The categories on its left are not ranked lower than the categories on its right.
$\vartheta(\cdot,\cdot)$	Distance between two categories.
$E_{O_r(s)}$	Entropy value of a category $O_r(s)$ in attribute A_r, $E_{O_r(s)} = -p_{O_r(s)} \log p_{O_r(s)}$.
$p_{O_r(s)}$	Occurrence probability of value $O_r(s)$ in A_r, $p_{O_r(s)} = \sigma_{O_r(s)}/N$.
$\sigma_{O_r(s)}$	Occurrence time of the value $O_r(s)$ in A_r.
$Dist(\cdot,\cdot)$	Distance between two data objects.
ω_{A_r}	Weight of A_r, $\omega_{A_r} = \omega^I_{A_r} \cdot \omega^S_{A_r}$.
$\omega^I_{A_r}$	Importance weight of A_r.
$\omega^S_{A_r}$	Scale weight of A_r.
R_{A_r}	Reliability of A_r, $R_{A_r} = \frac{E_{A_r}}{S_{A_r}}$.
E_{A_r}	Shannon entropy of A_r, $E_{A_r} = -\sum_{s=1}^{v_r} p_{O_r(s)} \log p_{O_r(s)}$.
S_{A_r}	Standard information of A_r, $S_{A_r} = -\log \frac{1}{v_r}$.

Thinking cost of each choice has been indicated using different widths in Fig. 2.

By making an analogy between the above-discussed choice choosing issue and the ordinal data distance measurement problem, the distance between two ordinal categories can be viewed as the cost for choosing a choice from two choices. Based on this, the whole data set can be viewed as the records of the answers given by the participants, and the distance measurement problem can be viewed as the cost prediction problem. Specifically, the distance between two categories $O_r(i_r)$ and $O_r(j_r)$ from A_r with $j_r > i_r$ can be measured by estimating the cost (distance) contributions of the $j_r - i_r + 1$ categories, i.e., $O_r(i_r)$, $O_r(i_r+1)$,..., $O_r(j_r)$.

According to the above-mentioned discussions, the vital problem in ordinal data distance measurement is how to measure the distance contributions of different categories. From the perspective of information theory, a higher entropy value usually indicates a larger amount of information or more uncertainty. A choice with more information or higher uncertainty level usually costs more thinking for a participant. Therefore, the entropy value of a category is suitable for indicating its distance contribution. More specifically, sum of the entropy values of categories $O_r(i_r)$, $O_r(i_r+1)$,..., $O_r(j_r)$ indicates the distance between $O_r(i_r)$ and $O_r(j_r)$.

Therefore, the distance between the rth value of two objects, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, from an ordinal data set X with N objects represented by d_{ord} ordinal attributes, is defined as

$$\vartheta(O_r(i_r), O_r(j_r)) = \begin{cases} \displaystyle\sum_{s=\min(i_r,j_r)}^{\max(i_r,j_r)} E_{O_r(s)}, & \text{if } i_r \neq j_r \\ 0, & \text{if } i_r = j_r \end{cases} \tag{1}$$

where $\vartheta(\cdot,\cdot)$ stands for the distance between two categories, and $E_{O_r(s)}$ stands for the entropy value of category $O_r(s)$, which can be written as

$$E_{O_r(s)} = -p_{O_r(s)} \log p_{O_r(s)} \tag{2}$$

where the item $p_{O_r(s)}$ stands for the occurrence probability of value $O_r(s)$ in attribute A_r, which can be written as

$$p_{O_r(s)} = \frac{\sigma_{O_r(s)}}{N} \tag{3}$$

where $\sigma_{O_r(s)}$ is the number of data objects in the data set X with their rth values equal to $O_r(s)$. Subsequently, the distance between two ordinal data objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be written as

$$Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{r=1}^{d_{\text{ord}}} \vartheta(O_r(i_r), O_r(j_r))^2}. \tag{4}$$

The distance between ordinal categories defined in (1) is consistent with the order relationships among the ordinal categories, because a pair of categories with larger order value difference will always have a larger distance according to (1). Since (4) is the L2-norm of the distances between intraattribute categories defined in (1), the distance between the ordinal objects defined in (4) is also consistent with the order relationships among ordinal categories, that is, the distances defined in (4) satisfy $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq Dist(\boldsymbol{x}_i, \boldsymbol{x}_l)$, if $O_r(i_r) \preceq O_r(j_r) \preceq O_r(l_r)$ or $O_r(l_r) \preceq O_r(j_r) \preceq O_r(i_r)$ where $i, j, l \in \{1, 2, \ldots, N\}$, $i_r, j_r, l_r \in \{1, 2, \ldots, v_r\}$ and $r \in \{1, 2, \ldots, d_{\text{ord}}\}$, as discussed in Section III-A. However, the defined distance still has two problems.

1) The importance of attribute is not considered. The defined distance treats each attribute equally, which is not always reasonable in practice.
2) The distances measured for the categories from different attributes do not have a unified scale. An attribute with larger number of categories are more likely to produce larger distance values, which makes the measured distance somewhat unreasonable.

To solve these two problems, we present a weighting scheme in Section III-C.

C. Attribute Weighting

From the perspective of information theory, higher entropy of an attribute means that this attribute offers more information [40]. Evidently, a decision made based on larger amount of information will be more convincible. Therefore, we weight the importance of the attributes according to the information amount they offer. Specifically, the weight value for weighting the importance of an attribute A_r is defined as

$$\omega^I_{A_r} = \frac{E_{A_r}}{\sum_{s=1}^{d_{\text{ord}}} E_{A_s}} \tag{5}$$

where E_{A_r} stands for the entropy of A_r, which is defined as

$$E_{A_r} = -\sum_{s=1}^{v_r} p_{O_r(s)} \log p_{O_r(s)}. \tag{6}$$

The attributes with larger number of categories may produce larger distance values and will contribute more to the distance between two data objects. To avoid this, the weight value for weighting the scale of an attribute A_r is defined as

$$\omega^S_{A_r} = \frac{\frac{1}{S_{A_r}}}{\sum_{s=1}^{d_{\text{ord}}} \frac{1}{S_{A_s}}} \tag{7}$$

where the factor S_{A_r}, called standard information, is defined as

$$S_{A_r} = -\log \frac{1}{v_r} \tag{8}$$

where S_{A_r} is the maximum entropy of an attribute. It calculates the entropy of an attribute when the occurrence probabilities of its categories are all the same. Here, we choose to use the standard information instead of the attribute's entropy E_{A_r} for scale weighting because the standard information makes the distances from different attributes comparable. For example, given $\vartheta(O_r(i_r), O_r(j_r)) = \vartheta(O_m(i_m), O_m(j_m))$ with $v_r = v_m$, the two distances will become unequal after the scale weighting using the weight values defined by $\omega^S_{A_r} = 1/E_{A_r}/\sum_{s=1}^{d_{\text{ord}}} 1/E_{A_s}$ and $\omega^S_{A_m} = 1/E_{A_m}/\sum_{s=1}^{d_{\text{ord}}} 1/E_{A_s}$ if $E_{A_r} \neq E_{A_m}$. By using standard information, this problem can be avoided.

To simultaneously weight the attributes using the two above defined weights $\omega^I_{A_r}$ and $\omega^S_{A_r}$, the integrated weight of an attribute A_r can be written as

$$\omega_{A_r} = \omega^I_{A_r} \cdot \omega^S_{A_r}. \tag{9}$$

To explain the physical meaning of the integrated weight, we also define another concept called reliability, which is written as

$$R_{A_r} = \frac{E_{A_r}}{S_{A_r}}. \tag{10}$$

The reliability indicates the percentage of the maximum information contained by attribute A_r. The higher the R_{A_r} is, the more convincible the distances measured according to attribute A_r will be. Based on (10), the weight of A_r can be rewritten as

$$\omega_{A_r} = \frac{R_{A_r}}{\sum_{s=1}^{d_{\text{ord}}} R_{A_s}}. \tag{11}$$

Since (11) is equivalent to the weight defined in (9), it can simultaneously weight the importance and scale of an attribute.

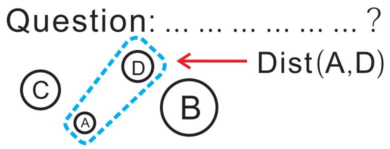


Fig. 3. Example of a multiple-choice question with nominal choices. In this question, A–D stand for “English,”“Machine Learning,” “Music,” and “Mathematics,” respectively.

Based on the distance between ordinal categories defined in (1) and the weight defined in (11), the weighted distances between ordinal categories can be written as

$$\vartheta(O_r(i_r), O_r(j_r)) = \begin{cases} \omega_{A_r} \cdot \sum\limits_{s=\min(i_r,j_r)}^{\max(i_r,j_r)} E_{O_r(s)}, & \text{if } i_r \neq j_r \\ 0, & \text{if } i_r = j_r \end{cases} \tag{12}$$

which has the following properties when $i, j, l \in \{1, 2, \ldots, N\}$, and $r \in \{1, 2, \ldots, d_{\text{ord}}\}$:

1) $\vartheta(O_r(i_r), O_r(j_r)) = 0$ iff $O_r(i_r) = O_r(j_r)$;
2) $\vartheta(O_r(i_r), O_r(j_r)) = \vartheta(O_r(j_r), O_r(i_r))$;
3) $0 \leq \vartheta(O_r(i_r), O_r(j_r)) \leq 1$;
4) $\vartheta(O_r(i_r), O_r(l_r)) \leq \vartheta(O_r(i_r), O_r(j_r)) + \vartheta(O_r(j_r), O_r(l_r))$;
5) $\vartheta(O_r(i_r), O_r(l_r)) = \vartheta(O_r(i_r), O_r(j_r)) + \vartheta(O_r(j_r), O_r(l_r)) - E_{O_r(j_r)} \cdot \omega_{A_r}$, iff $O_r(i_r) \preceq O_r(j_r) \preceq O_r(l_r)$ or $O_r(l_r) \preceq O_r(j_r) \preceq O_r(i_r)$; and
6) $\vartheta(O_r(i_r), O_r(j_r)) \leq \vartheta(O_r(i_r), O_r(l_r))$, if $O_r(i_r) \preceq O_r(j_r) \preceq O_r(l_r)$ or $O_r(l_r) \preceq O_r(j_r) \preceq O_r(i_r)$.

Based on (12), the distance between ordinal data objects defined in (4) has the following properties when $i, j, l \in \{1, 2, \ldots, N\}$.

1) $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ iff $\boldsymbol{x}_i = \boldsymbol{x}_j$.
2) $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = Dist(\boldsymbol{x}_j, \boldsymbol{x}_i)$.
3) $0 \leq Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 1$.
4) $Dist(\boldsymbol{x}_i, \boldsymbol{x}_l) \leq Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) + Dist(\boldsymbol{x}_j, \boldsymbol{x}_l)$.
5) $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq Dist(\boldsymbol{x}_i, \boldsymbol{x}_l)$, if $\forall_r \in \{1, 2, \ldots, d_{\text{ord}}\}$, $O_r(i_r) \preceq O_r(j_r) \preceq O_r(l_r)$ or $O_r(l_r) \preceq O_r(j_r) \preceq O_r(i_r)$.

D. Generalized EBDM for Categorical Data

We generalize the EBDM to make it capable for the distance measurement of categorical data with a mixture of nominal and ordinal attributes. For nominal data, we still treat the attributes as questions in a questionnaire. The difference is that the choices are not ordered. For example, there is a question “what is your favorite course?” in a questionnaire with four choices, i.e., “English,” “Machine Learning,” “Music,” and “Mathematics.” If a participant is trying to choose a choice from “English” and “Mathematics” as shown in Fig. 3, he/she will not consider the other two choices, i.e., “Machine Learning” and “Music,” because there is no order relationship among the choices.

Based on this, the concept of cost can be extended to a nominal-and-ordinal case. That is, the meaning of the cost is the thinking cost for all the choices that should be considered for choosing a choice from two choices, no matter the choices are ordered or not. Accordingly, the concept of distance induced by the concept of thinking cost can be generalized for categorical data. Specifically, given a data set X with d attributes ($d = d_{\text{ord}} + d_{\text{nom}}$, the former d_{ord} ones are ordinal attributes, and the latter d_{nom} are nominal attributes), the distance between two categories $O_r(i_r)$ and $O_r(j_r)$ can be written as

$$\vartheta(O_r(i_r), O_r(j_r)) = \begin{cases} \omega_{A_r} \cdot \sum_{s=\min(i_r,j_r)}^{\max(i_r,j_r)} E_{O_r(s)}, & \text{if } i_r \neq j_r, 0 < r \leq d_{\text{ord}} \\ \omega_{A_r} \cdot \sum_{s=i_r,j_r} E_{O_r(s)}, & \text{if } i_r \neq j_r, d_{\text{ord}} < r \leq d \\ 0, & \text{if } i_r = j_r \end{cases} \tag{13}$$

and the weight is also generalized as

$$\omega_{A_r} = \frac{R_{A_r}}{\sum_{s=1}^{d} R_{A_s}}. \tag{14}$$

The generalized distance metric defined in (13) and (14) can be utilized for calculating the distance between two categories, no matter they are ordinal or nominal. Given $i, j, l \in \{1, 2, \ldots, N\}$, and $r \in \{1, 2, \ldots, d\}$, the generalized weighted distances have the following properties.

1) $\vartheta(O_r(i_r), O_r(j_r)) = 0$ iff $O_r(i_r) = O_r(j_r)$.
2) $\vartheta(O_r(i_r), O_r(j_r)) = \vartheta(O_r(j_r), O_r(i_r))$.
3) $0 \leq \vartheta(O_r(i_r), O_r(j_r)) \leq 1$.
4) $\vartheta(O_r(i_r), O_r(l_r)) \leq \vartheta(O_r(i_r), O_r(j_r)) + \vartheta(O_r(j_r), O_r(l_r))$.

Based on (13), the distance between categorical data objects can be written as

$$Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{r=1}^{d} \vartheta(O_r(i_r), O_r(j_r))^2}, \tag{15}$$

which has the following properties when $i, j, l \in \{1, 2, \ldots, N\}$.

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

E. Discussions

This section discusses: 1) how to apply EBDM for distance measurement in clustering analysis; 2) time complexity for clustering analysis using EBDM; and 3) limitations of EBDM.

The work flow of distance measurement by using EBDM is shown in Fig. 4, and the corresponding distance measurement algorithm is shown in Algorithm 1. To save computation cost in clustering analysis, distance matrices containing the distances between each pair of categories of each attribute can be calculated according to lines 1–20 of Algorithm 1 in advance. Then, distances between data objects can be easily read off from these matrices according to line 21 of Algorithm 1.

Given a data set X with N data objects represented by d attributes, the computation procedures of EBDM-based distance measurement consists of four parts:

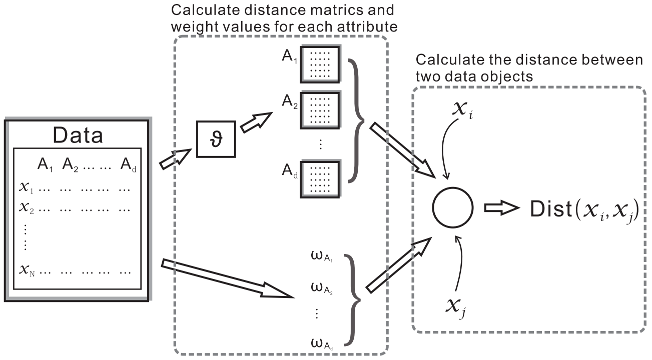


Fig. 4. Work flow of EBDM.

Algorithm 1 Distance Measurement Using EBDM

```
Input: Data set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$
Output: $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j \in \{1, 2, \ldots, n\}$
1:  for $r = 1$ to $d$ do
2:      $R_{A_r} = \frac{E_{A_r}}{S_{A_r}}$;
3:  end for
4:  for $r = 1$ to $d$ do
5:      $\omega_{A_r} = \frac{R_{A_r}}{\sum_{s=1}^{d} R_{A_s}}$;
6:  end for
7:  for $r = 1$ to $d_{\text{ord}}$ do
8:      if $i_r \neq j_r$ then
9:          $\vartheta(O_r(i_r), O_r(j_r)) = \omega_{A_r} \cdot \sum_{s=min(i_r,j_r)}^{max(i_r,j_r)} E_{O_r(s)}$;
10:     else
11:         $\vartheta(O_r(i_r), O_r(j_r)) = 0$;
12:     end if
13: end for
14: for $r = d_{\text{ord}} + 1$ to $d$ do
15:     if $i_r \neq j_r$ then
16:         $\vartheta(O_r(i_r), O_r(j_r)) = \omega_{A_r} \cdot \sum_{s=i_r,j_r} E_{O_r(s)}$;
17:     else
18:         $\vartheta(O_r(i_r), O_r(j_r)) = 0$;
19:     end if
20: end for
21: $Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{r=1}^{d} \vartheta(O_r(i_r), O_r(j_r))^2}$;
```

1) calculate the υ_r occurrence probabilities and entropy values of the categories from each attribute A_r, $r \in \{1, 2, \ldots, d\}$;
2) calculate a $\upsilon_r \times \upsilon_r$ distance matrix for each A_r according to the υ_r occurrence probabilities and entropy values;
3) calculate a weight value for each A_r according to the υ_r entropy values; and
4) read off the distance between two data objects according to the prepared distance matrices and the weights.

In clustering analysis, the computation in parts 1–3 should be executed once, and then being exploited in part 4 for distance reading off.

We analyze the time complexity of each of the four parts as follows.

1) For a data set X with N data objects represented by d attributes, time complexity for calculating the occurrence probabilities and entropy values of υ_r categories from

an attribute A_r is $O(N + v_r)$. For d attributes, the time complexity is $O(Nd + \sum_{r=1}^{d} v_r)$.

2) To calculate the distance between each pair of the v_r categories of an ordinal attribute A_r, $r \in \{1, 2, \ldots, d_{\text{ord}}\}$ according to (13), we first calculate the distances between $v_r - 1$ pairs of adjacent categories, e.g., A and B, B and C, C and D, D and E shown in Fig. 2. Then, the distances between $v_r - 2$ pairs of categories, i.e., A and C, B and D, C and E, can be calculated by simply adding the entropy values of C, D, and E to the distances between A and B, B and C, C and D, respectively. Therefore, the time complexity for producing the distance matrix of an ordinal attribute A_r is $O(v_r(v_r - 1)/2)$. If A_r is a nominal attribute, distances of $v_r(v_r - 1)/2$ pairs of its categories can be directly calculated using (13) by adding up the two entropy values of each pair. This procedure has the same time complexity as calculating a distance matrix for an ordinal attribute. Therefore, for d attributes, the time complexity is $O(\sum_{r=1}^{d} v_r(v_r - 1)/2)$.

3) To calculate the weight value of A_r, we should first add up the v_r entropy values of A_r's categories and divide it by the standard information of A_r according to (8) and (10) to obtain the reliability R_{A_r} of A_r. Then, the weight value of A_r can be obtained by dividing the reliability of A_r by the summation of the reliability of all the d attributes according to (11). Therefore, the time complexity for calculating the weight values of all the d attributes is $O(\sum_{r=1}^{d} v_r)$.

4) Time complexity for reading off the distance between two data objects according to the distance matrices and the weight values is $O(d)$.

According to the above-mentioned analysis, we will further discuss if the time complexity of EBDM will influence the time complexity of clustering analysis. Time complexity for calculating the distance matrices and weight values of all the d attributes is $O(Nd + \sum_{r=1}^{d}(2v_r + v_r(v_r - 1)/2))$. Since the value of v_r is different for each attribute, we use $V = max(v_r)$, $r \in \{1, 2, \ldots, d\}$ instead of v_r in the following analysis. Based on V, the time complexity can be rewritten as $O(Nd + Vd + V^2 d)$. Since V is usually a small constant, and $V^2 < N$ in most of the real categorical data sets, the time complexity for calculating the distance matrices using EBDM can be written as $O(Nd)$. If the distance matrices of all the attributes are given, the time complexity for partitioning the data objects into k groups in clustering analysis is at least $O(kdNI)$, where I is the number of iterations. Therefore, EBDM will not increase the overall time complexity of clustering analysis. In other words, even taking the time complexity of calculating the distance matrices for all the attributes using EBDM into account, the overall time complexity of clustering analysis is still $O(kdNI)$.

Since we focus on solving the most vital and fundamental problems in ordinal-and-nominal-attribute categorical data distance measurement, three limitations still exist in EBDM, which are discussed as follows.

1) Relationships among attributes have not been considered yet. Relevance between attributes, e.g., order correlation between ordinal attributes, and dependence between nominal attributes, may offer valuable information for distance measurement. To further exploit this kind of information, the clustering performance of EBDM could be improved.

2) Higher computation cost for analyzing streaming data. In this section, the time complexity is analyzed under the situation that the whole data set is known in advance. To cluster streaming data, the distance matrices and weight values should be updated dynamically, which may influence the clustering efficiency.

3) Attribute types should be specified in advance. EBDM does not have the ability to automatically detect which attribute is ordinal and which is nominal. For the categorical data sets with a large number of attributes, manually marking the type of each attribute will be laborious.

IV. Experiments

We embed the proposed EBDM and its counterparts into different representative clustering algorithms. Their performance on different real and benchmark data sets is evaluated using several popular validity indices. Various experiments are conducted to illustrate the efficacy of EBDM in clustering analysis.

A. Data Sets and Experimental Settings

Twelve data sets, including five real data sets, i.e., Internship, Photo, Assistant, Fruit, Pillow, and seven benchmark data sets, i.e., Employee, Lecturer, Hayth, Nursery, Solar, Voting, Tictac, are collected for the experiments. Among the twelve data sets, four of them, i.e., Internship, Photo, Employee, and Lecturer, are ordinal data sets. Another four of them, i.e., Assistant, Fruit, Hayth, and Nursery, are categorical data sets with a mixture of ordinal and nominal attributes. The remainder four, i.e., Pillow, Solar, Voting, and Tictac, are nominal data sets. Employee and Lecturer are collected from Weka web site [41]. Hayth, Nursery, Solar, Voting, and Tictac are collected from the UCI Machine Learning Repository [42]. Internship is collected from the students' questionnaires of the Education University of Hong Kong. Photo and Assistant are collected from the student questionnaires of the College of International Exchange of Shenzhen University. Fruit and Pillow are collected from the business survey of an advertising company. Statistics of the 12 data sets are shown in Table III. "Att.(O)" and "Att.(N)" indicate ordinal and nominal attributes, respectively.

To compare the metrics, we embed them into different distance-based clustering algorithms and then evaluate their clustering performance. The metrics, clustering algorithms, and validity indices are described as follows.

The commonly used HDM [13] is selected as a baseline. In addition, ADM [21], ABDM [20], CBDM [23], [24] and CDDM [25] are selected as state-of-the-art counterparts in the experiments.

KM clustering algorithm [13], [43], which is the most commonly used one for categorical data clustering, is selected as a baseline. The attribute weighted version of KM,

TABLE III

STATISTICS OF THE 12 DATA SETS

Data set	Data type	# Ins.	# Att.(O)	# Att.(N)	# Class
Internship	Ordinal	90	3	0	3
Photo	Ordinal	66	4	0	3
Employee	Ordinal	1,000	4	0	9
Lecturer	Ordinal	1,000	4	0	5
Assistant	Categorical	72	2	2	3
Fruit	Categorical	100	3	2	5
Hayth	Categorical	132	2	2	3
Nursery	Categorical	12,960	6	2	4
Pillow	Nominal	100	0	4	5
Solar	Nominal	323	0	9	6
Voting	Nominal	435	0	16	2
Tictac	Nominal	958	0	9	2

i.e., WKM [14], which can automatically weight the contributions of attributes for clustering, is also selected. In general, a subspace clustering algorithm can achieve better performance than the conventional ones. Therefore, the representative subspace clustering algorithm, i.e., entropy weighting (EW) k-means [15], and the state-of-the-art subspace clustering algorithm, i.e., weighted OCIL (WOC) [18], are also selected. Besides the above-mentioned four distance-based clustering algorithms, a representative evaluation-based clustering algorithm, i.e., EBC algorithm [17], is also selected.

For each data set, the performance of different metrics embedded in different clustering algorithms is averaged on ten runs. The clustering performance is evaluated using three powerful and popular validity indices, i.e., clustering accuracy (CA) [44], [45], adjusted rand index (ARI) [46], and normalized mutual information (NMI) [47].

CA measures the percentage of the data objects that can be correctly clustered, which is defined as CA $= \sum_{i=1}^{n} \delta(c_i, map(l_i))/n$, where c_i is the true label of the ith object, and the function $map(l_i)$ indicates the mapped label of ith object after mapping the obtained clusters to the true clusters [48]. If the mapped label and the true label of ith object are the same, $\delta(c_i, map(l_i)) = 1$, otherwise 0.

ARI is a more powerful version of Rand index (RI), which measures the agreement between the true labels and the obtained labels according to their expected agreement. ARI is defined as ARI $=$ (RI $-$ E(RI)) $\div$ (max(RI) $-$ E(RI)), where E(RI), and max(RI) stand for expected value of RI and maximum value of RI, respectively. More details about ARI and RI can be found in [44].

NMI measures the agreement between the true labels and the obtained labels from the perspective of information theory, which is defined as NMI $= (\sum_{r=1}^{k}\sum_{t=1}^{k} c_{r,t}\log(n\cdot c_{r,t}/c_r\cdot c_t)) \div ((\sum_{r=1}^{k} c_r \log c_r/N)(\sum_{t=1}^{k} c_t \log c_t/N))$, where k, $c_{r,t}$, c_r and c_t stand for the number of clusters, the number of data objects that are simultaneously assigned into the rth cluster by the obtained labels and the tth cluster by the true labels, the number of data objects that are assigned into the rth cluster by the obtained labels, and the number of data objects that are assigned into the tth cluster by the true labels, respectively.

For all these three indices, a larger value indicates better clustering performance. Both the CA and NMI have values from interval [0,1], while ARI has values from interval [-1,1]. If ARI value is less than 0, it indicates that the performance is lower than the expectation.

TABLE IV

AVERAGED CA ON FOUR ORDINAL DATA SETS

Alg.	Metric	Internship	Photo	Employee	Lecturer
KM	EBDM	**0.582±0.06**	**0.614±0.05**	**0.208±0.01**	**0.370±0.03**
	HDM	0.562±0.06	0.514±0.07	0.190±0.01	0.344±0.03
	ADM	0.533±0.01	0.503±0.04	**0.208±0.01**	0.313±0.03
	ABDM	0.528±0.01	0.538±0.08	0.200±0.01	0.316±0.02
	CBDM	0.507±0.01	0.541±0.06	0.198±0.01	0.308±0.03
	CDDM	0.558±0.02	0.486±0.07	0.189±0.01	0.331±0.04
WKM	EBDM	**0.571±0.04**	**0.535±0.09**	0.195±0.01	**0.364±0.03**
	HDM	0.559±0.05	0.486±0.09	0.192±0.01	0.339±0.04
	ADM	0.503±0.01	0.506±0.08	**0.200±0.01**	0.340±0.06
	ABDM	0.517±0.03	0.512±0.07	**0.200±0.01**	0.332±0.06
	CBDM	0.502±0.01	0.465±0.07	**0.200±0.01**	0.326±0.01
EW	EBDM	**0.608±0.07**	**0.609±0.06**	**0.207±0.01**	**0.377±0.04**
	HDM	0.558±0.03	0.532±0.06	0.193±0.01	0.344±0.04
	ADM	0.529±0.02	0.530±0.06	0.205±0.01	0.317±0.03
	ABDM	0.562±0.04	0.529±0.05	0.205±0.01	0.325±0.02
	CBDM	0.516±0.01	0.544±0.07	0.202±0.01	0.310±0.02
WOC	EBDM	**0.640±0.12**	**0.586±0.08**	**0.203±0.01**	**0.362±0.03**
	HDM	0.563±0.05	0.542±0.10	0.187±0.01	0.332±0.06
	ADM	0.508±0.02	0.498±0.08	0.201±0.01	0.325±0.02
	ABDM	0.500±0.00	0.515±0.08	0.198±0.01	0.337±0.01
	CBDM	0.500±0.00	0.568±0.04	0.197±0.01	0.318±0.02
	-	0.553±0.06	0.521±0.09	0.196±0.01	0.331±0.03
EBC	-	0.566±0.06	0.512±0.08	0.196±0.01	0.348±0.03

B. Clustering Performance on Ordinal and Categorical Data

To prove the superiority of the proposed EBDM in clustering categorical data sets with ordinal attributes, we embed EBDM and all its counterparts, i.e., HDM, ADM, ABDM, CBDM, and CDDM, into the four selected clustering algorithms, i.e., KM, WKM, EW, and WOC, and compare the clustering performance of them and the EBC clustering algorithm on the four ordinal data sets, i.e., Internship, Photo, Employee, and Lecturer, and the four categorical data sets, i.e., Assistant, Fruit, Hayth, and Nursery. In WKM, EW, and WOC clustering algorithms, distance between intraattribute categories should be calculated to update the weights of attributes. Because CDDM directly calculates the distance between objects, and cannot calculate the distance between intraattribute categories, CDDM is not embedded into them. WOC with its original object-cluster similarity measure is also compared in this experiment. In addition, since EBC is not a distance-based algorithm, we directly compare it with the other algorithms without embedding metrics into it.

Clustering performance in terms of CA, ARI, and NMI on the four ordinal data sets are compared in Tables IV–VI. Hereinafter, the experimental results highlighted by boldface and underline indicate the best and the second best results, respectively.

It can be observed that the performance of EBDM is the best on almost all the ordinal data sets, no matter which clustering algorithm is utilized. Only its CA performance on Employee data set and NMI performance on Internship data set by using WKM is not the best. However, it is still the second best, and the gap between it and the best result is very tiny, i.e., 0.005 for CA on Employee and 0.003 for NMI

TABLE V

AVERAGED ARI ON FOUR ORDINAL DATA SETS

Alg.	Metric	Internship	Photo	Employee	Lecturer
KM	EBDM	**0.024±0.04**	**0.245±0.08**	**0.026±0.00**	**0.068±0.03**
	HDM	0.004±0.05	0.096±0.07	0.009±0.01	0.045±0.01
	ADM	-0.005±0.00	0.117±0.06	0.018±0.00	0.036±0.01
	ABDM	-0.007±0.00	0.158±0.09	0.011±0.01	0.035±0.01
	CBDM	-0.014±0.00	0.117±0.06	0.014±0.00	0.033±0.02
	CDDM	0.004±0.01	0.071±0.06	0.013±0.00	0.042±0.02
WKM	EBDM	**0.011±0.02**	**0.108±0.09**	**0.018±0.01**	**0.055±0.03**
	HDM	0.008±0.03	0.072±0.09	0.013±0.00	0.042±0.03
	ADM	-0.012±0.00	0.086±0.08	0.013±0.01	0.040±0.03
	ABDM	-0.010±0.00	0.100±0.10	0.010±0.01	0.037±0.05
	CBDM	-0.014±0.00	0.050±0.07	0.017±0.00	0.027±0.00
EW	EBDM	**0.049±0.08**	**0.246±0.08**	**0.026±0.00**	**0.073±0.03**
	HDM	0.003±0.02	0.121±0.06	0.010±0.01	0.046±0.02
	ADM	-0.006±0.01	0.141±0.06	0.016±0.00	0.038±0.02
	ABDM	0.007±0.02	0.141±0.05	0.011±0.01	0.042±0.01
	CBDM	-0.012±0.00	0.115±0.06	0.015±0.00	0.035±0.01
WOC	EBDM	**0.109±0.12**	**0.171±0.09**	**0.024±0.01**	**0.060±0.01**
	HDM	0.008±0.03	0.131±0.09	0.012±0.00	0.046±0.04
	ADM	-0.010±0.00	0.082±0.08	0.016±0.01	0.033±0.01
	ABDM	-0.011±0.00	0.113±0.07	0.012±0.00	0.037±0.00
	CBDM	-0.017±0.00	0.124±0.04	0.015±0.01	0.032±0.01
	-	0.004±0.04	0.090±0.09	0.014±0.00	0.038±0.02
EBC	-	0.013±0.03	0.121±0.10	0.011±0.00	0.041±0.01

TABLE VI

AVERAGED NMI ON FOUR ORDINAL DATA SETS

Alg.	Metric	Internship	Photo	Employee	Lecturer
KM	EBDM	**0.018±0.02**	**0.281±0.09**	**0.083±0.01**	**0.092±0.04**
	HDM	0.015±0.02	0.126±0.06	0.048±0.01	0.070±0.02
	ADM	0.005±0.00	0.170±0.05	0.062±0.01	0.055±0.01
	ABDM	0.004±0.00	0.222±0.08	0.055±0.01	0.056±0.02
	CBDM	0.006±0.00	0.157±0.04	0.052±0.01	0.059±0.02
	CDDM	0.014±0.01	0.099±0.06	0.056±0.01	0.074±0.03
WKM	EBDM	0.015±0.02	**0.162±0.11**	**0.065±0.02**	**0.085±0.04**
	HDM	**0.018±0.01**	0.128±0.12	0.055±0.01	0.071±0.04
	ADM	0.003±0.00	0.146±0.09	0.051±0.01	0.065±0.04
	ABDM	0.001±0.00	0.157±0.10	0.050±0.01	0.068±0.06
	CBDM	0.004±0.00	0.113±0.07	0.061±0.01	0.052±0.01
EW	EBDM	**0.039±0.06**	**0.288±0.08**	**0.083±0.01**	**0.100±0.04**
	HDM	0.018±0.02	0.142±0.05	0.051±0.01	0.073±0.03
	ADM	0.005±0.00	0.194±0.06	0.058±0.01	0.056±0.02
	ABDM	0.012±0.01	0.213±0.06	0.057±0.01	0.065±0.02
	CBDM	0.008±0.00	0.158±0.05	0.057±0.01	0.061±0.01
WOC	EBDM	**0.068±0.08**	**0.223±0.10**	**0.077±0.01**	**0.088±0.02**
	HDM	0.016±0.01	0.196±0.11	0.052±0.01	0.072±0.05
	ADM	0.002±0.00	0.142±0.09	0.056±0.01	0.054±0.01
	ABDM	0.001±0.00	0.192±0.07	0.060±0.00	0.059±0.01
	CBDM	0.007±0.00	0.201±0.06	0.054±0.01	0.057±0.02
	-	0.020±0.02	0.140±0.10	0.054±0.01	0.061±0.03
EBC	-	0.015±0.02	0.160±0.11	0.049±0.01	0.058±0.02

TABLE VII

AVERAGED CA ON FOUR CATEGORICAL DATA SETS

Alg.	Metric	Assistant	Fruit	Hayth	Nursery
KM	EBDM	**0.603±0.07**	0.540±0.05	**0.417±0.05**	**0.384±0.02**
	HDM	0.538±0.07	0.456±0.04	0.386±0.03	0.360±0.04
	ADM	0.556±0.10	0.516±0.03	0.381±0.03	-
	ABDM	0.579±0.08	**0.548±0.06**	0.397±0.03	-
	CBDM	0.586±0.06	0.527±0.05	0.389±0.05	-
	CDDM	0.526±0.06	0.451±0.06	0.380±0.02	0.329±0.03
WKM	EBDM	0.539±0.11	0.504±0.03	**0.497±0.06**	**0.429±0.11**
	HDM	0.525±0.10	0.449±0.03	0.439±0.05	0.387±0.05
	ADM	0.560±0.12	**0.509±0.02**	0.440±0.03	-
	ABDM	**0.596±0.15**	0.499±0.01	0.359±0.04	-
	CBDM	0.499±0.07	0.494±0.03	0.405±0.08	-
EW	EBDM	**0.604±0.08**	0.546±0.05	0.402±0.06	**0.372±0.04**
	HDM	0.567±0.08	0.460±0.03	0.392±0.04	0.333±0.00
	ADM	0.565±0.09	0.516±0.03	0.379±0.03	-
	ABDM	0.568±0.07	**0.561±0.05**	**0.415±0.03**	-
	CBDM	0.581±0.06	0.525±0.05	0.386±0.05	-
WOC	EBDM	**0.628±0.10**	0.521±0.05	0.403±0.07	**0.365±0.03**
	HDM	0.508±0.10	0.496±0.05	0.379±0.05	0.360±0.04
	ADM	0.539±0.10	0.513±0.02	0.381±0.05	-
	ABDM	0.531±0.08	**0.537±0.05**	**0.413±0.05**	-
	CBDM	0.553±0.04	0.505±0.04	0.396±0.08	-
	-	0.565±0.10	0.484±0.06	0.402±0.07	0.355±0.05
EBC	-	0.522±0.07	0.447±0.04	0.360±0.04	0.360±0.04

TABLE VIII

AVERAGED ARI ON FOUR CATEGORICAL DATA SETS

Alg.	Metric	Assistant	Fruit	Hayth	Nursery
KM	EBDM	**0.210±0.07**	0.293±0.05	**0.012±0.04**	**0.068±0.02**
	HDM	0.111±0.07	0.195±0.07	-0.002±0.02	0.044±0.03
	ADM	0.161±0.08	0.283±0.04	-0.003±0.01	-
	ABDM	0.196±0.09	**0.319±0.06**	0.001±0.01	-
	CBDM	0.168±0.06	0.283±0.04	0.002±0.02	-
	CDDM	0.102±0.05	0.176±0.07	-0.004±0.01	0.030±0.02
WKM	EBDM	0.124±0.11	0.240±0.02	**0.061±0.03**	**0.124±0.17**
	HDM	0.107±0.09	0.207±0.04	0.021±0.02	0.043±0.06
	ADM	0.145±0.14	0.253±0.02	0.025±0.01	-
	ABDM	**0.215±0.20**	**0.266±0.01**	-0.009±0.01	-
	CBDM	0.061±0.06	0.261±0.03	0.011±0.04	-
EW	EBDM	**0.213±0.07**	0.297±0.05	0.007±0.04	**0.055±0.03**
	HDM	0.145±0.09	0.208±0.06	0.001±0.02	0.000±0.00
	ADM	0.172±0.06	0.283±0.04	-0.004±0.01	-
	ABDM	0.186±0.07	**0.329±0.06**	**0.011±0.01**	-
	CBDM	0.165±0.06	0.278±0.04	0.000±0.02	-
WOC	EBDM	**0.228±0.12**	0.241±0.05	**0.015±0.04**	**0.056±0.03**
	HDM	0.091±0.10	0.180±0.06	-0.001±0.02	0.044±0.03
	ADM	0.121±0.10	0.256±0.03	0.001±0.02	-
	ABDM	0.141±0.10	**0.282±0.03**	0.011±0.02	-
	CBDM	0.094±0.04	0.240±0.03	0.013±0.04	-
	-	0.142±0.09	0.205±0.06	0.012±0.04	0.023±0.02
EBC	-	0.094±0.07	0.160±0.05	0.044±0.03	0.044±0.03

on Internship. Among all the compared metrics, only EBDM has the mechanism to especially exploit the order information of ordinal attributes. This is the reason why EBDM is superior to the other counterparts in ordinal data clustering.

To further evaluate the performance of EBDM on ordinal-and-nominal-attribute data, clustering performance in terms of CA, ARI, and NMI on the four categorical data sets are compared in Tables VII–IX.

According to the results, it can be found that EBDM is still competitive for categorical data clustering, because most of the best and second best results are achieved by EBDM-based clustering algorithms. However, the superiority of EBDM in clustering categorical data is not as significant as its superiority in clustering ordinal data. This is because all the other compared metrics are actually designed for nominal attributes. A data set with more nominal attributes will, therefore, shorten the performance gap between a nominal data distance metric and EBDM. In addition, performance of ADM, ABDM and CBDM based algorithms on Nursery data set is not reported, because Nursery data set has very low inter-attribute dependence degree, which makes the ADM, ABDM, and CBDM metrics completely fail to measure the

TABLE IX

AVERAGED NMI ON FOUR CATEGORICAL DATA SETS

Alg.	Metric	Assistant	Fruit	Hayth	Nursery
KM	EBDM	**0.246±0.06**	0.460±0.04	**0.027±0.05**	**0.081±0.02**
	HDM	0.136±0.07	0.358±0.08	0.019±0.03	0.047±0.02
	ADM	0.209±0.08	0.446±0.04	0.011±0.01	-
	ABDM	0.239±0.10	**0.492±0.05**	0.013±0.01	-
	CBDM	0.190±0.05	0.447±0.04	0.017±0.03	-
	CDDM	0.122±0.07	0.322±0.09	0.012±0.01	0.036±0.03
WKM	EBDM	0.163±0.13	0.428±0.01	**0.065±0.03**	**0.156±0.23**
	HDM	0.147±0.12	0.383±0.03	0.032±0.03	0.052±0.08
	ADM	0.194±0.13	0.438±0.03	0.030±0.01	-
	ABDM	**0.256±0.18**	**0.458±0.02**	0.005±0.01	-
	CBDM	0.106±0.07	0.455±0.04	0.023±0.03	-
EW	EBDM	**0.249±0.06**	0.464±0.04	**0.024±0.05**	**0.063±0.03**
	HDM	0.166±0.09	0.371±0.08	0.021±0.04	0.000±0.00
	ADM	0.218±0.07	0.446±0.04	0.010±0.01	-
	ABDM	0.231±0.07	**0.501±0.04**	0.022±0.01	-
	CBDM	0.185±0.06	0.441±0.05	0.013±0.02	-
WOC	EBDM	**0.267±0.11**	0.419±0.04	0.026±0.03	**0.073±0.05**
	HDM	0.131±0.12	0.346±0.08	0.011±0.02	0.047±0.02
	ADM	0.171±0.11	0.436±0.02	0.012±0.02	-
	ABDM	0.197±0.09	**0.455±0.03**	0.019±0.01	-
	CBDM	0.159±0.04	0.421±0.03	0.024±0.04	-
	-	0.201±0.12	0.379±0.08	**0.030±0.05**	0.045±0.04
EBC	-	0.122±0.07	0.293±0.07	0.047±0.02	0.047±0.02

distances among categories. However, since EBDM exploits intra-attribute information for distance measurement, EBDM-based clustering algorithms are still workable on Nursery data set.

It has been pointed out in [49] and [50] that distance metric is data-sensitive and cannot always outperform the others on different data sets. Therefore, although the clustering performance of EBDM is not always the best on the above-mentioned eight data sets, the above experimental results are still sufficient to prove the effectiveness and robustness of EBDM in clustering analysis.

According to the comparison of the clustering performance of different clustering algorithms, EBC performs slightly better than the traditional KM in general. The other three state-of-the-art algorithms, i.e., WKM, EW, and WOC, are obviously more powerful because the best clustering results of each distance metric on different data sets are usually produced by one of them. Due to the space limitation, all the metrics are embedded into WOC for comparison in the following experiments.

C. Clustering Performance on Nominal Data

To illustrate that the proposed metric is also competent in clustering nominal data, we compare the clustering performance of EBDM with the other counterparts on the four nominal data sets, i.e., Pillow, Solar, Voting, and Tictac. Corresponding clustering performance is shown in Fig. 5. In this experiment, the original object-cluster similarity measure of WOC (denoted by MWOC) is also compared for completeness.

According to the results, we can find that even all the four data sets are nominal data, and all the other compared metrics are originally designed for nominal data, EBDM is still competitive. The performance of EBDM is always in the top 3, and it even outperforms the other counterparts on Solar and Tictac data sets.

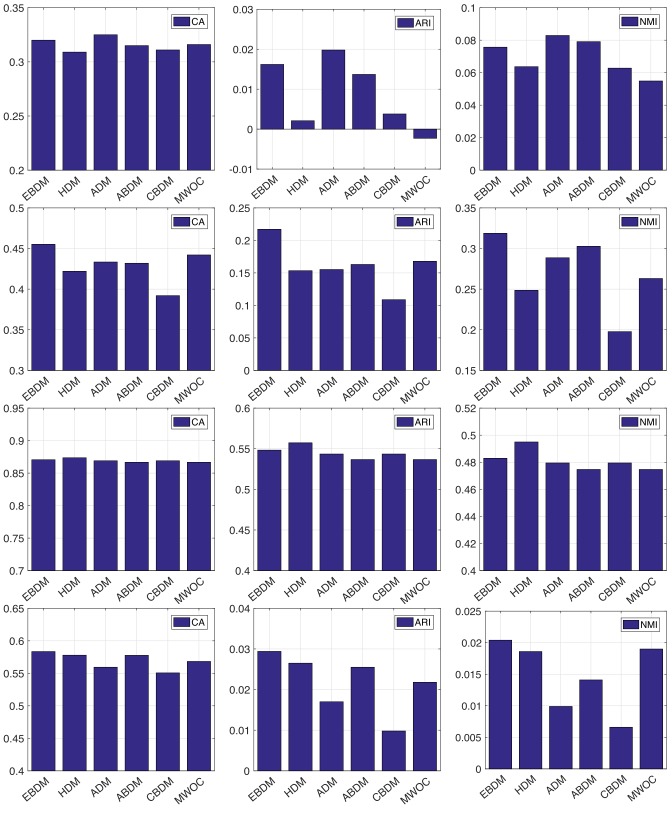


Fig. 5. Performance of EBDM, HDM, ADM, ABDM, CBDM, and the original MWOC under WOC clustering algorithm on Pillow (row 1), Solar (row 2), Voting (row 3), and Tictac (row 4) data sets.

D. EBDM and EBDM^{nom} Evaluation

The core idea of the proposed EBDM is to exploit the order information of ordinal attributes in categorical data for distance measurement. To verify the reasonableness of the order information exploiting mechanism, we compare the clustering performance of EBDM with EBDM$^{\text{nom}}$, which is the nominal version of EBDM. EBDM treats ordinal and nominal attributes differently according to (13), while EBDM$^{\text{nom}}$ treats all types of attributes as nominal one. If performance of EBDM outperforms EBDM$^{\text{nom}}$, effectiveness of the order information exploiting mechanism can be proved. Since clustering performance of EBDM and EBDM$^{\text{nom}}$ on nominal data sets are identical, experimental results on the four nominal data sets are omitted. Clustering performance of EBDM and EBDM$^{\text{nom}}$ on the eight data sets with ordinal attributes are compared in Figs. 6–8.

It can be observed from the histograms that EBDM outperforms EBDM$^{\text{nom}}$ on all the eight data sets. This indicates that EBDM is effective in exploiting the order information of ordinal attributes for more accurate clustering analysis. Since EBDM$^{\text{nom}}$ treats all the attributes as nominal ones, order information is completely ignored by it. Results of this experiment also once again proved the reasonableness of our core idea, i.e., ordinal attributes should be treated differently to exploit more valuable information for clustering analysis.

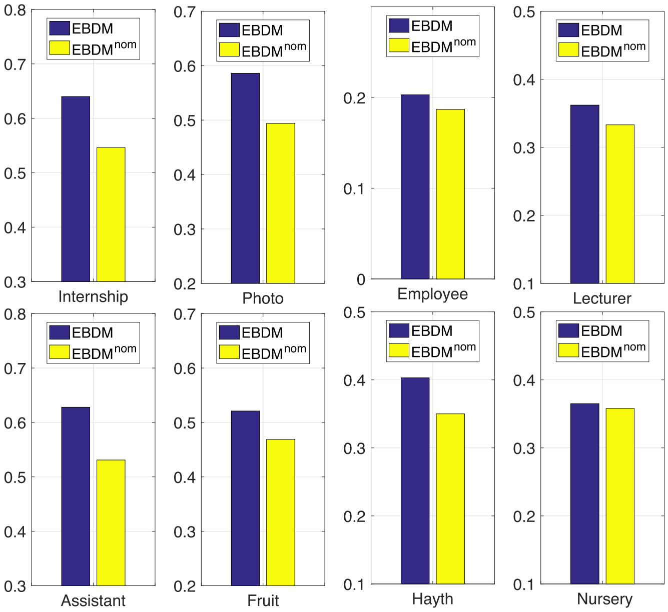


Fig. 6. Averaged CA of EBDM and EBDM^{nom} on four ordinal (row 1) and four categorical (row 2) data sets.

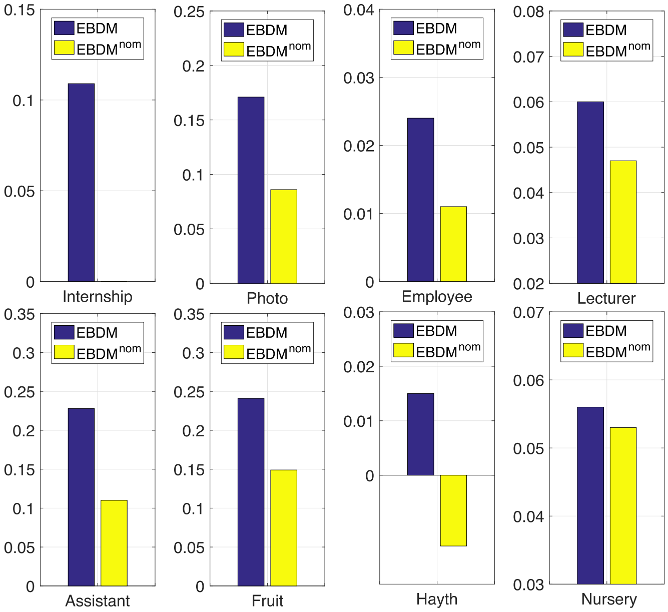


Fig. 7. Averaged ARI of EBDM and EBDM^{nom} on four ordinal (row 1) and four categorical (row 2) data sets.

E. Weighting Scheme Evaluation

To illustrate the effectiveness of the attribute weighting scheme in EBDM, we compare the clustering performance of EBDM and its no-weighting version (denoted by EBDM°) on all the twelve data sets. Their performance are compared in Tables X–XII. The best results are highlighted using boldface.

It can be observed that the clustering performance of EBDM with attribute weighting outperforms the version without attribute weighting on most of the data sets, which indicates that the attribute weighting scheme can effectively weight the contributions of different attributes during the distance measurement.

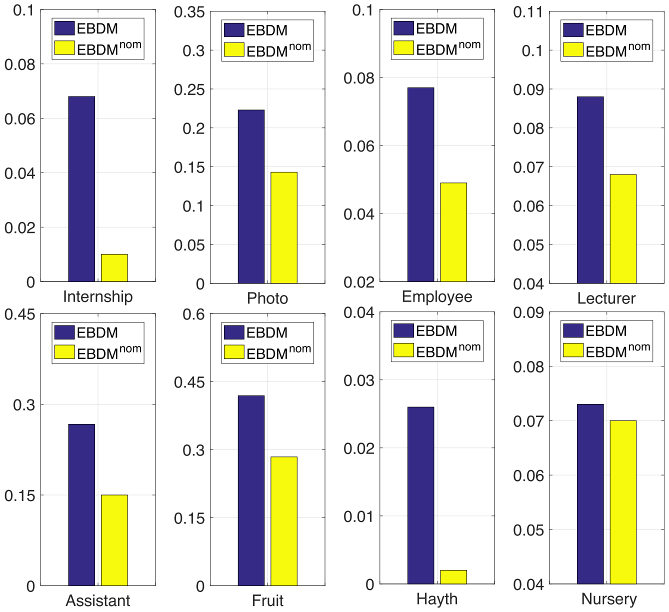


Fig. 8. Averaged NMI of EBDM and EBDM^{nom} on four ordinal (row 1) and four categorical (row 2) data sets.

TABLE X

Performance of EBDM and EBDM° on Four Ordinal Data Sets

Index	Metric	Internship	Photo	Employee	Lecturer
CA	EBDM	**0.640±0.12**	**0.586±0.08**	**0.203±0.01**	**0.362±0.03**
	EBDM^o	0.618±0.10	0.573±0.08	**0.203±0.01**	0.352±0.02
ARI	EBDM	**0.109±0.12**	**0.171±0.09**	**0.024±0.01**	**0.060±0.01**
	EBDM^o	0.073±0.09	0.166±0.09	**0.024±0.01**	0.057±0.01
NMI	EBDM	**0.068±0.08**	**0.223±0.10**	**0.077±0.01**	0.088±0.02
	EBDM^o	0.057±0.06	0.215±0.10	**0.077±0.01**	**0.091±0.02**

TABLE XI

Performance of EBDM and EBDM° on Four Categorical Data Sets

Index	Metric	Assistant	Fruit	Hayth	Nursery
CA	EBDM	**0.628±0.10**	**0.521±0.05**	**0.403±0.07**	**0.365±0.03**
	EBDM^o	0.622±0.09	0.512±0.04	0.402±0.06	0.360±0.03
ARI	EBDM	**0.228±0.12**	**0.241±0.05**	**0.015±0.04**	0.056±0.03
	EBDM^o	0.220±0.10	0.236±0.04	0.011±0.03	**0.057±0.03**
NMI	EBDM	0.267±0.11	**0.419±0.04**	**0.026±0.03**	**0.073±0.05**
	EBDM^o	**0.270±0.10**	0.416±0.04	0.023±0.03	0.070±0.04

TABLE XII

Performance of EBDM and EBDM° on Four Nominal Data Sets

Index	Metric	Pillow	Solar	Voting	Tictac
CA	EBDM	**0.320±0.02**	**0.455±0.07**	**0.871±0.00**	**0.584±0.03**
	EBDM^o	0.316±0.02	0.425±0.03	**0.871±0.00**	0.577±0.03
ARI	EBDM	**0.016±0.02**	**0.217±0.10**	0.548±0.00	**0.029±0.02**
	EBDM^o	0.013±0.03	0.186±0.06	**0.549±0.00**	0.026±0.02
NMI	EBDM	**0.076±0.02**	**0.319±0.10**	**0.483±0.00**	**0.020±0.01**
	EBDM^o	0.074±0.02	0.289±0.06	**0.483±0.00**	0.018±0.01

F. Distance Matrices Demonstration

To intuitively observe if the distances produced by different metrics are consistent with the natural distance structure of the data sets, we compare the distance matrices produced by different metrics. All the distance values are normalized into the interval [0,1], and the distance matrices are converted into

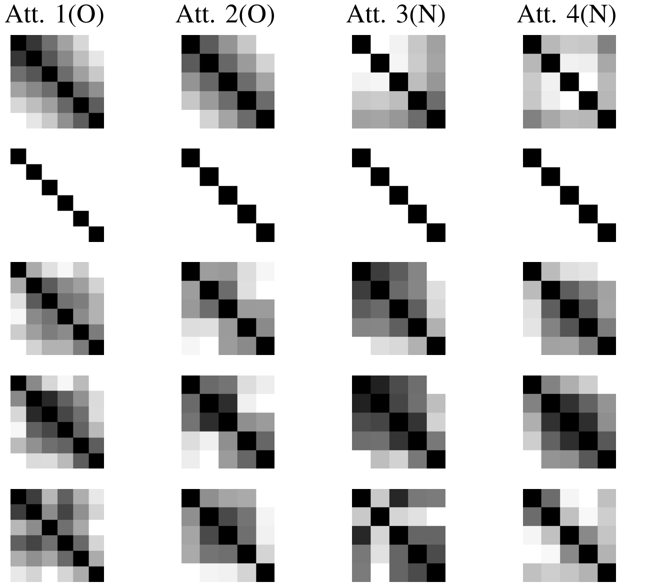


Fig. 9. Distance matrices produced by EBDM (row 1), HDM (row 2), ADM (row 3), ABDM (row 4), and CBDM (row 5) for each attribute of Assistant data set.

gray-scale maps accordingly. Lighter pixels indicate larger distance and vice versa. Therefore, for the distance matrix of an ordinal attribute, pixels on the diagonal from left-top corner to the right-bottom corner should be pure black, while the pixels locate toward the right-top and left-bottom corners should be lighter. Distance matrices of Assistant data set are demonstrated in Fig. 9. "(O)" and "(N)" indicate ordinal and nominal attributes, respectively. CDDM metric is not compared in this experiment because it cannot directly compute the distance between intraattribute categories.

It can be observed that only the distance matrices produced by EBDM are completely consistent with the order relationship among the categories of the two ordinal attributes. Since HDM assigns distances "0" and "1" to all the pairs of identical and different categories, it is incapable to indicate the distance structures of ordinal attributes. The distance matrices produced by ADM, ABDM, and CBDM can roughly indicate the distance structures of the two ordinal attributes, but a certain amount of distances produced by them are still disordered, i.e., the pixels are not gradually lighter toward the right-top and left-bottom corners from the diagonal. This is also the reason why their performance is superior to HDM, but is inferior to EBDM as shown in the experimental results in Section IV-B. For the other two nominal attributes, it is reasonable that their distance matrices produced by all the compared metrics are unordered.

V. Conclusion

In this paper, we have proposed a distance metric for categorical data clustering, called EBDM, from the perspective of information entropy. In contrast with the existing categorical data metrics, the proposed one treats ordinal attributes and nominal attributes differently but unifies the concept of the distance and importance of them, which avoids information loss during the distance measurement. For ordinal attributes, the order information is taken into account for the distance measurement, while for the nominal attributes, statistical information is exploited. Since the distance concepts of ordinal and nominal attributes are unified, it is unnecessary to separately compute the distances on ordinal and nominal attributes, and then weight and combine them to produce the final distances. Moreover, the proposed metric is easy to use and nonparametric, which can be easily applied for the clustering analysis of different types of categorical data. Experiments have shown that the proposed EBDM metric outperforms its counterparts on different real and benchmark categorical data sets.

References

[1] V. E. Johnson and J. H. Albert, *Ordinal Data Modeling*. New York, NY, USA: Springer, 2006.

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

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

[4] M. Pérez-Ortiz, P. A. Gutiérrez, and C. Hervás-Martínez, "Projection-based ensemble learning for ordinal regression," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 681–694, May 2014.

[5] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.

[6] Y. Xiao, B. Liu, and Z. Hao, "A maximum margin approach for semisupervised ordinal regression clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1003–1019, May 2016.

[7] Q. Hu *et al.*, "Feature selection for monotonic classification," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 69–81, Feb. 2012.

[8] 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.

[9] 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.

[10] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 585–592.

[11] F. Fernández, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2075–2085, Nov. 2014.

[12] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.

[13] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[14] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.

[15] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.

[16] D. Barbará, Y. Li, and J. Couto, "COOLCAT: An entropy-based algorithm for categorical clustering," in *Proc. 11th Int. Conf. Inf. Knowl. Manage.*, Nov. 2002, pp. 582–589.

[17] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering," in *Proc. 21st Int. Conf. Mach. Learn.*, Jul. 2004, p. 68.

[18] 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.

[19] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, Apr. 1950.

[20] 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.

[21] 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.

[22] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, Nov. 2007.

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

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

[25] 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.

[26] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, Jun. 1938.

[27] G. W. Corder and D. I. Foreman, *Parametric Statistics: A Step-by-Step Approach*. Hoboken, NJ, USA: Wiley, 2014.

[28] W. Pirie, "Spearman rank correlation coefficient," in *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: Wiley, 2006, pp. 502–505.

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

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

[32] S. Kullback, *Information Theory and Statistics*. North Chelmsford, MA, USA: Courier Corporation, 1997.

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

[34] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 856–863.

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

[36] 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, Aug. 1999.

[37] S. Greco, B. Matarazzo, and R. Slowinski, "Rough sets methodology for sorting problems in presence of multiple attributes and criteria," *Eur. J. Oper. Res.*, vol. 138, no. 2, pp. 247–259, Apr. 2002.

[38] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation by dominance relations," *Int. J. Intell. Syst.*, vol. 17, no. 2, pp. 153–171, Feb. 2002.

[39] J. W. T. Lee, D. S. Yeung, and E. C. C. Tsang, "Rough sets and ordinal reducts," *Soft Comput.*, vol. 10, no. 1, pp. 27–33, Jan. 2006.

[40] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[41] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools With Java Implementations*. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[42] D. Dua and E. K. Taniskidou, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

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

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

[45] K. Wagstaff, C. Claire, R. Seth, and S. Stefan, "Constrained k-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach. Learn.*, vol. 1, Jun. 2001, pp. 577–584.

[46] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Oct. 2010.

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

[48] L. Lovasz and M. D. Plummer, *Matching Theory*. Amsterdam, The Netherlands: North Holland, 1986.

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

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

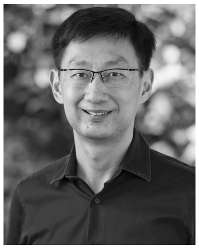
Yiqun Zhang received the B.Eng. degree from the School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China, in 2013, and the M.Sc. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Computer Science.

His current research interests include machine learning, data mining, and pattern recognition.

Yiu-ming Cheung (SM'06–F'18) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

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

Dr. Cheung is the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section and the Chair of the Technical Committee on Intelligent Informatics of the IEEE Computer Society. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, PATTERN RECOGNITION, KNOWLEDGE AND INFORMATION SYSTEMS. He is a fellow of IET, BCS, and RSA.



Kay Chen Tan (SM'08–F'14) received the B.Eng. degree (Hons.) in electronics and electrical engineering and the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is currently a Full Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. He has authored or co-authored more than 200 refereed articles and 6 books.

Dr. Tan is an Elected Member of the IEEE CIS AdCom from 2017 to 2019. He served as the Editor-in-Chief for *IEEE Computational Intelligence Magazine* from 2010 to 2013. He serves as the Editor-in-Chief for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and as the Editorial Board Member of more than ten journals.