Learnable Weighting of Intra-Attribute Distances for Categorical Data Clustering with Nominal and Ordinal Attributes

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

Abstract—The success of categorical data clustering generally much relies on the distance metric that measures the dissimilarity degree between two objects. However, most of the existing clustering methods treat the two categorical subtypes, i.e., nominal and ordinal attributes, in the same way when calculating the dissimilarity without considering the relative order information of the ordinal values. Moreover, there would exist interdependence among the nominal and ordinal attributes, which is worth exploring for indicating the dissimilarity. This paper will therefore study the intrinsic difference and connection of nominal and ordinal attribute values from a perspective akin to the graph. Accordingly, we propose a novel distance metric to measure the intra-attribute distances of nominal and ordinal attributes in a unified way, meanwhile preserving the order relationship among ordinal values. Subsequently, we propose a new clustering algorithm to make the learning of intra-attribute distance weights and partitions of data objects into a single learning paradigm rather than two separate steps, whereby circumventing a suboptimal solution. Experiments show the efficacy of the proposed algorithm in comparison with the existing counterparts.

Index Terms—Categorical data clustering, nominal-and-ordinal attribute, intra-attribute distance, learnable weighting

1 Introduction

Widespread categorical data can be easily collected from questionnaires, medical scales, scoring systems, and so on [1]. As one of the most widely used machine learning and pattern recognition techniques, clustering that partitions data objects into homogeneous groups in unsupervised environment [2], [3] has been commonly adopted for the analysis of categorical data [4], [5]. In order to better discover homogeneous clusters, weighting attributes according to their importance to the clustering task [6] is adopted by many existing clustering algorithms [7], [8], [9], [10], [11]. Since weighting an attribute is equivalent to uniformly weighting all the intra-attribute distances measured on this attribute, these algorithms are actually based on the hypothesis that all the intra-attribute distances are well defined, which is reasonable for numerical data with well-defined distance measure [12]. However, for categorical data whose distance measure is generally not well-defined, uniformly weighting the intra-attribute distances is surely unreasonable [13]. To solve this problem, most existing methods focus on exploring appropriate distance measures [14], [15] and attribute weighting mechanisms [11].

Successful attempts in exploring appropriate distance measures include Lin's [16] similarity measure, coupled [17] similarity metric, association-based [18], Ahmad's [19], context-based [20], [21], and Jia's [22] distance metrics. The above-mentioned measures define intra-attribute distances according to the possible value statistics, e.g., the occurrence frequencies and conditional occurrence probabilities. Lin's measure computes the cumulative entropy of a range of ordered possible values (i.e., the adjacent possible values {good, neutral, bad} of an ordinal attribute with possible values {very-good, good, neutral, bad, very-bad}) to indicate the corresponding intra-attribute distance (i.e., the distance between good and bad) with preserving the order relationship, which is suitable for the distance measurement of ordinal data. The others define intra-attribute distances according to the context information reflected by conditional probability distributions between interdependent attributes, which works well for nominal data. In recent years, more powerful representation-based methods including structure-based [23], coupled [24], [25], and heterogeneous coupling [26] representations, have been proposed to represent categorical data by embedding more informative and complex relationships existing in the level of values, attributes, and objects, so as to achieve a more reasonable distance measurement. Unfortunately, they still work well for nominal data only.

In summary, all the above mentioned measures are proposed without considering a very common situation that real categorical data are usually composed of a mixture of nominal and ordinal attributes [27], [28]. As the fragment of medical scale data set shown in Table 1, the values of

- *Yiqun Zhang is with the School of Computers, Guangdong University of Technology, Guangzhou 510080, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: yqzhang@gdut.edu.cn.*
- *Yiu-ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: ymc@comp.hkbu.edu.hk.*

Manuscript received 10 July 2020; revised 28 Jan. 2021; accepted 30 Jan. 2021. Date of publication 3 Feb. 2021; date of current version 3 June 2022.
(Corresponding author: Yiu-ming Cheung.)
Recommended for acceptance by S. Li.
Digital Object Identifier no. 10.1109/TPAMI.2021.3056510

TABLE 1
Fragment of Lymphography Data Set

No.	Attribute 1 (enlarge)	Attribute 2 (form)	Class (diagnosis)
1	↑	non-special	normal find
2	↑	vesicles	fibrosis
3	↑↑	vesicles	fibrosis
4	↑↑↑	chalices	metastasis
5	↑↑↑	chalices	malign lymph
6	↑↑↑↑	vesicles	malign lymph

ordinal Attribute 1 stand for the degrees of lymph enlargement, the values of nominal Attribute 2 stand for the special form of lymph, and the values of the Class attribute indicate the diagnosis results, which are the desired true cluster labels in cluster analysis.

In medicine, it is generally believed that the severity of fibrosis, metastasis, and malign lymph increases in sequence. Apparently, if we treat the ordered values of Attribute 1 as nominal values, information provided by the monotonic relationship between the values of Attribute 1 and the true class labels will be lost [29], which will directly affect the clustering accuracy. Moreover, there also exists an awkward gap between the cluster information provided by nominal and ordinal attributes, because the values of an ordinal attribute contain the relative order information, but the values of a nominal one do not. Hence, to avoid the loss of important information, entropy-based distance metrics [30], [31] have been proposed to quantify intra-attribute distances of nominal and ordinal attributes as information entropy [32] in a unified way. However, they have not established an essential connection between nominal and ordinal attributes for data clustering.

As for attribute weighting mechanism, most efforts have tried to weight attributes for each cluster, which is called subspace clustering. Typical subspace approaches include [9], [10], [11], which learn the different weight combinations of attributes for each cluster to explore more appropriate subspaces for gathering homogeneous data objects. Nevertheless, they uniformly weight all intra-attribute distances measured on the same attribute, which still makes these approaches incompetent in adapting the contributions of different intra-attribute distances to search for more appropriate clustering results. Most recently, a distance weighting-based clustering algorithm [33] has been proposed to learn the weights of intra-attribute distances automatically during clustering. This algorithm has remarkable performance on ordinal data sets, but it relies on the order relationship among attribute values for learning the distance weights, which makes it applicable to ordinal data only. To the best of our knowledge, clustering algorithm that can learn the weights of intra-attribute distances for categorical data with nominal and ordinal attributes has yet to be proposed.

In this paper, we will propose a new clustering method composed of a novel distance definition and an automatic distance weighting mechanism for any-type categorical data clustering, i.e., clustering data composed of any combination of nominal and ordinal attributes. Specifically, we study the intrinsic difference and connection of nominal and ordinal attributes, and convert each possible value of nominal attributes, e.g., “vesicles” of Attribute 2 as shown in Table 1, into a Boolean attribute with two possible values “vesicles” and “not vesicles”. Such Boolean attribute is a special case of ordinal attribute, i.e., an ordinal attribute with two extreme degrees “vesicles” and “not vesicles”. Thus, the heterogeneous clustering information provided by nominal and ordinal attributes becomes homogeneous information provided by ordinal attributes. On this basis, the information provided by interdependent attributes in three cases (i.e., (i) both attributes are nominal, (ii) both attributes are ordinal, and (iii) one is nominal and the other is ordinal) is utilized to measure intra-attribute distances of nominal and ordinal attributes in a unified way. Since the defined distances are not connected to a certain clustering task, we also propose a novel intra-attribute distance weighting mechanism to learn the distance weights iteratively based on the present data partition result to search for better clustering results. The proposed distance definition and weighting mechanism are complementary to each other in clustering. It turns out that the clustering algorithm utilizing them is competent for the cluster analysis of any-type categorical data. The main contributions of this paper are summarized below:

- Inherent connection of nominal and ordinal attributes is studied, and a novel measure suitable for intra-attribute distance measurement of any-type categorical data clustering is proposed accordingly.
- An intra-attribute distance weighting mechanism that iteratively updates the distance weights to search for better data partitions, if any, is designed to make the measured intra-attribute distances learnable.
- A new categorical data clustering algorithm is presented by utilizing the learnable distance measure. Given the number of sought clusters (which is a common setting in cluster analysis), this algorithm is parameter-free and has superior clustering performance on any-type categorical data.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 formulates the research problems. A design of homogeneous distance metric is proposed in Section 4. Then, Section 5 introduces a new clustering algorithm with the novel distance weighting mechanism as the core. Experimental results are given in Section 6. Finally, we draw a conclusion in Section 7.

2 Related Work

This section makes an overview of the existing related works on categorical data clustering.

2.1 Distance Measure

The distance measures for categorical data clustering can be generally categorized as the direct, context-based, and representation-based ones. The simplest direct measure [34] directly assigns distances 0 and 1 to identical and different intra-attribute values, respectively. The other direct measures [16], [30], [35] compute the intra-attribute distance between two possible values according to their occurrence frequencies. Direct measures are easy to use and have demonstrated great computational efficiency because their computation does not involve parameter selection, context information extraction, iterative learning, etc. However, since the valuable information provided by the correlated attributes is totally ignored, intra-attribute distances defined by them are not always reasonable in indicating the real dissimilarity degrees.

In contrast, the context-based measures [17], [18], [19], [20], [21], [22], [31] compute the distance between two intra-attribute values based on the context information, i.e., the statistical information provided by the other attributes that are correlated with the target one. In general, these measures outperform the direct ones, but their performance dependents more on the interdependence of attributes. For the data composed of independent attributes, some measures [18], [19], [20], [21] that are based on the sole information provided by the interdependent attributes would even fail for distance measurement.

Among all the above-mentioned direct and context-based measures, the two measures [30], [31] that unify the distance concept of nominal and ordinal attributes as the information divergence to avoid information loss are suitable for any-type categorical data clustering. Nevertheless, they only provide scale-level distance unification, but have yet to consider the intrinsic connection between nominal and ordinal attributes.

The representation-based distance measures encode categorical values into numerical ones, and then the advanced distance measures and clustering algorithms proposed for numerical data can be utilized. In many practical application scenarios, the encoding is performed by domain experts, which makes the performance sensitive to the prior knowledge. Further, for large-scale, high-dimensional, and multivariate categorical data, the encoding process is a laborious and non-trivial task. A commonly adopted way to circumvent these issues is to simply encode each possible value of nominal attributes into a binary-valued numerical attribute and the ordered possible values of each ordinal attribute into consecutive integers, which is called simple coding. It turns out that simple coding is applicable to any-type categorical data. Nevertheless, since it ignores the original statistical information of possible values, and it assigns the identical distance to different possible value pairs, empirical studies in [33] have shown that its performance is generally worse than the measures specially designed for categorical data. Recently, representation learning methods [23], [25], [26] have been proposed for automatically encoding categorical data in unsupervised environment. The one called SBC [23] reconstructs the original data set according to the inter-object dissimilarities. CDE in [25] encodes the original data set by performing k-means clustering and PCA on intra- and inter-attribute couplings. The newly proposed UNTIE [26] represents data set by using more types of couplings learned in multiple kernel spaces, and achieves superior clustering performance. However, all the above-mentioned representation learning methods are actually designed for nominal data only, and their performance somewhat depends on the non-trivial selection of parameters or kernel functions.

2.2 Clustering Algorithm

From the perspective of attribute weighting, the existing categorical data clustering algorithms can be roughly categorized as the non-attribute-weighting and attribute-weighting ones, respectively. As a non-attribute-weighting algorithm, the conventional k-modes [36] adopts Hamming distance [34] as a distance measure to compute the distance between data objects and the k modes. Based on the object-mode distances, it iteratively searches for better partitions of data set. Furthermore, some of its variants also focus on improving its robustness and scalability [37], [38]. In addition, clustering algorithm adopting entropy as a measure [39] has been proposed in the literature. It computes the entropy value of the present partition after moving an object into a cluster, and performs cluster analysis by searching for the partition with the minimum entropy value [40]. In general, all the above-mentioned algorithms assume that the attributes are of identical importance for clustering tasks, which is, however, not always true in practice.

In the literature, an attribute weighting-based categorical data clustering algorithm [7] has been proposed provided that the attributes are of different importance. It assigns different weights to the attributes according to their contributions in forming more compact clusters. That is, if the total distance between data objects and their clusters measured on a certain attribute is low, it indicates that this attribute contributes more than the others in forming the clusters with similar objects. Subsequently, a higher weight is thus assigned to this attribute in the next iteration to search for more compact clusters. Nevertheless, this weighting mechanism finds only a certain attributes' subset that is important to a certain subset of clusters, which is evidently incompetent in a more complex case. Therefore, subspace clustering algorithms [9], [10], [11] that weight each attribute according to its contribution in forming each certain cluster have been proposed.

In general, weighting an attribute is equivalent to uniformly weighting all the distances measured on it. Thus, all the above-mentioned attribute weighting-based algorithms actually assume that the distance measure can accurately indicate the intra-attribute distances. If the adopted distance measure is not appropriately defined, uniformly weighting the intra-attribute distances measured by them will just bring more irrationality into the clustering process. Therefore, the most recently proposed clustering algorithm [33] addresses this issue by iteratively weighting the importance of intra-attribute distances according to the present partition to search for more appropriate clustering results of the data set. Unfortunately, distance weighting of this algorithm relies on the order relationship among intra-attribute values, which makes it only applicable to the categorical data sets composed of ordinal attributes.

3 Problem Statement

We formulate the problem of distance weighting-based clustering of categorical data in this section. Table 2 lists the notations and symbols used in this paper.

A categorical data set S can be represented as a tuple $S = < X, A, O >$, where $X = \{\mathbf{x}_i | i \in N_X\}$ is the object set with n elements, and $N_X = \{1, 2, \ldots, n\}$ is the index set of X. For attribute set A composed of d attributes, we assume that the former $d^{(\mathrm{ord})}$ attributes are ordinal and the latter $d^{(\mathrm{nom})}$ attributes are nominal for convenience without loss of generality,

TABLE 2
Style of Notations and Explanation of Symbols

Notation (example)	Style
Attribute index (e.g., A^r)	Superscript
Value note (e.g., $d^{(\mathrm{ord})}$)	Superscript with parentheses
Function (e.g., $\mathrm{dist}(\cdot,\cdot)$)	Parentheses
Space (e.g., $\mathcal{R}_0^+$)	Uppercase, calligraphic font
Vector (e.g., $\mathbf{p}_l^r$)	Lowercase, bold font
Matrix (e.g., $\mathbf{Q}$)	Uppercase, bold font
Symbol (example)	**Explanation of example**
$\emptyset$ (e.g., $A^{(\mathrm{ord})} = \emptyset$)	$A^{(\mathrm{ord})}$ is an empty set
$\top$ (e.g., $[x_i^1, x_i^2, \ldots, x_i^d]^\top$)	Transpose of $[x_i^1, x_i^2, \ldots, x_i^d]$
$\succ$ (e.g., $o_1^r \succ o_2^r$)	o_1^r ranks higher than o_2^r
$\neg$ (e.g., $\neg o_g^s$)	A^s's possible values excluding o_g^s

and we have $d^{(\mathrm{ord})} + d^{(\mathrm{nom})} = d$. Formally, $A^{(\mathrm{ord})} = \{A^r | r \in N_A^{(\mathrm{ord})}\}$ is the ordinal attribute set, $A^{(\mathrm{nom})} = \{A^s | s \in N_A^{(\mathrm{nom})}\}$ is the nominal attribute set, $N_A^{(\mathrm{ord})} = \{1, 2, \ldots, d^{(\mathrm{ord})}\}$ and $N_A^{(\mathrm{nom})} = \{d^{(\mathrm{ord})}+1, d^{(\mathrm{ord})}+2, \ldots, d\}$ are the index sets of $A^{(\mathrm{ord})}$ and $A^{(\mathrm{nom})}$, respectively. $A = A^{(\mathrm{ord})} \cup A^{(\mathrm{nom})}$ is the complete attribute set, and $N_A = N_A^{(\mathrm{ord})} \cup N_A^{(\mathrm{nom})}$ is the complete index set of A. Accordingly, three types of categorical data can be distinguished by

$$\mathrm{datatype}(S) = \begin{cases} \text{mixed}, & A^{(\mathrm{ord})} \neq \emptyset, \quad A^{(\mathrm{nom})} \neq \emptyset \\ \text{ordinal}, & A^{(\mathrm{ord})} \neq \emptyset, \quad A^{(\mathrm{nom})} = \emptyset \\ \text{nominal}, & A^{(\mathrm{ord})} = \emptyset, \quad A^{(\mathrm{nom})} \neq \emptyset. \end{cases} \tag{1}$$

Hereinafter, a categorical data set composed of a mixture of ordinal and nominal attributes, pure ordinal attributes, and pure nominal attributes is called mixed, ordinal, and nominal data set, respectively. $O^r = \{o_m^r | m \in N_O^r\}$ is the set of v^r possible values of attribute A^r, and $N_O^r = \{1, 2, \ldots, v^r\}$ is the index set of A^r's possible values. The ith object of X is represented as $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^d]^\top$ with $x_i^r \in O^r$, $r \in N_A$. If A^r is an ordinal attribute (i.e., $r \leq d^{(\mathrm{ord})}$), its possible values satisfy $o_1^r \succ o_2^r \succ \ldots \succ o_{v^r}^r$ where the symbol "$\succ$" indicates that the values on its left are rank higher than the values on its right.

In crisp partitional clustering task, X is partitioned into k clusters, which can be represented as a cluster set $C = \{C_l | l \in N_C\}$ with $N_C = \{1, 2, \ldots, k\}$. Accordingly, X can be represented as a collection of k disjoint subsets $X = \bigcup_{l=1}^k X_{C_l}$ where X_{C_l} is the object set corresponding to the lth cluster. The k clusters are represented by their corresponding statistical information $P = \{P_l | l \in N_C\}$ where $P_l = \{\mathbf{p}_l^r | r \in N_A\}$ is the statistical information of C_l and $\mathbf{p}_l^r = [p_{l1}^r, p_{l2}^r, \ldots, p_{lv^r}^r]^\top$ is the probability distribution of the rth values of the objects in C_l. Values of P are dependent on $\mathbf{Q}$, which is an $n \times k$ matrix indicating the partition of X. The (i, l)th entry of $\mathbf{Q}$ is denoted as q_{il}. If $\mathbf{x}_i$ belongs to C_l, we have $q_{il} = 1$, otherwise, $q_{il} = 0$. To learn the importance of intra-attribute distances, we solve the clustering problem in a distance weighting framework. The weights of intra-attribute distances are denoted as a set of matrices $W = \{\mathbf{W}^r | r \in N_A\}$ where $\mathbf{W}^r$ is a $v^r \times v^r$ symmetric matrix storing the weights of intra-attribute distances of A^r. The (m, h)th entry of $\mathbf{W}^r$ is denoted as w_{mh}^r, which represents the weight of the distance between possible values o_m^r and o_h^r. The clustering problem can be formulated as minimizing the objective function

$$Z(\mathbf{Q}, P, W) = \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il} \mathrm{dist}(\mathbf{x}_i, C_l) \tag{2}$$

$$s.t. \begin{cases} \sum_{l=1}^k q_{il} = 1, \quad q_{il} \in \{0, 1\}, & i \in N_X, \\ \sum_{r=1}^d \sum_{m=1}^{v^r - 1} \sum_{h=m+1}^{v^r} w_{mh}^r = 1, & w_{mh}^r \in \mathcal{R}_0^+. \end{cases}$$

The object-cluster distance $\mathrm{dist}(\mathbf{x}_i, C_l)$ is defined as

$$\mathrm{dist}(\mathbf{x}_i, C_l) = \sum_{r=1}^{d} \mathrm{dist}^r(\mathbf{x}_i, C_l), \tag{3}$$

and $\mathrm{dist}^r(\mathbf{x}_i, C_l)$ is the object-cluster distance measured on attribute A^r. If $x_i^r = o_m^r$, $\mathrm{dist}^r(\mathbf{x}_i, C_l)$ can be written as

$$\mathrm{dist}^r(\mathbf{x}_i, C_l) = \sum_{h=1}^{v^r} w_{mh}^r \mathrm{dist}^r(o_m^r, o_h^r) p_{lh}^r, \tag{4}$$

and the intra-attribute distance $\mathrm{dist}^r(o_m^r, o_h^r)$ is defined as

$$\mathrm{dist}^r(o_m^r, o_h^r) = \frac{1}{d} \sum_{s=1}^{d} \mathrm{dist}^{rs}(o_m^r, o_h^r), \tag{5}$$

where the superscript "rs" of $\mathrm{dist}^{rs}(o_m^r, o_h^r)$ indicates that this is the intra-attribute distance between A^r's possible values with respect to A^s. We define distance in the form of Eq. (5) in order to exploit context information provided by interdependent attributes for distance measurement as most categorical data distance measures do [17], [18], [19], [21], [22], [31]. The exact definition of $\mathrm{dist}^{rs}(o_m^r, o_h^r)$ will be given in Section 4.3.

Similar to most existing k-modes-type algorithms, the minimization problem of Eq. (2) can be solved by iteratively computing one variable and fixing the others. Since the values of P are completely dependent on the values of $\mathbf{Q}$, we can iteratively solve the following two problems:

- **P.1**: Fix $W = \hat{W}$ and $P = \hat{P}$, solve the reduced problem $Z(\mathbf{Q}, \hat{P}, \hat{W})$, update P according to $\mathbf{Q}$;
- **P.2**: Fix $\mathbf{Q} = \hat{\mathbf{Q}}$ and $P = \hat{P}$, solve the reduced problem $Z(\hat{\mathbf{Q}}, \hat{P}, W)$.

4 Homogeneous Distance Measurement

For cluster analysis, the adopted distance measure usually dominates clustering performance. In this section, we study the differences and commonalities of ordinal and nominal attributes, and then propose a homogeneous intra-attribute distance definition for them.

4.1 Attribute Structure

We first discuss the difference between ordinal and nominal attributes. As shown in Fig. 1, if we treat the intra-attribute possible values as nodes connected by edges, since nodes of an ordinal attribute are naturally ordered, one node cannot be reached along the edges from another non-adjacent node without crossing its adjacent node, while for a nominal attribute, a node can be directly reached along an edge from any node without involving such "crossing". We construct graphs for studying the heterogeneity between ordinal and

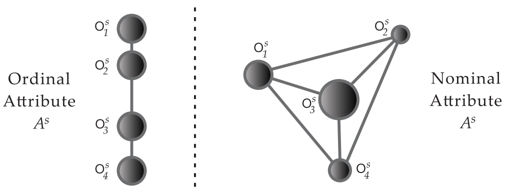


Fig. 1. Structural difference between ordinal and nominal attributes from the perspective of graph. The black nodes stand for possible values and the edges reflect the spatial relationships among possible values.

nominal attributes because graph is effective in modeling complex relationships between nodes [41], [42], and has been successfully applied to different machine learning tasks, such as sketch synthesis [43], item recommendation [44], and object retrieval [45]. It can be seen according to Fig. 1 that the structure of ordinal attribute is line-like while the structure of nominal attribute is net-like. These structures are consistent with the relationships among intra-attribute possible values of ordinal and nominal attributes from the practical point of view. For example, if we compare two choices, i.e., bad and very-good, of the review result regarding the novelty of a manuscript with the five choices $\{\text{very-good}, \text{good}, \text{neural}, \text{bad}, \text{very-bad}\}$. We will not skip neutral and good to directly compare bad and very-good, because all the choices are clearly ordered. In contrast, if we compare two choices that belong to a choice set without such order relationship, we will directly compare the two choices without involving the other choices. It is obvious that the structures of ordinal and nominal attributes are heterogeneous, which makes their intra-attribute distances difficult to be defined in a homogeneous way.

4.2 Homogeneous Learning

For mixed categorical data, there are two cases for Eq. (5): 1) $A^s \in A^{(\mathrm{ord})}$, and 2) $A^s \in A^{(\mathrm{nom})}$. Since the possible values of an ordinal attribute represent the different degrees of a concept while the possible values of a nominal attribute represent different concepts, we convert possible values of a nominal attribute into ordinal attributes as shown in Fig. 2 so that the original nominal attribute becomes homogeneous with ordinal attributes. Specifically, for $A^s \in A^{(\mathrm{nom})}$ with v^s possible values, we convert it into a set of v^s ordinal attributes

$$B^s = \{A^g | g \in N_O^s\}, \tag{6}$$

where A^g is a newly generated ordinal attribute corresponding to the possible value o_g^s of A^s. Each A^g has two possible values $o_1^g = o_g^s$ and $o_2^g = \neg o_g^s$ where $v^g = 2$ and $o_1^g \succ o_2^g$. Here, $\neg o_g^s$ stands for all the possible values of A^s except o_g^s. Each A^g can be viewed as a special case of ordinal attribute, in which there are only two possible values indicating two extreme degrees, i.e., "is o_g^s" and "is not o_g^s". In this way, all the nominal attributes can be converted into ordinal attributes, and the intra-attribute distances can then be measured according to the same type of information provided by the attributes.

4.3 Design of Proposed Distance Metric

The distance between two possible values (e.g., o_m^r and o_h^r of attribute A^r) with respect to another attribute (e.g., A^s) is defined in this part. Before presenting the details of this distance definition, let us first define the conditional probability distribution of an attribute (e.g., A^s) with respect to a possible value (e.g., o_m^r), which can be written as

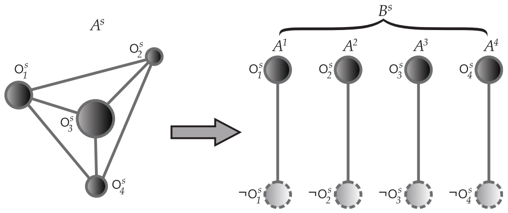


Fig. 2. Converting a nominal attribute A^s into a set of ordinal attributes B^s: Each nominal possible value o_g^s is converted into an ordinal attribute A^g with two ordered possible values o_g^s and $\neg o_g^s$.

$$\mathbf{u}_m^{rs} = [p(o_1^s|o_m^r), p(o_2^s|o_m^r), \ldots, p(o_{v^s}^s|o_m^r)]^\top, \tag{7}$$

where $p(o_g^s|o_m^r)$ is the conditional probability of o_g^s with respect to o_m^r following Bayes' theorem:

$$p(o_g^s|o_m^r) = \frac{\mathrm{card}(X_g^s \cap X_m^r)}{\mathrm{card}(X_m^r)}. \tag{8}$$

Here, $X_g^s = \{\mathbf{x}_i | x_i^s = o_g^s, i \in N_X\}$ is a subset of X with the sth values of all its objects equal to o_g^s, and the function $\mathrm{card}(\cdot)$ counts the cardinality of a set. Then, we define the distance between two possible values (e.g., o_m^r and o_h^r of attribute A^r) with respect to another attribute (e.g., A^s) as follows:

$$\mathrm{dist}^{rs}(o_m^r, o_h^r) = \begin{cases} \psi(\mathbf{u}_m^{rs}, \mathbf{u}_h^{rs}), & A^s \in A^{(\mathrm{ord})} \\ \frac{1}{v^s}\sum_{g=1}^{v^s} \psi(\mathbf{u}_m^{rg}, \mathbf{u}_h^{rg}), & A^s \in A^{(\mathrm{nom})}, \end{cases} \tag{9}$$

where $\psi(\cdot,\cdot)$ computes the distance between two probability distributions. For the nominal case (i.e., $A^s \in A^{(\mathrm{nom})}$), the distance with respect to A^s is computed as the mean of the distances with respect to the ordinal attributes $A^g \in B^s$ that are converted from A^s as shown in Fig. 2. See Eq. (6) and corresponding discussions in Section 4.2 for more details. As both A^s in the ordinal case and A^g in the nominal case are ordinal attributes, we only need to discuss how to define $\psi(\cdot,\cdot)$ in the ordinal case.

In the literature, although the distance between two probability distributions is commonly computed in the form of l_1-norm (i.e., $||\mathbf{u}_m^{rs} - \mathbf{u}_h^{rs}||_1$) or l_2-norm (i.e., $||\mathbf{u}_m^{rs} - \mathbf{u}_h^{rs}||_2$), they are not suitable here because they cannot preserve order relationship among possible values of an ordinal attribute. For example, given $\mathbf{u}_1 = [1,0,0,0]^\top$, $\mathbf{u}_2 = [0,1,0,0]^\top$, $\mathbf{u}_3 = [0,0,0,1]^\top$, we have $||\mathbf{u}_1 - \mathbf{u}_2||_1 = ||\mathbf{u}_1 - \mathbf{u}_3||_1$. However, if $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are obtained from an ordinal attribute, it is obvious that $\mathbf{u}_1$ and $\mathbf{u}_2$ are more similar than $\mathbf{u}_1$ and $\mathbf{u}_3$, because the two possible values that rank 1st and 2nd are more similar than the two possible values that rank 1st and 4th. To preserve the order relationship, we define $\psi(\cdot,\cdot)$ as the cost of transforming a probability distribution into another according to the structure of ordinal attribute shown in Fig. 1, and $\psi(\mathbf{u}_m^{rs}, \mathbf{u}_h^{rs})$ can be written as

$$\psi(\mathbf{u}_m^{rs}, \mathbf{u}_h^{rs}) = \frac{\sum_{t=1}^{v^s-1} |\sum_{g=1}^{t}\Big(p(o_g^s|o_m^r) - p(o_g^s|o_h^r)\Big)|}{v^s - 1}. \tag{10}$$

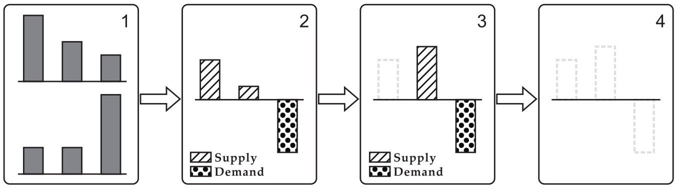


Fig. 3. Computation process of Eq. (10). In step 1, we have $\mathbf{u}_m^{rs} = [0.5, 0.3, 0.2]^\top$, i.e., the upper histogram, and $\mathbf{u}_h^{rs} = [0.2, 0.2, 0.6]^\top$, i.e., the lower histogram. To transform $\mathbf{u}_m^{rs}$ into $\mathbf{u}_h^{rs}$, we first subtract them and obtain the histogram $[0.3, 0.1, -0.4]^\top$ in step 2. The slash-filled bins indicate supplies, and the dot-filled bin indicates demand. Then, 0.3 supply at the first place is moved to the second place with 0.1 supply, the moving cost is (0.3×1)/2 = 0.15. In step 3, the total 0.4 supply at the second place is moved to the third place with 0.4 demand, the moving cost is (0.4×1)/2 = 0.2. Since the supply and demand exactly offset each other after step 3, the transforming is completed in step 4, and the total transforming cost is 0.15 + 0.2 = 0.35.

The distance defined in Eq. (10) computes the minimum moving cost for transforming $\mathbf{u}_m^{rs}$ into $\mathbf{u}_h^{rs}$ (or $\mathbf{u}_h^{rs}$ into $\mathbf{u}_m^{rs}$), where $|\sum_{g=1}^{t}(p(o_g^s|o_m^r) - p(o_g^s|o_h^r))|$ in Eq. (10) is the total 'supplies' or 'demands' at location o_t^s that should be moved to locations $o_{t+1}^s, o_{t+2}^s, \ldots, o_{v^s}^s$ for offsetting. During the above computation, the 'moving distance' between adjacent values is 1 because the prior knowledge we have is that the rank of a possible value is different from its adjacent possible value(s) by 1. A toy example shown in Fig. 3 intuitively illustrates the computation process.

Eq. (10) elaborately reflects the distance between two probability distributions obtained from an ordinal attribute, and we discuss it in detail below:

- According to our design, 'supplies' and 'demands' are moved strictly according to the structure of ordinal attribute as shown in Fig. 1. It turns out that the order relationship among the possible values is taken into account in computing the distance between two distributions by Eq. (10).
- It is intuitive that two more different distributions yield more 'supplies' and 'demands' for moving, and thus result in a larger distance computed by Eq. (10), which is consistent with the general distance definitions like Manhattan and euclidean distance.
- In terms of the form, Eq. (10) can be viewed as a special case of the Earth Movers' Distance (EMD) [46], [47], [48], as Eq. (10) only permits 'moving' between adjacent bins of histograms. However, Eq. (10) is designed under the guidance of the proposed graph structure shown in Fig. 1, which is very different from the motivation and principle for designing EMD.

Eqs. (9) and (10) have defined the distance between two possible values with respect to an attribute. Then, according to the structures of nominal and ordinal attributes studied in Section 4.1, we define the overall distance between two possible values by combining their distances with respect to each attribute as follows:

$$\mathrm{dist}^r(o_m^r, o_h^r) = \begin{cases} \frac{1}{d}\sum_{s=1}^{d}\sum_{t=\min(m,h)}^{\max(m,h)-1}\mathrm{dist}^{rs}(o_t^r, o_{t+1}^r) & \\ & A^r \in A^{(\mathrm{ord})} \\ \frac{1}{d}\sum_{s=1}^{d}\mathrm{dist}^{rs}(o_m^r, o_h^r), & A^r \in A^{(\mathrm{nom})}. \end{cases} \tag{11}$$

Based on Eq. (11), the distance between two data objects $\mathbf{x}_i$ and $\mathbf{x}_j$ with their rth values denoted as $x_i^r = o_m^r$ and $x_j^r = o_h^r$, respectively, can be written as

$$\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{d}\sum_{r=1}^{d}\mathrm{dist}^r(o_m^r, o_h^r). \tag{12}$$

Theorem 1. *Distance measure defined in Eqs. (9), (10), (11), and (12) is a distance metric.*

Proof According to Eqs. (9), (10), and (11), it is clear that the defined intra-attribute distance satisfies the following properties for any $m, h, t \in N_O^r$ and $r \in N_A$:

1) $\mathrm{dist}^r(o_m^r, o_h^r) \geq 0$;
2) $o_m^r = o_h^r \Leftrightarrow \mathrm{dist}^r(o_m^r, o_h^r) = 0$;
3) $\mathrm{dist}^r(o_m^r, o_h^r) = \mathrm{dist}^r(o_h^r, o_m^r)$;
4) $\mathrm{dist}^r(o_m^r, o_h^r) \leq \mathrm{dist}^r(o_m^r, o_t^r) + \mathrm{dist}^r(o_t^r, o_h^r)$.

According to Eq. (12), it is clear that the following properties hold for any $i, j, l \in N_X$:

1) $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \geq 0$;
2) $\mathbf{x}_i = \mathbf{x}_j \Leftrightarrow \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) = 0$;
3) $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{dist}(\mathbf{x}_j, \mathbf{x}_i)$;
4) $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_l) + \mathrm{dist}(\mathbf{x}_l, \mathbf{x}_j)$.

The defined distance measure satisfies all the distance metric properties. □

In practice, a set of distance matrices, i.e., $D = \{\mathbf{D}^r | r \in N_A\}$ where $\mathbf{D}^r$ is a $v^r \times v^r$ symmetric matrix storing intra-attribute distances of A^r, can be computed before clustering. The (m,h)th entry of $\mathbf{D}^r$ is denoted as d_{mh}^r where $d_{mh}^r = \mathrm{dist}^r(o_m^r, o_h^r)$. With D, distances can be directly read off during clustering.

Theorem 2. *Time complexity for computing the distance matrices D is $O(nd^2 + d^2V^3)$.*

Proof. Conditional probability distribution $\mathbf{u}_m^{rs}$ with $r, s \in N_A$ and $m \in N_O^r$ should be obtained before distance computation. For each $\mathbf{u}_m^{rs}$, o_m^r's corresponding values on A^s should be scanned once with time complexity $O(\mathrm{card}(X_m^r))$, and for all the $\mathbf{u}_m^{rs}$ with $m \in N_O^r$, the scan is with time complexity $O(n)$. Such scan should be performed for each pair of attributes, and thus the time complexity is $O(nd^2)$.

Given a pair of possible values o_m^r and o_h^r with $m, h \in N_O^r$ and $r \in N_A$, the time complexity for computing the distance between them based on the known $\mathbf{u}_m^{rs}$ and $\mathbf{u}_h^{rs}$ is $O(V)$ in both the two cases of Eq. (9). Note that $V = \max(v^1, v^2, \ldots, v^d)$ is the maximum number of possible values among all the attributes, which is adopted to simplify the time complexity analysis. To obtain $\mathbf{D}^r$, the time complexity for computing the $V(V-1)/2$ intra-attribute distances is $O(dV^3)$ in the case $A^r \in A^{(\mathrm{nom})}$ of Eq. (11). In the case $A^r \in A^{(\mathrm{ord})}$ of Eq. (11), distance between possible values with order difference 1 can be computed first, and then the distance between possible values with order difference 2, 3, ..., $V-1$ can be successively computed based on the distances computed in the previous step. Therefore, the time complexity for computing the $V(V-1)/2$ intra-attribute distances is also $O(dV^3)$ in the case $A^r \in A^{(\mathrm{ord})}$ of Eq. (11). The time complexity for computing a total of d distance matrices $\mathbf{D}^r$ is thus $O(d^2V^3)$.

Hence, the overall time complexity for obtaining D is $O(nd^2 + d^2V^3)$. □

5 Clustering Based on Intra-Attribute Distance Weighting

Often, separately treating the cross-coupled distance defining and data clustering results in a suboptimal solution. This section will therefore propose a learning mechanism that adjusts the defined intra-attribute distances to suit certain clustering tasks. We have constructed graph-like structure for the intra-attribute possible values to define their distances in Section 4.3, and will learn the weights of the distances in an iterative way with data clustering in the following subsections. Before introducing the details of our algorithm, let us conceptually discuss the existing methods whose learning paradigms are intuitively similar to ours.

Several clustering of bandits algorithms [44], [49], [50] have been proposed to construct graph for the objects (i.e., users in their application scenarios) and dynamically perform graph clustering according to the item preference of users over time. Further, the one in [51] captures the collaborative effects of the users, and the one in [52] captures the bi-collaborative effects between users and items by iteratively partitioning user and item graphs. The commonality of our method and the above-mentioned ones is that they all iteratively learn (1) the relationship between objects and (2) certain predictions for the objects. The differences are: (1) we construct graphs only for studying the distance definition between possible values, while most of the above-mentioned methods construct graphs for objects and cutting the graphs for object partitioning, (2) we optimize the prediction from objects to object clusters, while they optimize the prediction of items to be recommended to users. Although the above-mentioned methods are not solving the same type of problem as ours, their paradigms can provide inspiration for applying our method in more complex environments in the future, for example, in on-line or distributed situations. In the following, we will elaborate how to solve the two problems **P. 1** and **P. 2** stated in Section 3, and present the clustering algorithm, together with the time-complexity analysis.

5.1 Update Q As Given W and P

The process of solving problem $\mathbf{P}.\mathbf{1}$ is to obtain a data partition according to a certain distance measure and cluster representation, which actually adopts the same basic idea as most k-modes-type algorithms. The difference is that we use the statistical information P (defined in Section 3) instead of cluster modes in representing the clusters, which ensures the extraction of more rich information for learning distance weights W in solving problem $\mathbf{P}.\mathbf{2}$. Specifically, the details of solving $\mathbf{P}.\mathbf{1}$ are presented as follows. According to the objective function defined in Eq. (2), $\mathbf{P}.\mathbf{1}$ is solved by fixing $W = \hat{W}$ and $P = \hat{P}$, and computing $\mathbf{Q}$. Given distance matrices D, $\mathbf{Q}$ is computed by

$$q_{il} = \begin{cases} 1, & \text{if } l = \arg\min_y \mathrm{dist}(\mathbf{x}_i, C_y) \\ & \qquad = \arg\min_y \sum_{r=1}^{d} \sum_{h=1}^{v^r} w_{mh}^{\hat{r}} d_{mh}^{r} p_{yh}^{\hat{r}}\,, \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

for $\mathbf{x}_i$ with $x_i^r = o_m^r$. Since we represent the clusters using their probability distributions by P instead of using cluster modes, the form of the solution in Eq. (13) is different form the conventional k-modes-type algorithms. Thus, solution to $\mathbf{P}.\mathbf{1}$ is also rigorously given in Theorem 3.

Theorem 3. *Let W and P be fixed, $Z(\mathbf{Q}, \hat{P}, \hat{W})$ is minimized iff $\mathbf{Q}$ is computed utilizing Eq. (13).*

Proof. For any given $W = \hat{W}$ and $P = \hat{P}$, all the inner sums of the quantity

$$Z(\mathbf{Q}, \hat{P}, \hat{W}) = \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il} \mathrm{dist}(\mathbf{x}_i, C_l),$$

are nonnegative and independent. Let $o_m^r = x_i^r$, we can write the inner sum contributed by $\mathbf{x}_i$ as

$$z_i = \sum_{l=1}^{k} q_{il} \mathrm{dist}(\mathbf{x}_i, C_l) = \sum_{l=1}^{k} q_{il} \sum_{r=1}^{d} \sum_{h=1}^{v^r} w_{mh}^{\hat{r}} d_{mh}^{r} p_{lh}^{\hat{r}}.$$

Let $z_{il} = \sum_{r=1}^{d} \sum_{h=1}^{v^r} w_{mh}^{\hat{r}} d_{mh}^{r} p_{lh}^{\hat{r}}$, which is the inner sum contributed by $\mathbf{x}_i$ in C_l. We then obtain

$$z_i = \sum_{l=1}^{k} q_{il} z_{il}.$$

Since $\sum_{l=1}^{k} q_{il} = 1$ and $q_{il} \in \{0, 1\}$, it is clear that z_i is minimized iff the minimum z_{il} is assigned with $q_{il} = 1$ where l is determined by

$$l = \arg\min_y z_{iy} = \arg\min_y \sum_{r=1}^{d} \sum_{h=1}^{v^r} w_{mh}^{\hat{r}} d_{mh}^{r} p_{yh}^{\hat{r}},$$

and the other z_{il}s are assigned with $q_{il} = 0$. The result follows. □

We have presented the solution of updating $\mathbf{Q}$. Each time a new $\mathbf{Q}$ is obtained, the cluster representation P is updated accordingly, and such process is iterated until convergence.

5.2 Update W As Given Q and P

In Section 5.1, we have presented the solution of $\mathbf{P}.\mathbf{1}$. Then, $\mathbf{P}.\mathbf{2}$ should be solved based on the present $\mathbf{Q}$ and P to learn distance weights W. In this part, a novel learning scheme is designed by mining the latent interaction between data partition and intra-attribute distances, so as to seek for more appropriate data partition in the next iteration based on the defined intra-attribute distances and the newly learned W. The details of solving $\mathbf{P}.\mathbf{2}$ are presented as follows. Given fixed $\hat{\mathbf{Q}}$ and $\hat{P}$, the objective function defined by Eq. (2) can be written as

$$\begin{aligned} Z(\hat{\mathbf{Q}}, \hat{P}, W) &= \sum_{i=1}^{n} \sum_{l=1}^{k} \hat{q}_{il} \mathrm{dist}(\mathbf{x}_i, C_l) \\ &= \sum_{i=1}^{n} \sum_{l=1}^{k} \hat{q}_{il} \sum_{r=1}^{d} \sum_{h=1}^{v^r} w_{mh}^{r} d_{mh}^{r} p_{lh}^{\hat{r}} \\ &= \sum_{r=1}^{d} \sum_{m=1}^{v^r} \sum_{h=1}^{v^r} w_{mh}^{r} d_{mh}^{r} \sum_{l=1}^{k} \frac{f_{lm}^{r} f_{lh}^{r}}{f_l}, \end{aligned} \tag{14}$$

where $f^r_{lm} = \mathrm{card}(X^r_m \cap X_{C_l})$ and $f^r_{lh} = \mathrm{card}(X^r_h \cap X_{C_l})$ are the total number of objects in C_l with their rth values equal to o^r_m and o^r_h, respectively, $f_l = \mathrm{card}(X_{C_l})$ is the number of objects in C_l, and we have $f^r_{lh}/f_l = \hat{p}^r_{lh}$. In most k-modes-type algorithm with attribute weighting mechanism [7], [8], [10], Lagrangian multiplier is used to convert the constrained weights computation problem into an unconstrained problem so that the optimal attribute weights can be computed directly in each iteration. However, solving our intra-attribute distance weighting problem in this way may encounter two awkward issues:

- *Frequency Effect:* For attribute weighting, each attribute has the identical number of values, which is the basis for success in making the computed weights comparable. However, the occurrence frequencies of intra-attribute distances (i.e., $\sum_{l=1}^{k} f^r_{lm} f^r_{lh}$ of d^r_{mh}) are usually different from each other, which makes the computed weights of intra-attribute distances incomparable.
- *Co-occurrence Sparsity:* It is common for a real categorical data set that an intra-attribute distance (i.e., d^r_{mh}) never occur in a cluster, so that no statistical information is provided for the computation of its corresponding weight w^r_{mh}. If we set such weights to 0, the problem still cannot be fixed because there are many such weights preventing the algorithm from convergence.

We propose a novel intra-attribute distance weight updating scheme to circumvent the above-discussed issues. In general, a larger d^r_{mh} indicates that the two corresponding possible values o^r_m and o^r_h are more dissimilar. That is, d^r_{mh} is expected to contribute more in partitioning the objects in $X^r_m = \{\mathbf{x}_i | x^r_i = o^r_m, i \in N_X\}$ and the objects in $X^r_h = \{\mathbf{x}_i | x^r_i = o^r_h, i \in N_X\}$ into different clusters. Thus, given a data partition $\hat{\mathbf{Q}}$, if d^r_{mh} is larger but more objects in X^r_m and X^r_h are assigned into the same cluster, it is indicated that d^r_{mh} does not contribute in partitioning the objects X^r_m and X^r_h into different clusters as expected. Accordingly, the weight of d^r_{mh} should be estimated as its expectation in reducing $Z(\hat{\mathbf{Q}}, \hat{P}, W)$, and we have

$$w^r_{mh} \propto \left[d^r_{mh} \sum_{l=1}^{k} \frac{f^r_{lm} f^r_{lh}}{f^r_m f^r_h} \right]^{-1}, \tag{15}$$

where $f^r_m = \mathrm{card}(X^r_m) = \sum_{l=1}^{k} f^r_{lm}$ and $f^r_h = \mathrm{card}(X^r_h) = \sum_{l=1}^{k} f^r_{lh}$ are the intra-cluster occurrence frequencies of o^r_m and o^r_h, respectively. The term $(f^r_{lm} f^r_{lh})/(f^r_m f^r_h)$ quantifies the occurrence of d^r_{mh} in C_l as a joint occurrence probability of o^r_m and o^r_h in C_l, which avoids the *Frequency Effect*. If we directly update W by $w^r_{mh} = [d^r_{mh} \sum_{l=1}^{k} (f^r_{lm} f^r_{lh})/(f^r_m f^r_h)]^{-1}$ according to Eq. (15), the *Co-occurrence Sparsity* issue may make the values of different weights vary greatly in the interval $[1/d^r_{mh}, \infty)$, which may cause non-convergence. Thus, we discuss how to novelly circumvent the *Co-occurrence Sparsity* issue in the following.

Lemma 1. *Given an arbitrary partition* $\mathbf{Q}$ *of data set* S*, sum of the intra-cluster distance* $E^{r(\mathrm{intra})}_{mh}$ *and inter-cluster distance* $E^{r(\mathrm{inter})}_{mh}$ *contributed by* d^r_{mh} *is a constant.*

Proof. We first note that $E^{r(\mathrm{intra})}_{mh} = d^r_{mh} \sum_{l=1}^{k} f^r_{lm} f^r_{lh}$ and $E^{r(\mathrm{inter})}_{mh} = d^r_{mh} \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})$. Let $E^{r(\mathrm{total})}_{mh} = E^{r(\mathrm{intra})}_{mh} + E^{r(\mathrm{inter})}_{mh}$, we have

$$\begin{aligned}
E^{r(\mathrm{total})}_{mh} &= d^r_{mh} \left(\sum_{l=1}^{k} f^r_{lm} f^r_{lh} + \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} \left(f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh} \right) \right) \\
&= d^r_{mh} \left(\sum_{s=1}^{k} \sum_{u=s} f^r_{sm} f^r_{uh} \right. \\
&\quad \left. + \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} f^r_{sm} f^r_{uh} + \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} f^r_{um} f^r_{sh} \right) \\
&= d^r_{mh} \sum_{s=1}^{k} \sum_{u=1}^{k} f^r_{sm} f^r_{uh} = d^r_{mh} \sum_{s=1}^{k} f^r_{sm} \sum_{u=1}^{k} f^r_{uh} \\
&= d^r_{mh} f^r_m f^r_h .
\end{aligned}$$

Since d^r_{mh}, f^r_m, and f^r_h are constants for a given data set S, it is clear that $E^{r(\mathrm{total})}_{mh}$ is a constant. The result follows. □

Based on Lemma 1, Eq. (15) can be transformed to avoid the *Co-occurrence Sparsity* issue.

Lemma 2. *Given Eq. (15),* $w^r_{mh} \propto d^r_{mh} \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})/(f^r_m f^r_h)$ *holds when* $\exists\, l \in N_C$ *so that* $f^r_{lm} f^r_{lh} \neq 0$.

Proof. Let $H^r_{mh} = d^r_{mh} \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})/(f^r_m f^r_h)$.

We first prove that $H^r_{mh} < E^{r(\mathrm{total})}_{mh}/(f^r_m f^r_h)$. According to the proof of Lemma 1, we derive

$$\frac{E^{r(\mathrm{total})}_{mh}}{f^r_m f^r_h} - H^r_{mh} = \frac{d^r_{mh} \sum_{l=1}^{k} f^r_{lm} f^r_{lh}}{f^r_m f^r_h}. \tag{16}$$

Since $\exists\, l \in N_C$ so that $f^r_{lm} f^r_{lh} \neq 0$, and f^r_{lm} and f^r_{lh} are nonnegative integers, we have $\sum_{l=1}^{k} f^r_{lm} f^r_{lh} > 0$; Since f^r_m and f^r_h are positive constants and $d^r_{mh} > 0$ for two different possible values o^r_m and o^r_h, we then have

$$\frac{d^r_{mh} \sum_{l=1}^{k} f^r_{lm} f^r_{lh}}{f^r_m f^r_h} > 0 \;\Rightarrow\; H^r_{mh} < \frac{E^{r(\mathrm{total})}_{mh}}{f^r_m f^r_h}.$$

From Eqs. (15) and (16), we derive

$$w^r_{mh} \propto \frac{1}{\frac{E^{r(\mathrm{total})}_{mh}}{f^r_m f^r_h} - H^r_{mh}}.$$

Since we have proved $H^r_{mh} < E^{r(\mathrm{total})}_{mh}/(f^r_m f^r_h)$, and $E^{r(\mathrm{total})}_{mh}/(f^r_m f^r_h)$ is a constant, it is clear that the value of w^r_{mh} is proportional to the value of H^r_{mh}, which can be written as

$$w^r_{mh} \propto \frac{d^r_{mh} \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})}{f^r_m f^r_h}. \tag{17}$$

The result follows. □

According to Lemma 2, the weights of intra-attribute distances are updated by

$$w^{r(\mathrm{new})}_{mh} = \frac{d^r_{mh} \sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})}{f^r_m f^r_h}. \tag{18}$$

Eq. (18) is obtained with restriction $\exists\, l \in N_C$ so that $f^r_{lm} f^r_{lh} \neq 0$. We also demonstrate that when $\forall\, l \in N_C$, $f^r_{lm}, f^r_{lh} = 0$, Eq. (18) is still meaningful. $\forall\ l \in N_C$, $f^r_{lm}, f^r_{lh} = 0$ indicates that the objects in X^r_m and the objects in X^r_h never appear in the same cluster, which means that the contribution of d^r_{mh} in partitioning the objects in X^r_m and the objects in X^r_h into different clusters reaches the maximum, i.e., $\sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})/(f^r_m f^r_h) = 1$. Since the value of $\sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})/f^r_m f^r_h$ is in the interval [0,1], it is clear that weights updating utilizing Eq. (18) will not be influenced by the *Co-occurrence Sparsity* issue. We also use soft-max, i.e., $w^r_{mh} = w^{r(\text{new})}_{mh} / \sum_{s=1}^{d} \sum_{g=1}^{v^s-1} \sum_{t=g+1}^{v^s} w^{s(\text{new})}_{gt}$, to make the updated weights satisfy $\sum_{r=1}^{d} \sum_{m=1}^{v^r-1} \sum_{h=m+1}^{v^r} w^r_{mh} = 1$.

Advantages of the proposed weights updating scheme are summarized below:

- *Frequency Dominance* issue is avoided.
- *Co-occurrence Sparsity* issue is novelly circumvented.
- It is parameter-free, and the clustering algorithm based on it (see Section 5.3) always converge quickly, which has been illustrated in Section 6.

5.3 Complete Clustering Algorithm

The complete clustering algorithm called HD-NDW integrates the solutions of $\mathbf{P}.\mathbf{1}$ and the Novel Distance Weighting (NDW) mechanism for solving $\mathbf{P}.\mathbf{2}$. As described in Algorithm 1, it iteratively updates the data partitions and weights of the distances defined by the Homogeneous Distance (HD) metric for data partitioning. More specifically, $\mathbf{Step\,1}$ acts as a complete clustering algorithm that learns a data partition, which provides information for updating the weights of distances in $\mathbf{Step\,2}$. This is why we put $\mathbf{Step\,1}$ before $\mathbf{Step\,2}$ in HD-NDW. After reasonable distance weights are learned according to the data partition, the weights are fed back to $\mathbf{Step\,1}$ for learning more appropriate data partition, and such procedures iterate until convergence. Time complexity of HD-NDW is analyzed in Theorem 4.

Algorithm 1. HD-NDW Clustering Algorithm

```
Input: Data set $S$, number $k$ of clusters, distance matrices $D$.
Output: Partition $\mathbf{Q}$.
Step 0: Initialize the time-step by $\tau = 0$; Initialize $P^{(\tau)}$ and $W^{(\tau)}$;
Step 1: Fix $W^{(\tau)}$ and $P^{(\tau)}$, iteratively update $\mathbf{Q}^{(\tau)}$ by Eq. (13) and
    update $P^{(\tau)}$ according to $\mathbf{Q}^{(\tau)}$ until convergence, obtain
    $\mathbf{Q}^{(\tau+1)}$ and $P^{(\tau+1)}$; If $\mathbf{Q}^{(\tau+1)} \neq \mathbf{Q}^{(\tau)}$, go to Step 2; Otherwise,
    stop and Output $\mathbf{Q}^{(\tau)}$.
Step 2: Fix $\mathbf{Q}^{(\tau+1)}$ and $P^{(\tau+1)}$, update $W^{(\tau)}$ by Eq. (18), obtain
    $W^{(\tau+1)}$; Update the time-step by $\tau = \tau + 1$, go to Step 1;
```

Theorem 4. *Time complexity of Algorithm 1 is* $O(E(kdVnI + dn + dV^2k))$, *supposing* $\mathbf{Step\,1}$ *needs* I *iterations to converge, and the loop of* $\mathbf{Step\,1}$ *and* $\mathbf{2}$ *needs* E *iterations to converge.*

Proof. In $\mathbf{Step\,1}$, time complexity for computing the values of a row of $\mathbf{Q}^{(\tau)}$ is $O(kdV)$ because there are k clusters to be considered, and for each cluster, the distance is computed based on the d intra-attribute distances stored in D, and each attribute has a maximum of V possible values. See the proof of Theorem 3 for more details of the computing of $\mathbf{Q}^{(\tau)}$. Since there are n rows in $\mathbf{Q}^{(\tau)}$ and $\mathbf{Step\,1}$ repeats I times, the total time complexity of $\mathbf{Step\,1}$ is $O(kdVnI)$.

According to the proof of Lemma 1, the term $\sum_{s=1}^{k-1} \sum_{u=s+1}^{k} (f^r_{sm} f^r_{uh} + f^r_{um} f^r_{sh})/(f^r_m f^r_h)$ in Eq. (18) can be directly computed by $1 - \sum_{l=1}^{k} f^r_{lm} f^r_{lh}/(f^r_m f^r_h)$. Before the computation, we should first obtain the set of occurrence frequency matrices $F = \{\mathbf{F}^1, \mathbf{F}^2, \ldots, \mathbf{F}^d\}$ where $\mathbf{F}^r$ is a $k \times v^r$ matrix storing the occurrence frequencies of A^r's possible values in each cluster, and the (l,m)th entry of $\mathbf{F}^r$ is f^r_{lm}. To obtain F, the d values of each data object $\mathbf{x}_i$ should be scanned once according to the corresponding $q_{il} = 1$ in $\mathbf{Q}$. Since there are n objects in total, the time complexity for obtaining F is $O(dn)$. It is therefore clear that the time complexity for computing the $dV(V-1)/2$ weights according to each of the k clusters using Eq. (18) is $O(dn + dV^2k)$ in $\mathbf{Step\,2}$.

Since the loop of $\mathbf{Step\,1}$ and $\mathbf{2}$ repeats E times, the time complexity of HD-NDW is $O(E(kdVnI + dn + dV^2k))$. □

6 EXPERIMENTS

We conduct a series of experiments on various benchmark and real data sets to evaluate the proposed clustering method. We first describe the experimental settings. Then, we demonstrate and discuss the experimental results.

6.1 Experimental Settings

6.1.1 Experimental Design

Five experiments are designed as follows:

- *Clustering Performance of HD-NDW.* We compare HD-NDW with various clustering algorithms on mixed, ordinal, and nominal categorical data sets to illustrate the superiority of HD-NDW.
- *Effectiveness of HD.* HD is a core component of HD-NDW. We compare HD and various distance measures by combining them with the simplest k-modes clustering algorithm to illustrate the effectiveness of HD.
- *Effectiveness of NDW.* NDW is also a core component of HD-NDW. We compare HD-NDW and its non-weighting version to prove the effectiveness NDW.
- *Convergence Evaluation.* Convergence curves of HD-NDW on various data sets are demonstrated to illustrate its effectiveness and fast convergence.
- *Computational Efficiency Evaluation.* We compare the execution time of various clustering methods on synthetic data sets to illustrate the efficiency of HD-NDW.

For all the experiments, the number k of the clusters is set at the true number k^* of the clusters according to the data label. We run all the experiments 50 times and report the average results.

6.1.2 Validity Indices

We select the commonly used Adjusted Rand Index (ARI) [53] because it is powerful in discriminating clustering performance [54], [55]. Normalized Mutual Information (NMI) [22], [56] is selected to evaluate clustering performance from the perspective of information theory [57]. To make the evaluation comprehensive, the traditional Clustering Accuracy (CA) [58], [59] is also selected. NMI and CA are in the interval [0,1] and ARI is in the interval $[-1, 1]$. For all these selected validity indices, a higher value indicates a better clustering performance. We also adopt Wilcoxon signed-rank test and Bonferroni-Dunn test [60] to evaluate the statistical significance of the difference between clustering performance of different methods. In addition, we compute the averaged Intra- and Inter-Cluster Distance (ICD for short) [19] to intuitively demonstrate the cluster discrimination ability of different methods.

6.1.3 Counterpart Selection

The most representative partitional clustering algorithms are selected as counterparts for the experiments. We select k-modes (KMD) [36] because it is the most conventional one. We select Entropy-based Categorical data Clustering (ECC) [39] because it is conventional and representative among the entropy-based clustering algorithms. We also select attribute Weighting k-modes (WKM) [7], Mixed-attribute Weighting k-modes (MWKM) [10], and attribute Weighting and Object-cluster-similarity-based Clustering (WOC) [11] algorithms as another three counterparts. WKM and MWKM are two representative algorithms in the attribute-weighting clustering stream, and WOC is the most state-of-the-art one that extends the attribute weighting into subspace. Space structure-Based Clustering (SBC) [23], Coupled Data Embedding-based clustering (CDE) [25], and UNsupervised heTerogeneous couplIng lEarning-based clustering (UNTIE) [26] are also chosen as the counterparts in the stream of data representation-based clustering. SBC has two versions, denoted as SBC-1 and SBC-2, whose difference is to adopt the different distance functions only. For simplicity, we just therefore report the performance of the one with better performance on each data set. Also, the state-of-the-art Distance Learning-based Clustering (DLC) [33] algorithm is selected. Since it is designed for ordinal data only, we first perform the 'simple coding' as discussed in Section 2.1 to encode the nominal attributes of mixed data sets, and then perform DLC for clustering.

We select categorical data distance measures as counterparts of the proposed HD distance metric. We select Hamming distance metric [34] because it is the most commonly used one in categorical data clustering. We also select Lin's Similarity Measure (LSM) [16] as a representative for the stream of entropy-based measures, and Context-Based Distance Metric (CBDM) [21] as a representative for the stream of context-based metrics. We also select three state-of-the-art categorical data distance metrics, i.e., Jia's Distance Metric (JDM) [22], Entropy-Based Distance Metric (EBDM) [31], and Coupled Metric Similarity (CMS) [17] as counterparts. We set the parameters of the above-mentioned counterparts at the values suggested by the corresponding papers.

6.1.4 Data Sets

We collect 15 data sets for the experiments, and the data statistics are shown in Table 3.

Among the six mixed categorical data sets (mixed data sets for short), Lenses, Breast Cancer (abbreviated as Cancer), Hayes-Roth (abbreviated as Hayes), Lymphography (abbreviated as Lym), and Nursery, are benchmark data sets collected from the UCI Machine Learning Repository (UCI-MLR)[1] [61], Assistant Evaluation (abbreviated as Assistant) is a real mixed categorical data set collected from university questionnaires. Among the five ordinal data sets, Lecturer Evaluation (abbreviated as Lecturer), Social Works (abbreviated as Social), and Employee Selection (abbreviated as Selection) are benchmark data sets collected from the Weka website[2] [62], Photo Evaluation (abbreviated as Photo) is a real ordinal data set collected from university questionnaires, and Car Evaluation (abbreviated as Car) is a benchmark data set collected from UCI-MLR. For all these five ordinal data sets, monotonic correlation exists among all the attributes, i.e., an object composed of higher ranked values always ranks higher in comparison with the other objects composed of lower ranked values [29]. To utilize such known monotonicity, the original object-cluster distance is replaced with $\mathrm{dist}(\mathbf{x}_i, C_l) = |\mathrm{dist}(\mathbf{x}_0, \mathbf{x}_i) - \mathrm{dist}(\mathbf{x}_0, C_l)|$ for the measures (i.e., LSM, EBDM, DLC, and HD) that are capable in distinguishing the order of values, in conducting the clustering experiment in Section 6.2. Note that $\mathbf{x}_0$ here is a constructed object composed of the highest ranked value of each attribute. All the four nominal data sets, i.e., Solar Flare (abbreviated as Solar), Zoo, Voting Records (abbreviated as Voting), and Soybean, are benchmark data sets collected from UCI-MLR.

TABLE 3
Statistics of the 15 Utilized Data Sets

Data type	Data set	# Instance	# Attribute	# Class
Mixed	Lenses	24	2+2	12
	Assistant	72	2+2	3
	Hayes	132	2+2	3
	Lym	148	3+15	4
	Cancer	286	4+5	2
	Nursery	12,960	7+1	4
Ordinal	Photo	66	4	3
	Selection	488	4	9
	Lecturer	1,000	4	5
	Social	1,000	10	4
	Car	1,728	7	4
Nominal	Soybean	47	21	4
	Zoo	101	16	7
	Solar	323	9	6
	Voting	435	16	2

"# Attribute" of mixed categorical data sets indicates "# ordinal attributes + # nominal attributes"

6.1.5 Initialization of HD-NDW

In **Step 0** of the proposed HD-NDW algorithm, values of P and W should be initialized. For P, although different initialization strategies can be utilized, we adopt a strategy similar to the random initialization of the conventional k-modes algorithm. That is, we randomly select k^* objects as modes, and then assign values to the $k^* \times d$ vectors of P accordingly. Taking the data set shown in Table 1 as an example, suppose we have $o_1^1 = \uparrow$, $o_2^1 = \uparrow\uparrow$, $o_3^1 = \uparrow\uparrow\uparrow$, $o_4^1 = \uparrow\uparrow\uparrow\uparrow$, $o_1^2 =$non-special, $o_2^2 =$vesicles, and $o_3^2 =$chalices. If the 6th object in Table 1 (i.e., $\mathbf{x}_6 = [\uparrow\uparrow\uparrow\uparrow, \mathrm{vesicles}]^\top$) is initialized as

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

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

TABLE 4
Clustering Performance of Various Clustering Algorithms on Mixed and Ordinal Categorical Data Sets

Index	Data Set	KMD	ECC	WKM	MWKM	SBC	WOC	CDE	UNTIE	DLC	HD-NDW	Δ
ARI	Assistant	0.111±0.06	0.133±0.09	0.113±0.08	0.138±0.09	0.153±0.04	0.194±0.08	0.131±0.06	0.152±0.04	0.152±0.09	**0.330±0.05**	70.1%
	Lenses	0.088±0.13	0.104±0.14	0.087±0.17	0.124±0.13	0.148±0.11	0.117±0.16	0.085±0.15	0.088±0.13	0.146±0.10	**0.227±0.21**	53.4%
	Cancer	0.018±0.05	0.050±0.07	0.014±0.04	0.056±0.06	0.083±0.08	0.076±0.07	0.083±0.08	0.085±0.11	0.035±0.05	**0.090±0.10**	6.5%
	Hayes	-0.001±0.03	0.017±0.05	0.020±0.02	0.016±0.01	-0.012±0.01	0.019±0.04	0.081±0.04	0.084±0.06	0.026±0.03	**0.091±0.03**	8.8%
	Lym	0.108±0.04	0.194±0.04	0.075±0.05	0.131±0.05	0.127±0.07	0.163±0.06	0.193±0.03	0.197±0.05	**0.200±0.06**	0.195±0.03	-2.4%
	Nursery	0.054±0.02	0.072±0.10	0.083±0.11	0.058±0.02	0.017±0.01	0.002±0.00	0.053±0.02	0.084±0.02	0.115±0.08	**0.133±0.07**	15.8%
	Photo	0.102±0.06	0.121±0.09	0.100±0.08	0.140±0.09	0.186±0.05	0.158±0.09	0.115±0.07	0.115±0.09	0.267±0.07	**0.318±0.06**	19.3%
	Lecturer	0.034±0.02	0.035±0.02	0.032±0.02	0.038±0.01	0.046±0.01	0.040±0.02	0.034±0.02	0.033±0.02	0.151±0.01	**0.154±0.01**	1.5%
	Social	0.043±0.02	0.059±0.02	0.043±0.01	0.047±0.02	0.093±0.02	0.036±0.02	0.068±0.02	0.071±0.02	0.108±0.00	**0.112±0.01**	3.2%
	Selection	0.151±0.04	0.181±0.03	0.173±0.03	0.171±0.03	0.200±0.01	0.183±0.04	0.219±0.03	0.221±0.03	0.313±0.01	**0.328±0.02**	5.1%
	Car	0.025±0.04	0.058±0.05	0.026±0.04	0.031±0.02	0.027±0.03	0.035±0.03	0.019±0.05	0.023±0.06	0.112±0.02	**0.128±0.04**	14.5%
Averaged Rank		8.55	5.82	8.18	6.18	5.09	5.64	6.45	4.91	3.00	**1.18**	
NMI	Assistant	0.152±0.07	0.182±0.10	0.160±0.10	0.172±0.10	0.184±0.05	0.262±0.09	0.159±0.07	0.188±0.06	0.212±0.11	**0.390±0.04**	48.8%
	Lenses	0.227±0.10	0.255±0.14	0.199±0.18	0.276±0.12	0.305±0.07	0.262±0.14	0.203±0.14	0.213±0.13	0.308±0.09	**0.342±0.16**	11.3%
	Cancer	0.011±0.02	0.029±0.04	0.008±0.02	0.024±0.03	0.040±0.03	0.034±0.03	0.045±0.04	0.046±0.05	0.014±0.02	**0.062±0.03**	37.1%
	Hayes	0.019±0.04	0.032±0.05	0.026±0.02	0.033±0.03	0.003±0.01	0.043±0.06	0.087±0.03	0.086±0.05	0.032±0.03	**0.103±0.03**	18.3%
	Lym	0.168±0.04	0.243±0.04	0.130±0.05	0.188±0.05	0.170±0.04	0.231±0.06	0.237±0.04	0.243±0.05	0.223±0.05	**0.258±0.03**	5.9%
	Nursery	0.059±0.02	0.103±0.13	0.105±0.13	0.103±0.03	0.032±0.02	0.006±0.00	0.056±0.02	0.101±0.03	0.117±0.11	**0.162±0.09**	39.2%
	Photo	0.143±0.06	0.177±0.09	0.151±0.10	0.180±0.10	0.221±0.05	0.222±0.11	0.181±0.08	0.200±0.10	0.339±0.03	**0.373±0.03**	10.1%
	Lecturer	0.054±0.02	0.059±0.02	0.057±0.02	0.060±0.02	0.073±0.02	0.064±0.03	0.056±0.02	0.059±0.02	0.215±0.01	**0.217±0.01**	0.8%
	Social	0.065±0.02	0.086±0.02	0.060±0.02	0.068±0.02	0.131±0.02	0.059±0.02	0.094±0.02	0.088±0.01	0.167±0.00	**0.168±0.01**	0.2%
	Selection	0.280±0.04	0.335±0.03	0.305±0.03	0.308±0.02	0.353±0.01	0.307±0.04	0.370±0.02	0.368±0.02	0.491±0.01	**0.510±0.01**	3.7%
	Car	0.047±0.02	0.121±0.07	0.062±0.05	0.064±0.03	0.071±0.04	0.079±0.06	0.091±0.07	0.106±0.06	0.219±0.01	**0.228±0.01**	3.9%
Averaged Rank		9.00	5.55	8.45	6.27	5.55	5.64	5.64	4.55	3.36	**1.00**	
CA	Assistant	0.522±0.07	0.536±0.08	0.527±0.09	0.546±0.09	0.568±0.08	0.621±0.07	0.531±0.06	0.549±0.05	0.570±0.09	**0.639±0.07**	2.9%
	Lenses	0.534±0.09	0.537±0.11	0.538±0.10	0.557±0.10	0.564±0.09	0.544±0.11	0.512±0.10	0.513±0.07	0.561±0.07	**0.588±0.13**	4.1%
	Cancer	0.564±0.06	0.586±0.09	0.536±0.07	0.614±0.08	0.624±0.11	0.629±0.10	0.630±0.10	0.630±0.10	0.584±0.08	**0.651±0.09**	3.4%
	Hayes	0.384±0.03	0.414±0.06	0.439±0.05	0.416±0.02	0.354±0.02	0.413±0.08	0.442±0.05	0.452±0.04	0.446±0.05	**0.487±0.05**	7.7%
	Lym	0.462±0.05	0.512±0.04	0.433±0.07	0.482±0.06	0.505±0.06	0.551±0.05	0.519±0.04	0.550±0.05	0.538±0.06	**0.601±0.07**	9.2%
	Nursery	0.378±0.04	0.368±0.07	0.395±0.09	0.359±0.03	0.323±0.03	0.292±0.02	0.366±0.00	0.397±0.01	0.404±0.04	**0.423±0.06**	4.8%
	Photo	0.511±0.07	0.524±0.08	0.517±0.09	0.553±0.09	0.557±0.05	0.584±0.09	0.501±0.06	0.500±0.09	0.668±0.06	**0.698±0.05**	4.4%
	Lecturer	0.335±0.03	0.328±0.03	0.319±0.03	0.322±0.03	0.339±0.02	0.331±0.04	0.319±0.03	0.338±0.05	0.455±0.04	**0.465±0.04**	2.2%
	Social	0.370±0.03	0.384±0.03	0.371±0.02	0.372±0.03	0.421±0.02	0.372±0.03	0.391±0.02	0.409±0.02	0.414±0.02	**0.453±0.04**	7.6%
	Selection	0.365±0.04	0.372±0.04	0.373±0.03	0.369±0.04	0.386±0.01	0.427±0.05	0.407±0.03	0.437±0.04	**0.505±0.03**	0.489±0.03	-3.2%
	Car	0.370±0.04	0.384±0.06	0.375±0.07	0.369±0.03	0.357±0.04	0.366±0.05	0.389±0.05	0.390±0.05	0.437±0.04	**0.453±0.06**	3.5%
Averaged Rank		8.18	6.55	7.45	6.91	5.64	5.45	6.32	4.41	3.00	**1.09**	

The column of 'Δ' reports the improvements achieved by HD-NDW in comparison with the best-performing counterparts on different data sets. Results of significance tests are shown in Table 6 and Fig. 4.

the mode of the 2nd cluster, then the corresponding two vectors in P_2 will be $\mathbf{p}_2^1 = [0,0,0,1]^\top$ and $\mathbf{p}_2^2 = [0,1,0]^\top$, respectively. For W, we uniformly initialize each weight of it to $1/(\sum_{r=1}^{d} v^r(v^r-1)/2)$. In this way, the sum of all the initialized weights equals to 1, which is equal to the sum of the weights after updating, as the updated weights will be processed using soft-max (see the discussions following Eq. (18) for more details). Another purpose of such a uniform initialization is to make the initialized weights have no effect on the learning of **Step 1** in Algorithm 1. If we randomly initialize W, inappropriate distance weights will prevent **Step 1** from learning reasonable data partition, which will further influence the subsequent learning iterations.

6.2 Clustering Performance Evaluation of HD-NDW

Since a key working principle of HD-NDW is to convert the nominal attributes into ordinal ones for more reasonable distance measurement, the superiority of HD-NDW will be more prominent on mixed and ordinal data sets. In order to conduct a more targeted evaluation, we report the clustering performance on mixed and ordinal data sets in Table 4. To ensure the completeness of the evaluation, the performance on nominal data sets is reported in Table 5. The best and second-best results are highlighted using boldface and underline, respectively. Improvements achieved by HD-NDW in comparison with the best-performing counterparts on different data sets are reported in the column of 'Δ'. For each data set, the compared methods are ranked according to their performance, and the averaged rank of each method is reported. From the results shown in Tables 4 and 5, we have the following observations:

- HD-NDW obviously outperforms the other counterparts on mixed categorical data sets, because the homogeneous distance definition and the distance weighting mechanism may have desired effects on mixed categorical data sets.
- HD-NDW and DLC significantly outperform the other counterparts on ordinal data sets, because they take into account the intra- and inter-attribute order relationship, by which the learned distances are more appropriate for clustering.
- In the comparison on nominal data sets, superiority of HD-NDW is not as significant as on mixed and ordinal data sets because the HD component that uniformly defines distances for ordinal and nominal attributes will not have desired impact when processing nominal data. Nevertheless, since NDW still

TABLE 5
Clustering Performance of Various Clustering Algorithms on Nominal Data Sets

Index	Data Set	KMD	ECC	WKM	MWKM	SBC	WOC	CDE	UNTIE	HD-NDW	Δ
ARI	Solar	0.223±0.06	0.194±0.06	0.132±0.09	0.199±0.06	0.126±0.03	0.229±0.10	0.237±0.08	0.255±0.10	**0.318±0.08**	24.8%
	Zoo	0.628±0.18	0.530±0.15	0.651±0.18	0.594±0.18	0.413±0.14	0.618±0.13	0.741±0.11	**0.748±0.13**	0.721±0.15	-3.6%
	Voting	0.520±0.02	0.544±0.01	0.535±0.00	0.542±0.01	0.560±0.03	0.537±0.00	0.534±0.08	0.558±0.07	**0.564±0.00**	0.6%
	Soybean	0.688±0.22	0.659±0.16	0.772±0.21	0.740±0.22	0.816±0.11	0.788±0.21	0.821±0.19	**0.829±0.17**	0.803±0.21	-3.1%
Averaged Rank		6.75	7.00	6.25	6.25	5.75	5.25	3.75	**1.75**	2.25	
NMI	Solar	0.300±0.05	0.278±0.06	0.218±0.10	0.271±0.06	0.196±0.03	0.331±0.09	0.319±0.08	0.348±0.10	**0.408±0.08**	17.2%
	Zoo	0.753±0.09	0.700±0.06	0.779±0.08	0.745±0.08	0.595±0.08	0.786±0.05	**0.810±0.05**	0.808±0.08	0.809±0.08	-0.1%
	Voting	0.448±0.02	0.476±0.01	0.452±0.01	0.473±0.01	0.473±0.00	0.475±0.00	0.462±0.07	0.458±0.08	**0.489±0.00**	2.7%
	Soybean	0.805±0.15	0.771±0.10	0.849±0.12	0.847±0.13	0.856±0.06	0.885±0.11	0.892±0.11	**0.902±0.10**	0.897±0.11	-0.6%
Averaged Rank		7.00	6.25	6.75	6.50	6.75	3.50	3.50	3.25	**1.50**	
CA	Solar	0.482±0.05	0.442±0.05	0.400±0.06	0.462±0.05	0.400±0.04	0.483±0.07	0.483±0.07	0.498±0.06	**0.540±0.05**	8.3%
	Zoo	0.676±0.13	0.623±0.10	0.703±0.12	0.647±0.13	0.554±0.09	0.669±0.10	0.758±0.08	**0.778±0.09**	0.760±0.10	-2.4%
	Voting	0.861±0.01	0.869±0.01	0.852±0.00	0.868±0.00	0.875±0.00	0.867±0.00	0.864±0.04	0.872±0.05	**0.876±0.00**	0.1%
	Soybean	0.791±0.17	0.773±0.14	0.837±0.17	0.811±0.17	0.874±0.10	0.821±0.17	0.874±0.14	**0.876±0.13**	0.849±0.16	-3.1%
Averaged Rank		6.50	7.00	6.75	6.25	5.25	5.50	4.00	**1.75**	2.00	

The column of 'Δ' reports the improvements achieved by HD-NDW in comparison with the best-performing counterparts on different data sets.

acts in booting the clustering performance, HD-NDW is still competitive in comparison with the state-of-the-art UNTIE, and obviously outperforms the others.

- Although UNTIE is not specially designed for representing data set with ordinal attributes, it still shows strong data representation ability, because it performs the best in comparison with the counterparts except the two methods (i.e., DLC and HD-NDW) that contain specially designed mechanisms for exploiting the information embedded in ordinal attributes. As for the performance on nominal data sets, UNTIE performs the best in general, while HD-NDW is still very competitive.

6.2.1 Significance Test

According to the averaged rank shown in Table 4, UNTIE and DLC are clearly the two most competitive counterparts. We conduct significance test using Wilcoxon signed-rank test and report the results in Table 6. It can be seen that even at 99 percent confidence interval, HD-NDW is still significantly better than the two counterparts in terms of all three validity indices.

To intuitively compare the proposed HD-NDW with the other counterparts, we further perform Bonferroni-Dunn test [60] on the performance of different methods and visualize the results in Fig. 4. The counterparts rank outside the Critical Difference (CD) intervals are believed to be significantly different from HD-NDW. It can be observed from Fig. 4a that HD-NDW is significantly better than almost all five methods proposed in recent five years at confidence interval 90 percent. In comparison with all nine counterparts in Fig. 4b, HD-NDW is still significantly better than eight counterparts at confidence interval 90 percent. Although HD-NDW is not significantly better than DLC, HD-NDW is capable in processing any-type categorical data, while DLC is designed for ordinal data only, which cannot be directly used for mixed data and is incompetent for nominal data. Therefore, HD-NDW also demonstrates superiority in comparison with DLC in terms of availability for nominal data clustering.

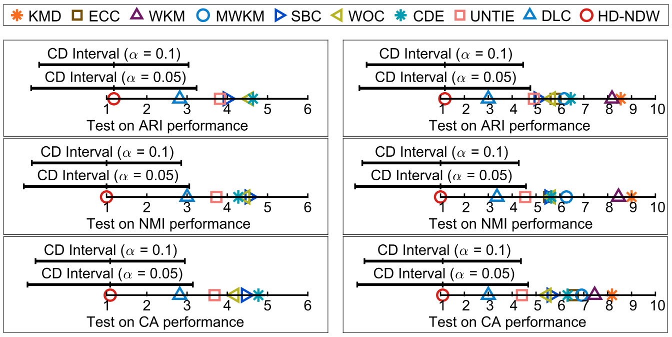
TABLE 6
Wilcoxon Signed-Rank Test on the Performance of HD-NDW versus DLC and HD-NDW versus UNTIE

Index	HD-NDW versus DLC	HD-NDW versus UNTIE
ARI	+	+
NMI	+	+
CA	+	+

The symbol "+" indicates that HD-NDW is significantly different from a certain counterpart for the two-tailed Wilcoxon signed-rank test at confidence interval 99 percent (i.e., α = 0.01).

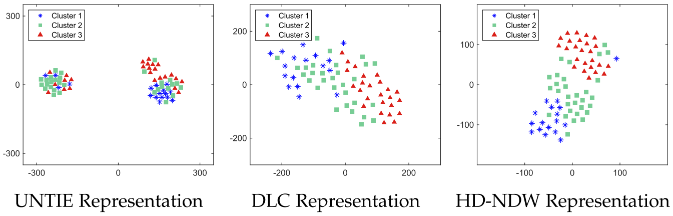
6.2.2 Visualization of Data Representation

To intuitively compare the reasonableness of the learned representation or distances of the three best performing methods, i.e., UNTIE, DLC, and HD-NDW, we visualize their representations in Fig. 5 by converting them into two-dimensional points using t-Distributed Stochastic Neighbor Embedding (t-SNE) [63]. Since DLC and HD-NDW are not



(a) Test for recent-five-year methods. (b) Test for all the compared methods.

Fig. 4. Bonferroni-Dunn (BD) test on the performance of (a) methods proposed in recent five years, and (b) all the compared methods. Critical Difference (CD) for the two-tailed BD tests in (a) at confidence interval 95 percent (α = 0.05) and 90 percent (α = 0.1) are 2.05 and 1.86, respectively. CD for the two-tailed BD tests in (b) at confidence interval 95 percent (α = 0.05) and 90 percent (α = 0.1) are 3.58 and 3.28, respectively. The counterparts rank outside the CD intervals are believed to be significantly different from HD-NDW.



UNTIE Representation DLC Representation HD-NDW Representation

Fig. 5. t-SNE visualization of the representations produced by UNTIE, DLC, and HD-NDW on Assistant data set. The three types of markers indicate data objects belonging to different true clusters.

representation-based methods, we first use them to learn intra-attribute distances, and then encode the data values using the learned distances for representation. Since the distances learned by DLC satisfy $\text{dist}(o_a, o_b) + \text{dist}(o_b, o_c) = \text{dist}(o_a, o_c)$ for $a < b < c$ or $a > b > c$, we directly encode the possible values by $o_1 = 0$, $o_2 = \text{dist}(o_1, o_2)$, $o_3 = \text{dist}(o_1, o_3)$, and so on, which will not twist the distances learned by DLC. For the distances learned by HD-NDW, we encode a possible value using the distances between it and all the intra-attribute possible values to preserve the information of the learned distances. For example, for an attribute with possible values $\{o_a, o_b, o_c\}$, the value o_b is encoded into a vector $[\text{dist}(o_a, o_b), \text{dist}(o_b, o_b), \text{dist}(o_c, o_b)]^\top$ by the HD-NDW learned distance metric. Note that the HD-NDW distance here is the one defined by HD multiplied by the corresponding distance weight learned by HD-NDW.

It can be observed that the true clusters in the HD-NDW-represented data set are obviously more separable in comparison with UNTIE and DLC. The reason should be that Assistant is a mixed categorical data set that is composed of nominal and ordinal attributes. For this kind of data, UNTIE is unable to take into account the order information embedded in ordinal attributes, while DLC is unsuitable for learning distances of nominal attributes.

6.2.3 *Visualization of Cluster Discrimination*

Averaged ICD computed based on the true cluster labels of a data set can intuitively indicate the discrimination ability of a distance metric. According to [19], averaged ICD between two clusters C_l and C_t with n_l and n_t data objects, respectively, is computed by $\sum_{\mathbf{x}_i \in C_l} \sum_{\mathbf{x}_j \in C_t} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)/(n_l n_t)$. When $l = t$, it computes the averaged intra-attribute distance; otherwise, it computes the averaged inter-attribute distance. Since different distance metrics may have different scales, computing the multiple relationship between the averaged intra- and inter-cluster distances is a feasible solution [22] to fairly compare the discrimination ability of different metrics. Therefore, we pre-process the ICD matrix of each distance metric by dividing all the values in the matrix by the minimum value in this matrix. Then we visualize the pre-processed ICD matrices as gray scale maps in Fig. 6. ICD matrix of a better distance metric should be darker on the main diagonal and lighter on the other locations, which indicates smaller averaged intra-cluster distances and larger averaged inter-cluster distances, respectively. From Fig. 6, it is clear that HD-NDW has better cluster discrimination ability than UNTIE and DLC.

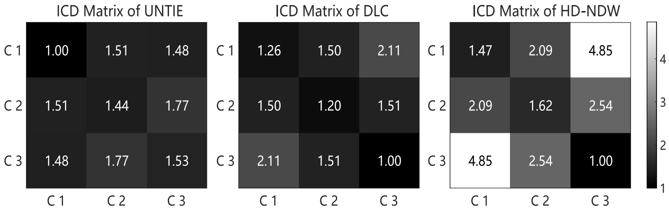


Fig. 6. Gray scale maps of the ICD matrices produced by UNTIE, DLC, and HD-NDW on Assistant data set. Darker on the main diagonal and lighter on the other locations indicate a better distance metric.

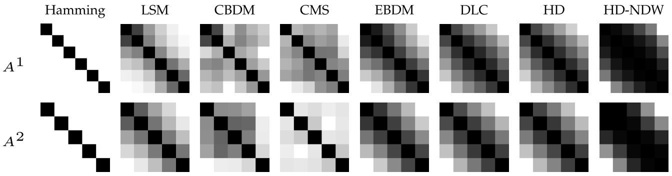


Fig. 7. Gray scale maps of the intra-attribute distance matrices of the two ordinal attributes of Assistant data set produced by various distance metrics.

6.3 Effectiveness Evaluation of HD

In Fig. 7, we visualize the intra-attribute distances produced by different distance measures to intuitively compare them. JDM is not compared in Fig. 7 because it directly measures object-cluster distance and does not produce intra-attribute distances. The produced distances are first normalized into the interval [0,1] using min-max scaling, and then the normalized distances are visualized by converting them into corresponding gray scale pixels. A lighter pixel represents a larger distance between two possible values, and a pure black pixel represents a distance value of 0. In Fig. 7, the pixel located at the mth column and hth row of a gray scale map represents the distance between the mth and hth possible values of the corresponding attribute. In general, two possible values with larger order difference should have larger distance, and thus the pixels should be darker on the main diagonal, and lighter towards the upper right and lower left corners in the gray scale maps.

It can be observed that Hamming distance is completely incapable in distinguishing the distances between different possible values. Although CBDM and CMS exploits more context information for distance measurement, they cannot reveal the order relationship among possible values of ordinal attributes. Obviously, the distances produced by LSM, EBDM, DLC, HD, and HD-NDW are consistent with the order relationship among possible values of the two ordinal attributes. Since the distances produced by HD-NDW is the weighted version of the distances produced by HD, gray scale maps of HD and HD-NDW are different in Fig. 7, but they both reflect the order relationship.

Fig. 8 compares clustering performance of different distance measures and illustrates that even not weighted by NDW, distance measured using HD is still very competent. More detailed observations are provided below:

- HD outperforms the other counterparts on mixed data sets because it is the only one that can measure intra-attribute distances of nominal and ordinal attributes in a homogeneous way. HD outperforms the other counterparts on ordinal data sets because it preserves the order relationship among ordered possible values.

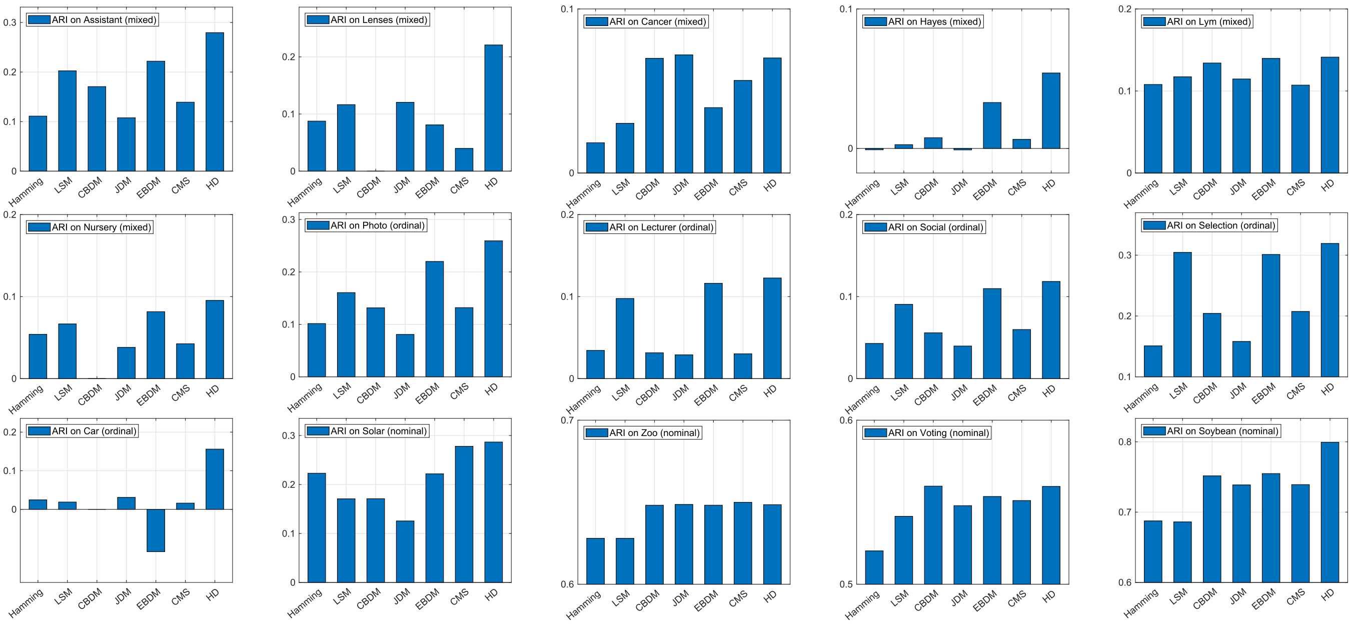


Fig. 8. Clustering performance of various distance measures on mixed, ordinal, and nominal data sets, where a better measure yields a higher value.

- On Zoo, Voting, Cancer, and Lym data sets, performance of HD is competitive but cannot be obviously better than the others. This may be because that the above-mentioned four data sets are composed of more nominal attributes, which weakens the advantages of HD accordingly.
- ARI performance of CBDM is exactly 0 on Lenses, Nursery, and Car data sets because these data sets are composed of independent attributes and CBDM fails in measuring distances for such data sets.

6.4 Effectiveness Evaluation of NDW

Clustering performance of the original version of HD-NDW and the version without NDW (abbreviated as non-NDW) is demonstrated in Fig. 9. By comparing them, effectiveness of the NDW mechanism can be empirically proved.

It can be observed that HD-NDW performs better than the non-NDW version on all the data sets, which indicates that the NDW mechanism does optimize the distance weights during the clustering of HD-NDW to obtain better clustering results. It can also be observed that HD-NDW does not outperform non-NDW a lot on Lenses, Voting and Soybean data sets. This may be because most attributes of these three data sets have only two possible values, and for such attributes, there is only one intra-attribute distance to be weighted during clustering, which makes NDW degrades into a conventional attribute weighting mechanism, and thus obscures the merits of NDW.

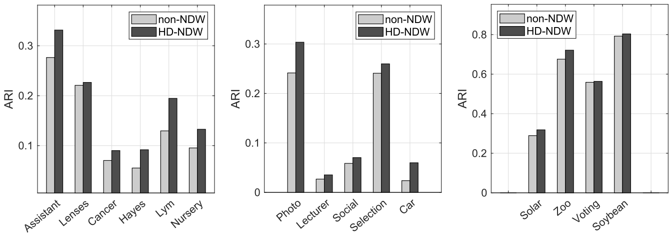


Fig. 9. Clustering performance of HD-NDW and its version without NDW (non-NDW for short) on mixed, ordinal, and nominal data sets. A higher value indicates a better clustering performance.

6.5 Convergence Evaluation

We plot the convergence curves of HD-NDW on each data set in Fig. 10. Specifically, after each iteration of **Step 1** in Algorithm 1, 'No. of Iteration' is added by 1, and the current 'Error' (i.e., the current value of objective function) is plotted. When **Step 1** converges and **Step 2** is triggered, the current 'Error' is marked by a circle. When the whole algorithm converges, the current 'Error' is marked by a box.

It can be seen that HD-NDW converges within 6 - 22 iterations on different data sets, which is very fast for learning a large number (i.e., $\sum_r^d v^r(v^r-1)/2$) of intra-attribute distance weights. Moreover, the convergence curves are monotonically decreasing, and 'Error' decreases sharply after updating the distance weights, which clearly illustrates the effectiveness of HD-NDW.

In our experiments, since the true number k^* of the clusters is utilized, partition learned by **Step 1** is relatively reasonable, which offers useful information for learning W in **Step 2**. This would be the reason why **Step 2** is always triggered 2 - 3 times for different data sets.

6.6 Computational Efficiency Evaluation

We randomly generate synthetic categorical data sets to evaluate the computational efficiency of different clustering methods in terms of four data factors: (1) number of data objects (n), (2) number of attributes (d), (3) number of possible values per attribute (V), and number of clusters (k). Synthetic data sets are generated by increasing the value of one factor and fixing the other three factors at the default values. The default values are set at $n = 10k$, $d = 10$, $V = 3$, and $k = 2$. The value ranges for increasing each factor are set at $n = \{10k, 20k, \ldots, 100k\}$, $d = \{10, 20, \ldots, 100\}$, $V = \{3, 10, 20, \ldots, 90\}$, and $k = \{2, 4, \ldots, 20\}$. As HD-NDW is proposed for mixed categorical data clustering, we let it treat each generated data set as comprising $d/2$ nominal and $d/2$ ordinal attributes in this

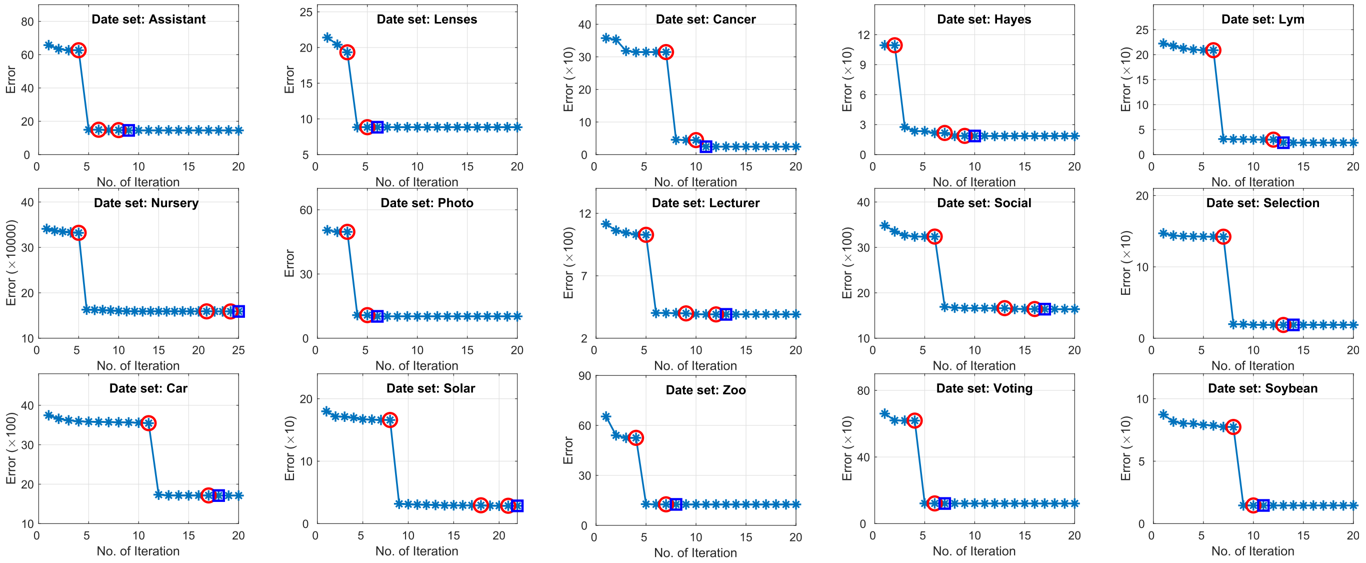


Fig. 10. Convergence curves of HD-NDW on mixed, ordinal, and nominal data sets. The circles indicate the moments that Step 2 of Algorithm 1 is triggered, and the boxes indicate the moments of convergence of Algorithm 1.

experiment. Since the data representation learning of SBC, CDE, UNTIE, and the distance computation of HD-NDW are necessarily processed for clustering, their execution time is counted in for comparison. We plot the execution time of different methods in terms of the four data factors in Fig. 11. It can be observed that the computation cost of HD-NDW has approximately linear relation with n and k, which are consistent with the time complexity analysis in Sections 4.3 and 5.3.

In comparison with the state-of-the-art methods (i.e., UNTIE and DLC), it can be observed that the trends and values of the computation cost of HD-NDW and DLC are almost the same in terms of n and k. Furthermore, HD-NDW has lower computation cost than UNTIE in terms of n. The computation cost of HD-NDW and DLC has higher increasing rate than UNTIE over k, because HD-NDW and DLC connect the distance learning with the target clustering task, and thus have better clustering performance in general as shown in Table 4. Since k is usually a very small value from the practical point of view, k will not have a big impact on the efficiency of HD-NDW. Moreover, although the computation cost of HD-NDW has higher increasing rate over d and V, the computations (e.g., the computation of each value in D, and the computation of each value in W) that are related to these two factors are independent and can be easily parallelized for acceleration.

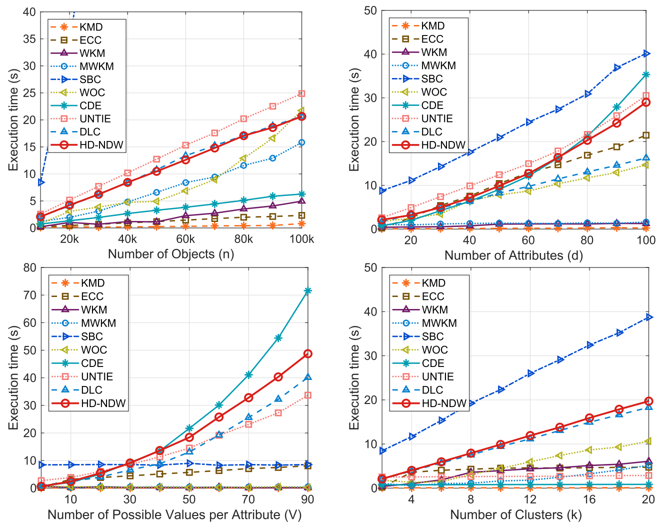


Fig. 11. Execution time of various clustering algorithms w.r.t. number of objects (n), number of attributes (d), number of possible values per attribute (V), and number of clusters (k).

In summary, HD-NDW does not bring much extra computation cost in comparison with the state-of-the-art methods, and its computation cost has a linear relation with n, which is generally the most concerned factor in terms of the computational efficiency of a clustering method.

7 Conclusion

In this paper, we have proposed HD intra-attribute distance definition and NDW distance weighting mechanism, both of which are utilized to present HD-NDW clustering algorithm for data clustering with nominal and ordinal attributes. HD is formed based on the intrinsic connection of ordinal and nominal attributes, and can therefore define their intra-attribute distances in a homogeneous way. In the clustering process of HD-NDW, NDW novelly quantifies and iteratively updates the weights of intra-attribute distances defined by HD according to the present data partition, thereby ensuring an effective learning of the importance of intra-attribute distances for searching optimal clustering results. It turns out that HD-NDW is capable of clustering categorical data composed of any combination of nominal and ordinal attributes. Extensive experimental results have demonstrated that HD-NDW always converges quickly and has superior clustering performance in comparison with the existing counterparts.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61672444, in part by Hong Kong Baptist University (HKBU), Research Committee, Initiation Grant Faculty Niche Research Areas

(IG-FNRA) 2018/19, under Grant RC-FNRA-IG/18-19/SCI/03, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Project ITS/339/18, and in part by the HKBU Interdisciplinary Research Clusters Matching Scheme (IRCMS) under Project: RC-IRCMs/18-19/01.

References

[1] A. Agresti, *Categorical Data Analysis*. Hoboken, NJ, USA: Wiley, 2003.

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

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

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

[5] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[6] D. S. Yeung and X. 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1223–1235, Aug. 2006.

[13] 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.

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

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

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

[17] S. Jian, L. Cao, K. Lu, and H. Gao, "Unsupervised coupled metric similarity for non-IID categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1810–1823, Sep. 2018.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, Oct. 2016.

[24] S. Jian, L. Cao, G. Pang, K. Lu, and H. Gao, "Embedding-based representation of categorical data by hierarchical value coupling learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1937–1943.

[25] S. Jian, G. Pang, L. Cao, K. Lu, and H. Gao, "CURE: Flexible categorical data representation by hierarchical coupling learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 853–866, May 2019.

[26] C. Zhu, L. Cao, and J. Yin, "Unsupervised heterogeneous coupling learning for categorical representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 21, 2020, doi: 10.1109/TPAMI.2020.3010953.

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

[28] V. E. Johnson and J. H. Albert, *Ordinal Data Modeling*. Berlin, Germany: Springer, 2006.

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

[30] Y. Zhang, Y.-M. Cheung, and K. 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.

[31] Y. Zhang and Y.-M. Cheung, "A new distance metric exploiting heterogeneous inter-attribute relationship for ordinal-and-nominal-attribute data clustering," *IEEE Trans. Cybern.*, early access, Apr. 27, 2020, doi: 10.1109/TCYB.2020.2983073.

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

[33] Y. Zhang and Y.-M. Cheung, "An ordinal data clustering algorithm with automated distance learning," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 6869–6876.

[34] V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, *Data Science and Classification*. Heidelberg, Germany: Springer, 2006.

[35] Y. Zhang and Y.-M. Cheung, "Exploiting order information embedded in ordered categories for ordinal data clustering," in *Proc. 24th Int. Symp. Methodologies Intell. Syst.*, 2018, pp. 247–257.

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

[37] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific-Asia Conf. Knowl. Discov. Data Mining*, 1997, pp. 21–34.

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

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

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

[41] S. E. Schaeffer, "Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, 2007.

[42] M. C. Nascimento and A. C. De Carvalho, "Spectral methods for graph clustering–A survey," *Eur. J. Oper. Res.*, vol. 211, no. 2, pp. 221–231, 2011.

[43] M. Zhang, N. Wang, Y. Li, and X. Gao, "Neural probabilistic graphical model for face sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2623–2637, Jul. 2020.

[44] S. Li, C. Gentile, and A. Karatzoglou, "Graph clustering bandits for recommendation," 2016, *arXiv:1605.00596*.

[45] V. Garro and A. Giachetti, "Scale space graph representation and kernel matching for non rigid and textured 3D shape retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1258–1271, Jun. 2016.

[46] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[47] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, 2001, pp. 251–256.

[48] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric for image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1590–1602, Aug. 2011.

[49] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, 2014, pp. 757–765.

[50] K. Mahadik, Q. Wu, S. Li, and A. Sabne, "Fast distributed bandits for online recommendation systems," in *Proc. 34th ACM Int. Conf. Supercomputing*, 2020, pp. 1–13.

[51] S. Li and P. Kar, "Context-aware bandits," 2015, *arXiv:1510.03164*.

[52] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 539–548.

[53] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Proc. 19th Int. Conf. Artif. Neural Netw.*, 2009, pp. 175–184.

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

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

[56] 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.

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

[58] L. Lovász, *Matching Theory*. Amsterdam, The Netherlands: North-Holland, 1986.

[59] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.

[61] D. Dua and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

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

[63] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Yiqun Zhang (Member, IEEE) received the BEng degree from the South China University of Technology, China, in 2013, and the MS and PhD degrees from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2014 and 2019, respectively. He is currently a lecturer at the School of Computers, Guangdong University of Technology, Guangzhou, China. He is also 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 PhD degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. He is currently a full professor at the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, computer vision, pattern recognition, data mining, nonlinear optimization, and information hiding. He serves as an associate editor of the *IEEE Transactions on Cybernetics*, *IEEE Transactions on Emerging Topics in Computational Intelligence*, *IEEE Transactions on Cognitive and Developmental Systems*, *IEEE Transactions on Neural Networks and Learning Systems* (2014-2020), *Pattern Recognition*, and *Neurocomputing*, to name a few. For more information, please visit http://www.comp.hkbu.edu.hk/∼ymc