2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650134

Clustering by Learning the Ordinal Relationships of Qualitative Attribute Values

Pengkai Wang[a‡], Yunfan Zhang[a‡], Yiqun Zhang[a*], Yang Lu[b], Mengke Li[c], and Yiu-ming Cheung[d]
[a]Guangdong University of Technology, China, [b]Xiamen University, China
[c]Guangming Lab & Shenzhen University, China, [d]Hong Kong Baptist University, Hong Kong
[‡]Co-first author, [*]Corresponding author: yqzhang@gdut.edu.cn

***Abstract*—In many real-world clustering tasks, data objects are described by both quantitative and qualitative attributes. Attributes with semantically ordered qualitative values are very common and are usually coded according to their order (i.e., consecutive integers) for clustering. However, semantic order is not always globally interdependent with a certain clustering task. An intuitive case is that level of income (attribute) is not always positively correlated with the level of mental health (label). Using mismatched order surely forms a bottleneck to clustering performance, and conversely, the unsupervised clustering process prevents understanding of "true" order. Therefore, we proposed a novel learning paradigm to tune the value order. More specifically, we adjust the intra-attribute orders, and let this process learn mutually with object clustering, thus bridging the gap between value order and clustering task. To the best of our knowledge, this is the first attempt to learn ordinal relationships among qualitative attribute values. Extensive experiments with significance tests show that our method outperforms the existing relevant clustering approaches on qualitative attribute data.**

I. Introduction

Clustering is a fundamental data analysis technique, which is commonly used in many machine learning and data mining tasks. Current diverse data acquisition pathways allow data objects to be described by both the numerical attributes with quantitative and the categorical attributes with qualitative values, where the semantically ambiguous qualitative values are often not well encoded as numerical values by experts prior to cluster analysis. This has gradually shifted clustering research from numerical method [1]–[3] to addressing problems posed by categorical attributes [4]–[6], especially how to define their distances [7], [8], in recent years.

For categorical attributes, a taxonomy further distinguishes the attributes into nominal and ordinal ones based on whether there is a semantic order of possible values, e.g., "strong-accept", "accept", "weak accept", etc., that reviewers might recommend for this submission. The ordinal values are usually simply treated as consecutive integers and thus form an identical distance "1" between each pair of adjacent values. Such distance structure is based on external semantics, and thus cannot serve different clustering tasks adequately. An intuitive survey example [9] is that the mental health classes (i.e., healthy and unhealthy) corresponding to high-, moderate-, and low-income groups are neutral between healthy and unhealthy, tend to be healthy, and tend to be unhealthy, respectively, as shown in Fig. 1. This indicates that the semantic order of attribute "income" somewhat mismatches the task of cluster analysis of mental health patterns, and the mismatch effect accumulates when all the ordinal attributes participate in clustering based on their external semantic orders. Although attributes interactions can somewhat mitigate the effect caused by inaccurate orders, clustering performance can be further enhanced by using the "true" order of the attribute values.

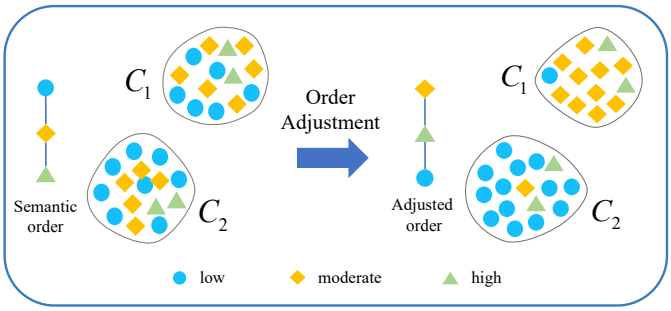

Fig. 1. An example illustrating that order adjustment matters the clustering. The C_1 and C_2 are the healthy and unhealthy clusters, respectively. After the order adjustment, the clustering result better conforms with reality.

Many advanced approaches in the literature are devoted to more comprehensively utilizing the information of categorical attributes, and have achieved considerable success in improving clustering accuracy. According to the way they exploit the information of categorical attributes in clustering, we can roughly divide them into statistical knowledge-based and learning-based approaches. Then we proceed from the perspective of how they handle ordinal attributes.

The statistical knowledge-based approaches attempt to obtain reasonable attribute representations based on the data statistics rather than the approaches that rely solely on semantic knowledge of values, e.g., conventional Hamming and order distance-based clusterings. Among this stream, entropy-based measures [10], [11] adopt statistical information entropy of the values to reflect their affiliations to the clusters. Later, more approaches [12]–[14] have been proposed on the basis of probability, adopting the common basic principle that two values with similar statistical context (e.g., occurrence frequency or conditional probability on the other attributes) should have a higher similarity. However, clustering based on all the above approaches relies fully on the data statistics but ignores

the semantic order. Therefore, some recent advances [15], [16] especially consider order information during distance measurement, thus achieving better clustering performance. Nevertheless, all the above measures work independently of the clustering, thereby limiting the clustering performance.

The learning-based approaches have thus appeared in the literature to jointly learn representations of attributes and clustering of objects. The conventional learning-based approaches either learn object-cluster similarities [17], [18] or attributes importance [19], [20] for clustering. However, they facilitate the learning based on a priori assumption on the distances among intra-attribute values, which still prevents them from approaching "true" distance representations w.r.t. clustering. Recently, a more advanced categorical data clustering approach [21] that introduces multiple kernels to comprehensively represent the attributes has been proposed. Since it does not consider ordinal attributes, the most recent approaches [22], [23] further represent and adaptively adjust the distance structures of ordinal attributes during clustering.

To the best of our knowledge, all the existing feasible solutions are based on the original semantic order, which may not suit the clustering as demonstrated by the left case in Fig. 1. Thus the original orders bottleneck the clustering performance, while the unsupervised setting conversely hinders the understanding of the "true" orders. These cross-coupled factors form the crux of the generally poor clustering performance on ordinal-attributed categorical data. Motivated by this, a method that lets the semantic order tuning and the clustering learn from each other is in urgent need.

This paper, therefore, proposes a novel method for clustering categorical data that bridges the gap between the semantic orders and the orders preferred by the clustering task. The key innovation is that we simultaneously remove the restrictions brought by the macro semantic order and the micro attribute value-level distances to the clustering through one learning paradigm. That is, we let the three objectives, i.e., (1) orders of values, (2) distances of values, and (3) partitions of objects, iteratively learn from each other through the proposed optimization algorithm. Three main contributions of this work are summarized below:

- A new paradigm, which is efficient, parameter-free, interpretable, and can be easily extended to most clustering methods for enhancement is proposed. It facilitates a high degree of freedom for learning representations of categorical data, thereby adequately eliminating the deterioration brought by the information ambiguity to clustering accuracy.
- We design an order understanding strategy to extract "hints for better orders" of each ordinal attribute from the current optimal clustering results. Through the strategy, a new order that better suits the current clusters can be inferred to provide a re-launch point that is closer to the global optimum in the next learning epoch.
- To efficiently fine-tune the newly learned order relationships, an inter-value distance learning mechanism is elegantly incorporated into the learning paradigm by concisely representing the v^2-scale distances of each attribute using v-scale weights between adjacent values (v for the number of possible values of an attribute).

II. Related Work

This section overviews existing statistical knowledge-based and learning-based categorical attribute representations for clustering, as these two topics are highly related to our work.

A. Statistical Knowledge-based Representation

Entropy-based measures [10], [11] quantify the object-cluster affiliations by adopting information entropy as a measure. [15] further preserves the order relationships of ordinal attributes by successively computing the entropy on each pair of adjacent possible values. To exploit the inter-dependence among attributes, approaches [12], [13], [24] measure the distance between two values as the differences between their corresponding conditional probability distributions reflected by the other attributes. Later, the measure [14] further selects a set of highly related attributes as context for more reasonable distance measurement. To cope with independent attributes, the recent work [16] proposes to simultaneously consider intra- and inter-attribute information, which achieves sufficient clustering performance improvement. Most recently, [25] uses entropy-based similarity and hamming distance-based metric to measure the numerical and categorical attributes.

However, all the above approaches represent object-cluster similarities independently of clustering, thus limiting the clustering performance. To address this issue, learning-based representations have been proposed in the literature.

B. Learning-based Representation

Conventional approaches [19], [26] learn the data representation from the perspective of attributes, i.e., update the weight of each attribute during clustering to obtain a better representation where the clusters appear in a more compact way. The two flexible subspace learning approaches [20], [27] learn the weights of an attribute w.r.t. different clusters. There are also approaches proposed from the perspective of objects [17], which learns object-cluster similarity based on the occurrence probability of object values in different clusters during clustering. The work [18] further considers the importance of different attributes, and achieves a better clustering performance, and [28] proposed an innovative loss function based on consensus clustering.

To more finely learn the value-level representations, [21] first represents each attribute using multiple kernels and then jointly learns the representations with clustering, which achieves very competitive clustering performance. The works [22], [29] have also been proposed to learn the inter-value distances of categorical attributes. Most recently, [23] proposes to first convert heterogeneous nominal and ordinal attributes into a homogeneous form, and then learn the inter-value distances in a decent way. However, all the learning-based approaches are still based on the semantic relationship among the possible values, which prevents them from achieving more satisfactory clustering accuracy.

III. Proposed Method

This section first introduces basic problem settings, and then presents how to infer the orders of attribute values according to a given partition of objects. Finally, the inference strategy is combined with the clustering task to form a learning paradigm.

A. Preliminaries

Given a categorical dataset denoted as $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ with data volume n, each of the n data objects is described by the values from d attributes $A = \{a_1, a_2, ..., a_d\}$. A data object $\mathbf{x}_i$ can be denoted in the form of a vector as $\mathbf{x}_i = [x_{i,1}, x_{i,2}, ..., x_{i,d}]^\top$. An attribute a_r describes a value domain $V_r = \{o_{r,1}, o_{r,2}, ..., o_{r,v_r}\}$ with v_r possible values. If a_r is an ordinal attribute, its values are with extra semantic order (also called rank hereinafter) in comparison with nominal attributes, and the order can be written as $\phi(o_{r,1}) > \phi(o_{r,2}) > ... > \phi(o_{r,v_r})$, where the function $\phi(\cdot)$ fetches the rank of a possible value by

$$\phi(o_{r,i}) = i. \tag{1}$$

Since a nominal attribute can be encoded into boolean-valued attributes by one-hot encoding, and the encoded attribute can be treated as ordinal attributes, we discuss by assuming all the attributes are ordinal hereinafter.

Partitional clustering is to divide the data object set X into k non-overlapping subsets called clusters $C = \{C^1, C^2, ..., C^k\}$. A cluster C^l can be represented by a d-dimensional vector $\mathbf{c}^l = [c_1^l, c_2^l, ..., c_d^l]^\top$ with values from the value domains corresponding to the d attributes. The general goal of clustering is to minimize the overall difference among intra-cluster data objects. The object-cluster affiliation is reflected by an $n \times k$ partition matrix $\mathbf{Q}$ with its the (i,j)th entry indicating if the ith object belongs to the jth cluster by

$$q_{i,j} = \begin{cases} 1, & \text{if } j = \arg\min\limits_{1 \le h \le k} \Gamma(\mathbf{x}_i, C^h) \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where $\Gamma(\mathbf{x}_i, C^j)$ is the dissimilarity between x_i and cluster C^j, which can be generally written as

$$\Gamma(\mathbf{x}_i, C^j) = \sum_{r=1}^{d} \gamma(x_{i,r}, c_r^j) \tag{3}$$

with $\gamma(x_{i,r}, c_r^j)$ being the distance between $\mathbf{x}_i$ and C^j reflected by attribute a_r. Accordingly, the objective function can be written as

$$E(\mathbf{Q}) = \sum_{i=1}^{n}\sum_{j=1}^{k} q_{i,j} \cdot \Gamma(\mathbf{x}_i, C^j) \quad \text{s.t.} \quad \sum_{j=1}^{k} q_{i,j} = 1. \tag{4}$$

For categorical data, distance $\Gamma(\mathbf{x}_i, C^j)$ is typically defined based on the fixed semantic order of ordinal attribute values, which bottlenecks the clustering performance. To remove such restriction, we denote the orders of each attribute as $\Phi = \{\Phi_1, \Phi_2, ..., \Phi_d\}$ where $\Phi_r = \{\phi(o_{r,1}), \phi(o_{r,2}), ..., \phi(o_{r,v_r})\}$ stores the ranking of possible values for a_r, and design a learning mechanism to make the rank values in Φ learnable with clustering tasks in the following subsections.

Algorithm 1 Order Updating of Possible Values

Input: X, $\mathbf{Q}$.
Output: Φ.

```
for r = 1, 2, ..., d do
    Obtain G^r by Eq. (5) and Eq. (6);
    set D_r = ∅;
    repeat
        Find i* and j* according to Eq. (7);
        D_r = D_r ∪ {o_{r,i*}, o_{r,j*}}, update Φ_r by inserting o_{r,i*}
        and o_{r,j*} between the last found pair;
    until V_r \ D_r = ∅;
end for
```

B. Order Inference

Our proposed approach aims to represent categorical data, thus providing more appropriate distances $\Gamma(\mathbf{x}_i, C^j)$ and $\gamma(x_{i,r}, c_r^j)$ for clustering. As discussed in Section I, orders between values of an ordinal attribute dominate the distance measurement, and thus our goal is to obtain the reasonable order of the possible values to better match the clustering task. Therefore, this subsection presents how to determine the current optimal order Φ^* based on a given partition of objects.

From the perspective of the given clusters C, two possible values are more similar if they co-occur more frequently in the same cluster, and vice versa. Accordingly, dissimilarity between a pair of possible values $o_{r,i}$ and $o_{r,j}$ is defined as

$$g_{i,j}^r = \sum_{l=1}^{k} \frac{|p_{r,i}^l - p_{r,j}^l| \cdot \sigma(C^l)}{n}, \tag{5}$$

which is specified as the (i,j)-th entry of a gap matrix $\mathbf{G}^r$ corresponding to a_r. $|p_{r,i}^l - p_{r,j}^l|$ is the difference between the occurrence probabilities of $o_{r,i}$ and $o_{r,j}$ in cluster C^l with $p_{r,i}^l$ defined as

$$p_{r,i}^l = \frac{\sigma(X_{r,i}^l)}{\sigma(C^l)}, \tag{6}$$

where $X_{r,i}^l = \{\mathbf{x}_h | x_{h,r} = o_{r,i}, \mathbf{x}_h \in C^l\}$, and $\sigma(\cdot)$ counts the cardinality of a set.

To determine the orders of the possible values of an ordinal attribute a_r, we first determine the two values $o_{r,i^*}, o_{r,j^*} \in V_r$ with the current largest difference, which satisfy

$$\begin{aligned} i^*, j^* &= \arg\max_{i,j} g_{i,j}^r \\ \text{s.t.} \quad & V_r \backslash D_r \neq \emptyset \text{ and } i, j \in \{1, 2, ..., v_r\}, \end{aligned} \tag{7}$$

where D_r stores the possible values that have been ordered. By putting the two possible values $o_{r,i}^*$ and $o_{r,j}^*$ at the two sides of the obtained order, and update D_r by $D_r = D_r \cup \{o_{r,i^*}, o_{r,j^*}\}$. We then consider the rest $v_r - 2$ possible values in $V_r \setminus D_r$ according to Eq. (7), and insert the current most different pairs into the order with updating D_r until $V_r \setminus D_r = \emptyset$ or there is only one possible value in $V_r \setminus D_r$ due to the odd number of possible values in V_r. For the latter case, the only one possible value is directly inserted between the last found pair of possible values. After determining the new orders of all

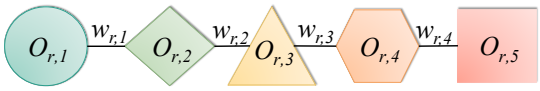

Fig. 2. Distance structure of an attribute represented by inter-value weights. In this toy example, we have the number of possible values $v_r = 5$.

the possible values of each attribute, we obtain Φ, and the above process is summarized in **Algorithm 1**. To learn Φ from data partitions, we treat it as a variable to participate in the optimization of the clustering objective in **Algorithm 2**, which will be discussed in the following subsections.

C. Distance Learning

Although we obtain new order Φ, it is still insufficient to fully utilize the information of the current object partition, and the distance structures will also change depending on Φ during clustering. Therefore, this subsection proposes an approach for learning the distance among the ordered values, and then we discuss how to learn both distances and orders during clustering in the next subsection.

To maintain an appropriate distance structure, a distance learning mechanism is derived based on a given order Φ and partition C. We introduce variables $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d\}$ to describe the distance structures of each attribute. More specifically, each $\mathbf{w}_r = [w_{r,1}, w_{r,2}, ..., w_{r,v_r-1}]^\top$ is a (v_r-1)-dimensional vector representing the weights of inter-value distances of a_r as shown in Fig. 2. By considering $\mathbf{W}$ and Φ, the new distance Γ_{w} can be written as

$$\Gamma_{\mathrm{w}}(\mathbf{x}_i, C^l; \Phi) = \sum_{r=1}^{d} \sum_{m=1}^{v_r} \gamma_{\mathrm{w}}(x_{i,r}, o_{r,m}; \Phi_r) \cdot u_{r,m}^l, \tag{8}$$

which is the distance between $\mathbf{x}_i$ and C^l. Here, $u_{r,m}^l$ is the weight of the possible value $o_{r,m}$ to cluster C^l, which is defined as

$$u_{r,m}^l = \frac{\sigma(X_{r,m}^l)}{\sigma(C^l)} \tag{9}$$

where $X_{r,m}^l = \{\mathbf{x_j} | x_{j,r} = o_{r,m}, \mathbf{x_j} \in C^l\}$ is a set of objects in C^l with their r-th values equal to $o_{r,m}$, which ranks m-th in Φ_r. Please note that once new Φ is obtained, subscripts of all the intra-attribute possible values in a value domain V_r are changed accordingly to reflect new orders in the corresponding Φ_r. In Eq. (8), $\gamma_{\mathrm{w}}(x_{i,r}, o_{r,m}; \Phi_r)$ is the distance between $x_{i,r}$ and $o_{r,m}$ according to the corresponding distance structure as shown in Fig. 2, and thus $\gamma_{\mathrm{w}}(x_{i,r}, o_{r,m}; \Phi_r)$ can be defined as

$$\gamma_{\mathrm{w}}(x_{i,r}, o_{r,m}; \Phi_r) = \begin{cases} \sum_{h=\min(\phi(x_{i,r}),m)}^{\max(\phi(x_{i,r}),m)-1} w_{r,h}, & \text{if } \phi(x_{i,r}) \neq m \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

For example, if $x_{i,r} = o_{r,2}$, i.e., $\phi(x_{i,r}) = 2$, then $\gamma_{\mathrm{w}}(x_{i,r}, o_{r,4}; \Phi_r)$ computes the distance between $o_{r,2}$ and $o_{r,4}$, which equals to $w_{r,2} + w_{r,3}$ according to the distance structure shown in Fig. 2.

As the whole distance structure can be described by $\mathbf{W}$, we update it according to the given data objects partition C by

$$w_{r,m}^{(\text{new})} = w_{r,m}^{(\text{old})} - \eta \cdot D_{r,m} \tag{11}$$

where η is a small learning rate, and $D_{r,m}$ is the overall intra-cluster distance contributed by $w_{r,m}$, which is computed by

$$D_{r,m} = w_{r,m}^{(\text{old})} \cdot \sum_{j=1}^{k} (u_{r,m}^j + u_{r,m+1}^j) \tag{12}$$

where the value of $(u_{r,m}^j + u_{r,m+1}^j)$ reflects the probability that the distance described by $w_{r,m}$ being accumulated through Eq. (10) in computing the total intra-cluster distance of C^j, i.e., total difference among intra-cluster objects of C^j. It is intuitive that the overall probability $\sum_{j=1}^{k}(u_{r,m}^j + u_{r,m+1}^j)$ multiplied by the distance $w_{r,m}$ (i.e., the distance between $o_{r,m}$ and $o_{r,m+1}$) in Eq. (12) computes the expectation of intra-cluster distance caused by $w_{r,m}$ in all the clusters. Therefore, $w_{r,m}$ with a larger contribution to the overall intra-cluster distance should be punished with greater force in Eq. (11) to ensure a steeper minimization of cluster objective, i.e., making the overall intra-cluster distance as small as possible. Moreover, after all the new weights of an attribute a_r are computed by Eq. (11), soft-max is adopted to normalize the updated weights by

$$w_{r,m} = \frac{w_{r,m}^{(\text{new})}}{\sum_{h=1}^{v_r-1} w_{r,h}^{(\text{new})}}. \tag{13}$$

The updating of $w_{r,m}$ can also be completed by using a more rigorously derived strategy in [22]. With the above weights updating strategies, we then discuss how to iteratively learn the weights with clustering.

D. Clustering Algorithm with Distance and Order Learning

The previous subsections illustrates how to reconstruct the distance structures of attributes (i.e., Φ and $\mathbf{W}$) given partition C. Our ultimate goal is to combine the order inference and distance learning processes with the clustering task to facilitate joint optimization. By combining the order adjustment in Section III-B and distance updating in Section III-C, the objective function in Eq. (4) can be rewritten as

$$E(\mathbf{Q}, \mathbf{W}, \Phi) = \sum_{i=1}^{n} \sum_{j=1}^{k} q_{i,j} \cdot \Gamma_{\mathrm{w}}(\mathbf{x}_i, C^j; \Phi). \tag{14}$$

The key process for minimizing Eq. (14) can be summarized into the following five steps: (1) Fix Φ and $\mathbf{W}$, compute $\mathbf{Q}$; (2) Fix Φ and $\mathbf{Q}$, update $\mathbf{W}$; (3) Iteratively implement (1) and (2) until $\mathbf{Q}$ remain unchanged; (4) Fix $\mathbf{Q}$ and $\mathbf{W}$, update Φ; (5) Iteratively implement (1)-(4) until Φ remain unchanged.

However, since the updating of Φ may result in new distance structures that prevent the learning from convergence, we replace Eq. (5) by

$$g_{i,j}^r = \sum_{l=1}^{k} \frac{|p_{r,i}^l - p_{r,j}^l| \cdot \sigma(C^l)}{n} \cdot |i-j|^\tau \tag{15}$$

Algorithm 2 CLORD: Clustering by Learning ORders and Distances

Input: X, k, η.
Output: $\mathbf{Q}$, $\mathbf{W}$, Φ.

Set $\Phi^{\tau} = \Phi$, $\tau = 0$, $Converge(\Phi) = False$;
while $Converge(\Phi) = False$ **do**
  Set $t = 0$, $Converge(\mathbf{Q}, \mathbf{W}) = False$; Initialize $\mathbf{W}^t$ by $w_{r,m} = \frac{1}{v_r - 1}$; Initialize $\mathbf{Q}^t$ by randomly selecting one entry from each row, and set the entry to 1;
  while $Converge(\mathbf{Q}, \mathbf{W}) = False$ **do**
    Fix Φ^{τ} and $\mathbf{W}^t$, obtain $\mathbf{Q}^{t+1}$ by Eq. (2);
    Fix Φ^{τ} and $\mathbf{Q}^{t+1}$, obtain $\mathbf{W}^{t+1}$ by Eqs. (11) and (13);
    if $\mathbf{Q}^{t+1}$=$\mathbf{Q}^t$ **then**
      $Converge(\mathbf{Q}, \mathbf{W}) = True$;
    end if
    $t = t + 1$;
  end while
  Fix $\mathbf{Q}^t$ and $\mathbf{W}^t$, obtain $\Phi^{\tau+1}$ by Algorithm 1;
  if $\Phi^{\tau+1}$=Φ^{τ} **then**
    $Converge(\Phi) = True$;
  end if
  $\tau = \tau + 1$;
end while

where τ is the number of learning iterations. The term $|i-j|^{\tau}$ gradually consolidate the previously learned orders, and thus weakens the instability caused updating of Φ.

The overall learning algorithm is summarized in **Algorithm 2**, and the complexity of the proposed method is provided in Theorem 1.

Theorem 1. *The overall time complexity of CLORD is $O(EIndk)$ at every iteration τ.*

Proof. Assuming there are n objects and d attributes in a dataset, the time complexity for obtaining $\Phi^{\tau+1}$ is $O(EndV)$, where E is the number of iterations for updating Φ, and $V = \max(v_1, v_2, ..., v_d)$ is adopted to simplify the analysis, as attributes may have different numbers of categories. Assuming I is the number of iterations to update $\mathbf{Q}^{t+1}$ and $\mathbf{W}^{t+1}$ when Φ is fixed, time complexity for obtaining $\mathbf{Q}^{t+1}$ is $O(EIndkV)$, and for obtaining $\mathbf{W}^{t+1}$ is $O(EIdkV^2)$, so the overall time complexity of CLORD is $O(EndV + EIndkV + EIdkV^2)$. Since $V << n$ and $V^2 < n$, the time complexity of CLORD can be simplified to $O(EIndk)$, which is similar to the state-of-the-art clustering methods, e.g., HD [29] and H2H [23]. □

IV. Experiments

The proposed CLORD is evaluated by comparing it with another 13 clustering methods on 10 real benchmark datasets. We introduce the experimental setup and then present the results of the designed experiments with observation analysis.

A. Experimental Setup

Comparative results are conducted from four perspectives: (1) compare CLORD with six existing methods with a significance test to statistically illustrate its superiority, (2) compare with five existing methods enhanced by the proposed core order learning mechanism to verify its scalability, (3) compare CLORD with its ablated versions to show the effectiveness of its components, and (4) compare CLORD with state-of-the-art method under different k values to demonstrate the necessity of order learning and the clustering flexibility brought by it. Moreover, changes in value ranks during clustering, execution time, cluster effect visualization, etc., are also provided to support the evaluation.

TABLE I
STATISTICS OF 10 DATASETS. n, d, AND k^* STAND FOR THE NUMBERS OF DATA OBJECTS, ATTRIBUTES (ORDINAL+NOMINAL), AND "TRUE" CLUSTERS USED FOR ALL THE EXPERIMENTS, RESPECTIVELY.

Datasets (Abbreviation)	n	d	k^*
Soybean Large (SY)	266	24+11	15
Balance scale (BS)	624	2+2	3
Fertility (FT)	100	4+3	2
Photo Evaluation (PE)	66	4+4	3
Shuttle Landing (SL)	15	5+1	2
Caesarian Section (CS)	80	3+2	2
Tic-Tac-Toe (TT)	958	9+0	2
Lense (LE)	999	4+0	5
Mammographic (MM)	824	4+0	2
Congressional Voting (VT)	434	16+0	2

The compared methods include the conventional clustering, i.e., k-means (KM), k-modes (KMD), clustering with object-cluster distance learning, i.e., OCIL [17], clustering based on the state-of-the-art distance metric UDM [16], and the most recent distance learning-based clustering approaches, i.e., HD [29] and H2H [23]. The above counterparts are chosen from different principle streams to form a more convincing performance comparison, and their hyper-parameters (if any) are set following the corresponding source papers. The learning rate η of CLORD is empirically set at 0.01. Five of the above methods are enhanced by our order learning for comparison. KMD is excluded as it treats nominal and ordinal attributes identically. Three ablated versions of CLORD are also compared, which will be introduced in the subsection "Ablation Study".

Two internal validity indices, i.e., ComPactneSs (CPS) and New Condorcet Criterion (NCC) [30], that are irrelevant to external labels and adopted distance metric have been chosen for fair evaluation, as we are doing unsupervised learning with different ks and the compared methods adopt various distance metrics that are incomparable during clustering. CPS quantifies the overall intra-cluster-object dissimilarity based on the value matching degree between two objects, and thus the lower the better. NCC simultaneously measures the intra-cluster similarity and inter-cluster dissimilarity, and thus the larger the better. Bonferroni-Dunn (BD) significance test [31] with Critical Difference (CD) interval is also adopted to provide statistical evidence for the superiority of CLORD.

Ten real datasets where PE from [22] and the remainder of nine real benchmark datasets from the UCI machine learning repository [32] are sorted out in Table I. Although the pro-

TABLE II

CPS PERFORMANCE (THE LOWER THE BETTER) COMPARISON. THE GREEN AND RED VALUES IN PARENTHESES REPRESENT THE IMPROVEMENT AND WEAKENING VALUES OBTAINED BY APPLYING OUR ORDER LEARNING MECHANISM TO THE CORRESPONDING METHOD, RESPECTIVELY.

Data	**KMD**	**KM** (Δ)	**OCIL** (Δ)	**UDM** (Δ)	**HD** (Δ)	**H2H** (Δ)	**CLORD**
SY	3.756	3.765 (+0.003)	3.823 (-0.017)	3.736 (+0.037)	3.686 (+0.038)	3.741 (-0.043)	**3.544**
BS	1.485	1.464 (+0.002)	1.457 (+0.000)	1.482 (-0.004)	1.454 (-0.005)	1.457 (-0.001)	**1.446**
FT	**1.665**	1.690 (-0.003)	1.690 (-0.002)	1.710 (-0.019)	1.687 (-0.016)	1.685 (-0.022)	1.672
PE	1.299	1.312 (+0.005)	1.311 (-0.001)	1.303 (+0.003)	1.311 (+0.003)	1.317 (-0.003)	**1.283**
SL	1.089	1.335 (-0.014)	1.318 (-0.040)	1.253 (-0.084)	1.133 (-0.053)	1.108 (-0.060)	**1.048**
CS	0.686	0.713 (-0.014)	0.722 (-0.011)	0.662 (-0.005)	0.667 (-0.028)	**0.623** (-0.005)	**0.623**
TT	2.748	2.743 (-0.014)	2.744 (-0.018)	2.781 (-0.010)	2.709 (-0.005)	2.777 (-0.007)	**2.700**
LE	1.298	1.333 (-0.002)	1.331 (-0.003)	1.341 (-0.010)	1.338 (-0.008)	1.315 (+0.005)	**1.294**
MM	0.766	0.714 (0.000)	0.714 (0.000)	0.716 (0.000)	0.737 (+0.033)	0.727 (-0.007)	**0.707**
VT	2.809	2.853 (-0.033)	2.935 (-0.085)	2.846 (-0.039)	3.161 (-0.269)	2.979 (-0.109)	**2.787**
Rank	3.80	4.95	4.85	4.80	4.20	4.30	1.10

TABLE III

NCC PERFORMANCE (THE HIGHER THE BETTER) COMPARISON. THE GREEN AND RED VALUES IN PARENTHESES REPRESENT THE IMPROVEMENT AND WEAKENING VALUES OBTAINED BY APPLYING OUR ORDER LEARNING MECHANISM TO THE CORRESPONDING METHOD, RESPECTIVELY.

Data	**KMD**	**KM** (Δ)	**OCIL** (Δ)	**UDM** (Δ)	**HD** (Δ)	**H2H** (Δ)	**CLORD**
SY	104.97	104.63 (-0.04)	104.93 (-0.37)	104.59 (-0.02)	105.87 (+0.42)	102.61 (+0.16)	**107.63**
BS	93.85	100.33 (-0.10)	100.00 (+0.11)	95.74 (+1.34)	100.09 (+0.35)	98.83 (-0.24)	**100.84**
FT	3.82	3.74 (+0.00)	3.73 (+0.00)	3.68 (+0.09)	3.76 (+0.05)	3.79 (+0.05)	**3.83**
PE	1.09	1.12 (-0.01)	1.11 (+0.00)	1.11 (-0.00)	1.10 (-0.00)	1.12 (-0.00)	**1.14**
SL	0.08	0.07 (+0.00)	0.08 (+0.00)	0.08 (+0.00)	0.08 (+0.00)	0.08 (+0.00)	**0.09**
CS	1.14	1.11 (+0.02)	1.09 (+0.03)	1.17 (+0.02)	1.17 (+0.05)	**1.24**(+0.01)	**1.24**
TT	438.90	442.71 (+2.92)	442.69 (+3.28)	428.35 (+0.06)	445.12 (+1.77)	432.37 (+2.79)	**446.44**
LE	279.98	287.05 (-0.24)	287.50 (+0.36)	275.16 (+4.07)	285.23 (+1.12)	287.92 (-1.71)	**289.83**
MM	172.46	179.76 (0.00)	179.76 (0.00)	179.62 (0.00)	176.19 (-5.04)	178.54 (+1.06)	**180.57**
VT	204.29	202.70 (+0.97)	199.38 (+3.00)	203.11 (+1.44)	190.51 (+10.21)	197.58 (+4.21)	**205.04**
Rank	4.60	4.15	4.65	5.20	4.30	4.10	1.00

posed method is discussed in terms of ordinal data, we use categorical data with mixed nominal and ordinal attributes to evaluate the proposed method in a more challenging scenario. Regarding data processing, since KMD does not have ordinal processing capabilities, the dataset is directly treated as nominal; KM and OCIL encode ordinal data into numerical data before clustering; UDM, HD, and H2H are originally proposed for mixed categorical data, so we follow the original settings. As for the CLORD method, we process nominal data using the same approach as KMD.

B. Clustering Performance Evaluation

Clustering performance evaluated by NCC and CPS is demonstrated in Tables II and III, respectively, with the result(s) of the best-performing method(s) marked in **boldface** on each dataset. The last rows in the two tables report the average ranks of the methods on all the data sets. Results in the brackets are the improvements achieved by enhancing the methods using our core order learning mechanism.

It can be observed that the proposed CLORD outperforms all the compared methods except for the CPS performance on FT dataset. CLORD outperforming all the state-of-the-art methods, i.e., UDM, HD, and H2H, further indicates its superiority in clustering categorical data.

According to the results in the brackets, we can make a statistic that our ordering mechanism successively improves the performance of the compared methods that take into account the value orders in 72 (i.e., the green results) out of 100 slots (i.e., results of 5 methods on 10 datasets evaluated by 2 validity indices). This indicates that the order learning mechanism can be easily extended to interactively learn the order of attribute values with the existing clustering methods, and effectively improve their clustering performance.

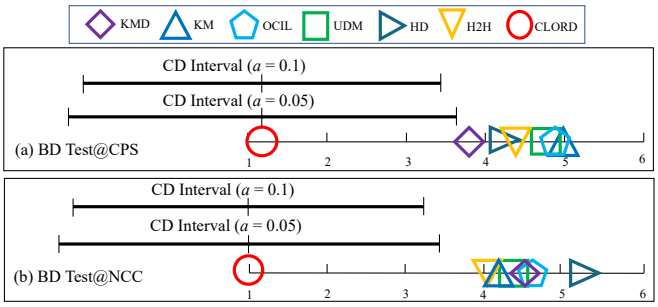

Fig. 3. Results of the two-tailed BD tests ($\alpha = 0.05$ and $\alpha = 0.1$) implemented on the "Rank" rows of Tables II and III, respectively.

C. Significance Study

To statistically illustrate the superiority of our method, we implement BD test on the "Rank" rows in Tables II and III under 95% and 90% confidence interval ($\alpha = 0.05$ and $\alpha = 0.1$). By computing the corresponding CD interval, the test results are visualized in Fig. 3. According to [31], the target method (i.e., CLORD) is believed to significantly outperform all the methods that appear out of its right-side CD interval.

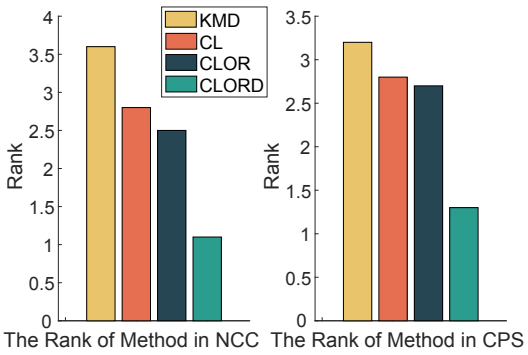

Fig. 4. Clustering performance of CLORD and its three ablated versions, i.e., CLOR, CL, and KMD.

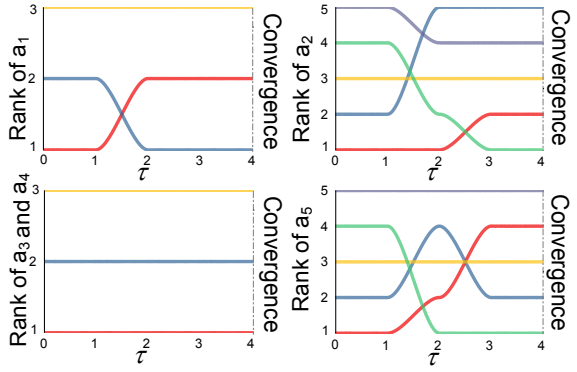

Fig. 5. Demonstration of the ordinal attribute ranks of SL dataset during the learning process of CLORD. Lines in different colors indicate the ranks of different possible values.

TABLE IV
PERFORMANCE OF H2H AND CLORD UNDER DIFFERENT kS.

Datesets		BS ($k^* = 3$)		LE ($k^* = 5$)	
Index	k	H2H	CLORD	H2H	CLORD
NCC	2	78.32	83.54↑	209.33	217.12↑
	3	98.83	100.84↑	253.48	258.60↑
	4	106.67	108.46↑	274.43	278.22↑
	5	112.09	112.82↑	287.92	289.83↑
	6	114.54	115.62↑	294.49	297.13↑
	7	116.92	117.81↑	300.73	302.48↑
	8	118.50	119.28↑	304.77	305.98↑
	9	119.58	120.27↑	307.18	308.97↑
	10	120.51	121.32↑	309.55	311.05↑
	11	120.88	121.79↑	311.07	312.66↑
	12	121.32	122.28↑	312.64	313.98↑

It can be seen that CLORD performs significantly better than all the counterparts in terms of both the two validity indices.

D. Ablation Study

The complete CLORD is compared with its three versions formed by successively removing the distance learning component (i.e., "D" of CLORD), the order learning component (i.e., "OR" of CLORD), and the component of distinguishing nominal and ordinal attributes. Accordingly, CLORD degrades to CLOR, CL, and KMD, respectively, where CLOR only learns the value orders without learning their distances, CL adopts the ranks to encode the ordinal attribute values into consecutive integers and adopts Hamming distance for the nominal attributes, and KMD indicates that when CL treats both ordinal and nominal attribute as nominal ones, it degrades to the conventional KMD. Performance of the above four versions of CLORD are evaluated by both CPS and NCC on all the datasets, and their average ranks are reported in Fig. 4.

It can be seen that the performance of CLORD, CLOR, CL, and KMD becomes worse in turn. The fact that CLOR performs worse than CLORD proves the rationality of distance learning; CLOR outperforming CL indicates the correctness of the learned orders; KMD performing worse than CL verifies the necessity of distinguishing ordinal and nominal attributes.

E. Performance under Different Numbers of Clusters

To illustrate the flexibility of CLORD in learning representations for different clustering tasks, we compare CLORD with the state-of-the-art H2H that does not learn the orders during clustering under different number of clusters, k, where $k = \{k', k'+1, k'+2, ..., k'+10\}$ with $k' = \max(k^* - 5, 2)$. It can be seen from the results on BS and LE datasets in Table IV that CLORD always outperforms H2H. This indicates that CLORD is more flexible as it customizes both the orders and distances among attribute values for each given k. This property also makes CLORD a more promising clustering solution in real applications where k is usually subject to different analysis purposes and users.

F. Convergence Visualization

As the core of this work is the learning of orders of attribute values, we visualize the order changing of different attributes of SL datasets in Fig. 5 to provide an impression about the learning process. The horizontal and vertical axes represent the number of iterations τ and the ranks of ordinal attribute values, respectively. The dashed vertical lines indicate the convergence iteration of CLORD. We visualize the same ranks of values together (i.e., a_3, a_4 in the SL dataset), which remain unchanged during learning. It can be observed that CLORD converges very quickly, i.e., within 4 iterations on all the datasets. Moreover, the ranks of values fluctuate very little and the overall change is monotonous during the learning, which intuitively indicates the reasonableness of the proposed order learning mechanism.

V. CONCLUSION

In this paper, an intuitive but yet-to-be-concerned semantic order mismatch phenomenon that brings performance bottlenecks for categorical data clustering has been studied and elaborately addressed. To eliminate the mismatch in the clustering task without discarding relevant semantic information, we design a new learning paradigm to gradually tune the semantic order, which iteratively adjusts the previous order and learns distances among ordered values based on the current optimal

clustering result. It turns out that the proposed method is robust to different clustering tasks, i.e., either clustering different datasets or clustering the same dataset with different sought numbers of clusters k. Time complexity analysis and comprehensive experiments illustrate the promising performance of the proposed method.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62102097, 62306181, and 62376233, the NSFC/Research Grants Council (RGC) Joint Research Scheme under the grant N_HKBU214/21, the Natural Science Foundation of Guangdong Province under grants: 2024A1515010163, 2023A1515012855, and 2022A1515011592, the General Research Fund of RGC under grants: 12201321, 12202622, and 12201323, the RGC Senior Research Fellow Scheme with grant: SRFS2324-2S02, and the Science and Technology Program of Guangzhou under grant 202201010548.

REFERENCES

[1] Yiu-Ming Cheung. k-means: A new generalized k-means clustering algorithm. *Pattern Recognition Letters*, 24(15):2883–2893, 2003.

[2] Liang Bai, Xueqi Cheng, Jiye Liang, Huawei Shen, and Yike Guo. Fast density clustering strategies based on the k-means algorithm. *Pattern Recognition*, 71:375–386, 2017.

[3] Mingjie Zhao, Yiqun Zhang, Yuzhu Ji, and Yang Lu. Unsupervised concept drift detection via imbalanced cluster discriminator learning. In *Proceedings of the 6th Chinese Conference on Pattern Recognition and Computer Vision*, pages 31–43. Springer, 2023.

[4] Lang Zhao, Yiqun Zhang, Yuzhu Ji, An Zeng, Fangqing Gu, and Xiaopeng Luo. Heterogeneous drift learning: classification of mix-attribute data with concept drifts. In *Proceedings of the 9th International Conference on Data Science and Advanced Analytics*, pages 1–10. IEEE, 2022.

[5] Shenghong Cai, Yiqun Zhang, Xiaopeng Luo, Yiu-ming Cheung, Hong Jia, and Peng Liu. Robust categorical data clustering guided by multi-granular competitive learning. In *Proceedings of the 44th International Conference on Distributed Computing Systems*, pages 1–12. IEEE, 2024.

[6] Fangqi Nie, Pengcheng Yan, Yiqun Zhang, Fangqing Gu, Yang Lu, and Yue Zhang. Space2: dual space learning for categorical data clustering. In *Proceedings of the 19th International Conference on Computational Intelligence and Security*, pages 1–5. IEEE, 2023.

[7] Madhavi Alamuri, Bapi Raju Surampudi, and Atul Negi. A survey of distance/similarity measures for categorical data. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, pages 1907–1914. IEEE, 2014.

[8] Yiqun Zhang and Yiu-Ming Cheung. Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6530–6544, 2023.

[9] Richard G Wilkinson. Income distribution and life expectancy. *BMJ: British Medical Journal*, 304(6820):165, 1992.

[10] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, page 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[11] Daniel Barbará, Yi Li, and Julia Couto. Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, pages 582–589, 2002.

[12] Si Quang Le and Tu Bao Ho. An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 26(16):2549–2557, 2005.

[13] Amir Ahmad and Lipika Dey. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*, 28(1):110–118, 2007.

[14] Dino Ienco, Ruggero G Pensa, and Rosa Meo. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1–25, 2012.

[15] Yiqun Zhang, Yiu-Ming Cheung, and Kay Chen Tan. A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):39–52, 2020.

[16] Yiqun Zhang and Yiu-Ming Cheung. A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Cybernetics*, 52(2):758–771, 2022.

[17] Yiu-ming Cheung and Hong Jia. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8):2228–2238, 2013.

[18] Hong Jia, Yiu-ming Cheung, and Jiming Liu. A new distance metric for unsupervised learning of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 27(5):1065–1079, 2015.

[19] Elaine Y Chan, Wai Ki Ching, Michael K Ng, and Joshua Z Huang. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*, 37(5):943–952, 2004.

[20] Hong Jia and Yiu-Ming Cheung. Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3308–3325, 2017.

[21] Chengzhang Zhu, Longbing Cao, and Jianping Yin. Unsupervised heterogeneous coupling learning for categorical representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):533–549, 2022.

[22] Yiqun Zhang and Yiu-ming Cheung. An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 6869–6876, 2020.

[23] Yiqun Zhang, Yiu-ming Cheung, and An Zeng. Het2hom: representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 1–8, 2022.

[24] Lang Zhao, Yiqun Zhang, Xiaopeng Luo, Yue Zhang, Yiu-Ming Cheung, and Kangshun Li. Selecting heterogeneous features based on unified density-guided neighborhood relation for complex biomedical data analysis. In *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine*, pages 771–778. IEEE, 2023.

[25] Naoki Masuyama, Yusuke Nojima, Hisao Ishibuchi, and Zongying Liu. Adaptive resonance theory-based clustering for handling mixed data. In *2022 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2022.

[26] Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

[27] Lifei Chen, Shengrui Wang, Kaijun Wang, and Jianping Zhu. Soft subspace clustering of categorical data with probabilistic distance. *Pattern Recognition*, 51:322–332, 2016.

[28] Jayanth Reddy Regatti, Aniket Anand Deshmukh, Eren Manavoglu, and Urun Dogan. Consensus clustering with unsupervised representation learning. In *2021 International Joint Conference on Neural Networks*, pages 1–9. IEEE, 2021.

[29] Yiqun Zhang and Yiu-ming Cheung. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3560–3576, 2022.

[30] T.R. Santos and Luis E. Zárate. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3):1247–1260, 2015.

[31] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[32] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.