

Break the Tie: Learning Cluster-Customized Category Relationships for Categorical Data Clustering

Mingjie Zhao¹, Zhanpei Huang², Yang Lu³, Mengke Li⁴,
Yiqun Zhang^{2*}, Weifeng Su⁵, Yiu-ming Cheung^{1*}

¹Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

²School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China

³School of Informatics, Xiamen University, Xiamen, China

⁴College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

⁵Guangdong Key Lab of AI and Multi-Modal Data Processing, BNU-HKBU United International College, Zhuhai, China
mjzhao@comp.hkbu.edu.hk, 2112405010@mail2.gdut.edu.cn, luyang@xmu.edu.cn, mengkeli@szu.edu.cn,

yqzhang@gdut.edu.cn, wfsu@uic.edu.cn, ymc@comp.hkbu.edu.hk

Abstract

Categorical attributes with qualitative values are ubiquitous in cluster analysis of real datasets. Unlike the Euclidean distance of numerical attributes, the categorical attributes lack well-defined relationships of their possible values (also called categories interchangeably), which hampers the exploration of compact categorical data clusters. Although most attempts are made for developing appropriate distance metrics, they typically assume a fixed topological relationship between categories when learning distance metrics, which limits their adaptability to varying cluster structures and often leads to suboptimal clustering performance. This paper, therefore, *breaks the intrinsic relationship tie* of attribute categories and learns customized distance metrics suitable for flexibly and accurately revealing various cluster distributions. As a result, the fitting ability of the clustering algorithm is significantly enhanced, benefiting from the learnable category relationships. Moreover, the learned category relationships are proved to be Euclidean distance metric-compatible, enabling a seamless extension to mixed datasets that include both numerical and categorical attributes. Comparative experiments on 12 real benchmark datasets with significance tests show the superior clustering accuracy of the proposed method with an average ranking of 1.25, which is significantly higher than the 5.21 ranking of the current best-performing method. Code and extended version with detailed proofs are provided below.

Code & Datasets —

<https://github.com/ZHAO-Mingjie/SCOF>

Extended Version — https://github.com/ZHAO-Mingjie/SCOF/Completed_verison

1 Introduction

Clustering analysis on real data usually deals with categorical attributes characterized by qualitative values without a well-defined category relationship like the Euclidean distance of numerical data (Agresti 2003; Zhang et al. 2025c;

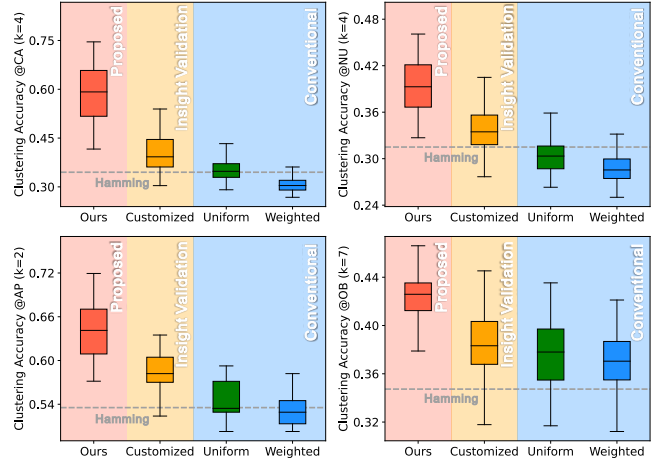


Figure 1: Clustering accuracy of our methods and three different distance measurement strategies using k -modes: Customized, Uniform, and Weighted on four datasets CA, NU, AP and TA in Table 1. Additionally, Hamming is also included as a baseline. k indicates the number of clusters. Our method outperforms other strategies on these datasets. The “Customized” strategy achieved better results than “Uniform” and “Weight” because it assigns customized category relationships to different clusters.

Zhao et al. 2022; Zhang et al. 2025a). For instance, the distance between “lawyer” and “driver” in an attribute “Occupation” of a categorical dataset is not universally defined as the distance between 0.5 and 1.2 of numerical data. Since distance metric is the basis for most clustering algorithms, lacking suitable distance metrics may surely prevent the existing clustering techniques designed for numerical data from achieving satisfactory performance on categorical data. Moreover, additionally considering the definition of an appropriate distance metric brings greater challenges to categorical data clustering (Johnson and Albert 2006; Fernandez-Navarro et al. 2013; Feng et al. 2025).

To uncover the implicit distance relationship of the pos-

*Corresponding Author

sible values of categorical data (Agresti 2010; Jain, Murty, and Flynn 1999; dos Santos and Zárate 2015), some research primarily focuses on developing distance measures based on value-mapping techniques (Hamming 1950), information entropy (Lin 1998; Zhang and Cheung 2022b; Zhang, Cheung, and Tan 2020), significance test (Hu et al. 2025e,b,a,d), and statistical characteristics (Ahmad and Dey 2007; Le and Ho 2005; Jian et al. 2018b; Ienco, Pensa, and Meo 2012; Bai and Liang 2022). However, they are all task-independent and thus incapable of adapting to different clustering tasks across various categorical datasets. To address this limitation, distance learning methods further connect the distance to downstream clustering tasks using approaches such as similarity learning based on the distribution probabilities of attribute values (Cheung and Jia 2013), kernel space mapping (Zhu, Cao, and Yin 2022), and graph-based learning (Zhang and Cheung 2022a, 2023; Zhao et al. 2024). While they achieve competitive clustering performance, the learned category relationships remain uniform for the global cluster distribution. This approach neglects the subspace specificity of clusters, i.e., some samples may exhibit a compact clustering effect in a certain subspace spanned by an attribute subset.

Although some subspace categorical data clustering methods (Bai et al. 2011; Cao et al. 2013; Jia and Cheung 2017) are proposed considering the attribute-cluster relationship recently, they primarily focus on learning the attributes’ importance while still adopting a uniform category relationship among the possible values of each attribute across all the clusters. In fact, the same attribute may exhibit diverse category relationships across different clusters (i.e., conventional part of **Figure 1**). For example, when considering the attribute “occupation” (a_o) in a pneumonia patient dataset, the category relationship of a_o ’s possible values {programmer, construction worker, driver, ...} differs in two diagnosis clusters, one for COVID-19 (C_{co}), and one for pneumoconiosis (C_{pn}). That is, for cluster C_{co} , the occupational difference between “programmer” and “construction worker” is relatively small, as their chances of contracting the highly contagious COVID-19 are comparable, while for cluster C_{pn} , the difference between these two occupations is huge because “construction worker”’s prolonged exposure to dust-laden environments significantly increases their likelihood of developing pneumoconiosis compared to programmers. Therefore, for categorical data, whose values are inherently more vague compared to numerical data, it is promising to define subspace category relationships for each attribute to better serve the exploration of different clusters.

To validate the effectiveness of the subspace category relationship for categorical data clustering, in **Figure 1**, the k -modes (Huang 1998) clustering accuracy corresponding to the following distance measurement strategies is plotted using box plots: 1) Customized: k -modes + random Customized category relationships, i.e., for a given attribute, randomly generate a series of category relationships for different clusters; 2) Uniform: k -modes + random Uniform category relationship; 3) Weighted: k -modes + Hamming distance + random Weighted attributes per the cluster; Additionally, k -modes + Hamming distance is also included as a baseline. The category relationships of the Customized,

Uniform, and Weighted are randomly generated and are executed 100 times. It can be seen that: 1) Customized outperforms the others, which demonstrates that clustering-customized category relationship brings better clustering results than uniform category relationships. 2) Uniform clearly surpasses Weighted and Hamming, which indicates Intra-attribute distance relationship modeling outperforms attribute weighting. Therefore, obtaining the optimal subspace category relationship i.e., learning clustering-customized relationship among intra-attribute values for each cluster, is the key to achieving accurate categorical data clustering.

In this paper, a new learning paradigm for categorical data clustering that learns clustering-customized category relationships from cluster (DISC) is proposed. The learning process is no longer constrained to adjusting attribute weights for clusters under uniform and fix category relationships. Instead, category relationship is initially assigned to all clusters, which then evolves into different structures for each cluster during the learning process. This allows the relationships between attribute values to be adaptively captured under a high degree of freedom, ensuring that the learned distance relationships are meaningful and tailored to improve the quality and compactness of the clusters, thus enhancing clustering performance, as can be seen in **Figure 1**. We model the distance relationship of intra-attribute possible values as a fully connected graph, and infer a tree-like category relationship to concisely and exactly reflect the relationship among attribute values. As a result, each attribute has a series of category relationships tailored for different clusters, significantly enhance the fitting ability for cluster optimization. Moreover, it is theoretically and empirically demonstrated that the learned category relationships exhibit strong compatibility with the Euclidean distance, enabling a natural extension of the method to mixed datasets, which include both numerical and categorical attributes, and resulting in enhanced clustering performance. The main contributions of this work are summarized into three-fold:

- **Revealing the effect of category relationships in clustering:** The coupling between clustering and category relationships is studied by extending the subspace clustering principles, and it is validated that optimal category relationship is the foundation for accurate clustering.
- **Proposing a novel paradigm for category relationships learning:** A novel learning strategy is proposed to modeling and tuning the category relationships of attributes. It infers more appropriate category relationships of intra-attribute values to suit current partition results, effectively improving the clusters through distance-cluster joint optimization.
- **Providing a rigorous metric with convergence guarantee:** A series of theoretical proofs are provided to rigorously demonstrate that the joint learning process of subspace category relationship exhibits strong compatibility with Euclidean distance, along with high efficiency and convergence.

2 Related Work

Distance Measures for categorical data are generally categorized into encoding-based and direct measurement approaches. Early methods, such as one-hot encoding with Hamming distance (Hamming 1950), fail to capture the full spectrum of dissimilarity between intra-attribute values. Subsequent methods (Ahmad and Dey 2007; Le and Ho 2005; Jian et al. 2018b) refine distance discrimination but still focus on pairwise distances without considering overall category relationships. Information entropy-based approaches (Lin 1998; Zhang, Cheung, and Tan 2020) improve this by better capturing intra-attribute relationships, while more recent methods (Cai et al. 2024; Zhang and Cheung 2022b; Zhang, Cheung, and Zeng 2022; Zhang et al. 2024) attempt to place attribute values along a distance axis, offering a linear representation similar to numerical attributes. However, these methods remain task-independent and may not adapt well to different clustering tasks.

Distance Learning combines metric optimization with clustering through iterative learning. Early work (Cheung and Jia 2013) focus on modeling sample-cluster similarities using probability distributions within clusters. Later methods, such as kernel-based techniques (Zhu, Cao, and Yin 2022) and joint embedding of numerical and categorical attributes (Jian et al. 2018a; Zhu et al. 2020; Chen et al. 2024), further refine distance metrics. More recent advances (Zhang and Cheung 2020, 2022a, 2023; Zhao et al. 2024) unify metric learning with adaptive graph weighting, optimizing both cluster structures and distance functions simultaneously.

Subspace Learning focuses on identifying attribute subsets that enhance clustering. Early methods (Bai et al. 2011) fix attribute weights, limiting adaptability, while subsequent approaches (Cao et al. 2013) iterate on weight and cluster assignments. Methods like (Jia and Cheung 2017) integrate attribute weighting for mixed data clustering, and extensions (Zhang et al. 2025b) unify categorical and numerical attributes via geometric projections, aligning latent dimensions across attribute types.

These advances improve clustering performance, but the learned category relationships remain uniform across the global cluster distribution, neglecting subspace-specific variations. This limits their capability in representing complex relationships between categorical attribute values.

3 Proposed Method

In this section, we first formulate the problem (§ 3.1). Then the subspace category relationship (§ 3.2) is then inferred from the modeled value-level relation structure, and the corresponding distance metric (§ 3.3) is defined. After that, a joint learning (§ 3.4) scheme is introduced to make the distance learnable with clustering.

3.1 Problem Formulation

The problem of categorical data clustering with subspace category relationship learning is formulated as follows. Given a categorical dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ with n data samples. Each data sample can be denoted as an l -dimensional row vector $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,l}]^\top$ repre-

sented by l attributes $A = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l\}$. Each attribute \mathbf{a}_r can be denoted as a column vector $\mathbf{a}_r = [x_{1,r}, x_{2,r}, \dots, x_{n,r}]$ composed of the r -th values of all n samples. The n values can be viewed as sampled from a limited number of possible values $V_r = \{v_r^1, v_r^2, \dots, v_r^{o_r}\}$ with o_r indicating the number of possible values of \mathbf{a}_r . The goal of our cluster-customized category relationships clustering can be formalized as minimizing the intra-cluster dissimilarity:

$$z(\mathbf{H}, \mathbf{M}, \mathbf{G}) = \sum_{j=1}^k \sum_{i=1}^n h_{i,j} \cdot \Phi(\mathbf{x}_i, \mathbf{m}_j), \quad (1)$$

where \mathbf{H} , \mathbf{M} and \mathbf{G} are a $n \times k$ cluster partition matrix, a $k \times l$ centroid matrix, and a matrix containing $k \times l$ cluster-customized category relationships (also called subspace category relationships) represented as graphs, respectively. Specifically, cluster partition \mathbf{H} is an affiliation matrix with its (i, j) -th entry $h_{i,j}$ indicating the affiliation between sample \mathbf{x}_i and cluster C_j , which can be written as:

$$h_{i,j} = \begin{cases} 1, & \text{if } j = \arg \min_y \Phi(\mathbf{x}_i, \mathbf{m}_y) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

with $\sum_{j=1}^k h_{i,j} = 1$. $\Phi(\mathbf{x}_i, \mathbf{m}_y)$ is the attribute-level dissimilarity between sample \mathbf{x}_i and cluster C_y described by the cluster center \mathbf{m}_y . The value of \mathbf{m}_y is computed following the way of the conventional k -modes clustering algorithm, i.e., the value of $m_{y,r}$ is equal to the mode from \mathbf{a}_r in C_y . The attribute-level dissimilarity is computed in the context of subspace category relationships can be written as:

$$\Phi(\mathbf{x}_i, \mathbf{m}_y) = \sum_{r=1}^l \phi(x_{i,r}, m_{y,r}; G_{y,r}), \quad (3)$$

where $G_{y,r} \in \mathbf{G}$ is the category relationship of r -th attribute in y -th cluster, and $\phi(x_{i,r}, m_{y,r}; G_{y,r})$ is the corresponding sample-cluster dissimilarity.

This paper focuses on defining the category relationship within each cluster to facilitate a more thorough optimization through clustering-customized distance metric learning. Therefore, the key is how to reasonably define the distance $\phi(x_{i,r}, m_{y,r}; G_{y,r})$ under an appropriate category relationship $G_{y,r}$ in the attribute aspect.

3.2 Subspace Category Relationship Inference

To infer subspace category relationships for each cluster, we first model the value-level relation through a fully-connected graph as it can flexibly reflect any relationship among possible values. A graph $G_{j,r}$ indicating the category relationship of r -th attribute in j -th cluster is denoted as $G_{j,r} = \langle V_r, B_r, \mathbf{W}_{j,r} \rangle$, where V_r includes o_r possible values as nodes. B_r is the set of $o_r(o_r - 1)/2$ edges connecting all the o_r nodes. $\mathbf{W}_{j,r}$ is the weight matrix reflecting edge lengths.

Given cluster partition \mathbf{H} , for each cluster, the weight between two nodes v_r^u and v_r^s within a cluster C_j is defined as the difference between their Conditional Probability Distributions (CPDs) extracted from the current cluster by:

$$\mathbf{W}_{j,r}(u, s) = |p(v_r^u | C_j) - p(v_r^s | C_j)|, \quad (4)$$

where $\mathbf{W}_{j,r}(u, s)$ is the (u, s) -th element of weight matrix $\mathbf{W}_{j,r}$, and $p(v_r^u|C_j)$ is the conditional probability of v_r^u given cluster C_j , which can be computed as:

$$p(v_r^u|C_j) = \frac{\text{card}(\{\mathbf{x}_i \in C_j | \mathbf{x}_{i,r} = v_r^u\})}{\text{card}(C_j)}, \quad (5)$$

where $\text{card}(\cdot)$ counts the number of samples in a set. A condition probability distribution $p(v_r^u|C_j)$ describes the distribution pattern of a value within C_j . Consequently, two values with similar patterns are considered to be more similar, and thus the weight of the edge connecting them is smaller. Although the weighted fully connected graph provides a flexible way to represent the relationships among the intra-attribute values in each cluster, it introduces uncertainties in the distance representation because the relationship between a pair of values can be described by multiple paths.

To refine the graph, and consider that the common definition of distance between two samples is based on the shortest path length (Bondy and Murty 2008; Munkres 2000), we extract the shortest paths that can connect all the possible values (i.e., all nodes in $G_{j,r}$) to form the Minimum Spanning Tree (MST) $\mathcal{T}_{j,r}$ (called relation tree hereinafter) by:

$$\mathcal{T}_{j,r} = \arg \min_{\mathcal{E}_t \in B_r} \sum_{b_r^{u,s} \in \mathcal{E}_t} \mathbf{W}_{j,r}(u, s), \quad (6)$$

where \mathcal{E}_t is a subset of B_r . \mathcal{E}_t contains $o_r - 1$ edges connecting all nodes V_r . $b_r^{u,s}$ from B_r is the weighted edge connect nodes v_r^u and v_r^s . Consequently, for a cluster C_j , a series of relation trees $\mathcal{T}_j = \{\mathcal{T}_{j,1}, \mathcal{T}_{j,2}, \dots, \mathcal{T}_{j,r}\}$ is inferred corresponding to all the attributes, and the subspace category relationship of all clusters can be represented as a set $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$ corresponding to graph set G .

Theorem 1. *The relation tree $\mathcal{T}_{j,r}$ defines a deterministic and Euclidean-compatible distance metric over the categorical attribute values.*

Proof. It is proved that the relation tree $\mathcal{T}_{j,r}$, inferred from the fully connected graph $G_{j,r}$, retains all pairwise distances among nodes, establishing its determinism and uniqueness. Then, it is demonstrated that the distance metric induced by the relation tree always forms a linear structure, naturally compatible with the one-dimensional Euclidean geometry. A detailed proof can be found in **Appendix A.1**. \square

Remark 1 (Determinism and Euclidean Compatibility of the Inferred Relation Tree). *The inferred relation tree exhibits determinism and strong compatibility with Euclidean space (as proven in **Theorem 4**). The determinism of the relation tree ensures that the learned distance metric is stable and interpretable, providing clarity and eliminating the need for users to manually define or adjust category relationships. The compatibility with Euclidean space allows for easy extension to mixed datasets, making the method adaptable to both categorical and numerical attributes. This is further validated in § 4.3, where our method achieves performance improvement and demonstrates superior results compared to other clustering methods.*

3.3 Clustering-Customized Subspace Distance

Given the subspace category relationship \mathcal{T} , we further tune them into customized distance metrics by modeling the distance between intra-attribute values to capture more compact clusters within subspaces. Specifically, for cluster C_j , given the relation tree $\mathcal{T}_{j,r}$ corresponding to attribute \mathbf{a}_r , the distance between values v_r^u and v_r^s can be computed by:

$$\mathbf{D}_{j,r}(u, s) = \sum_{\mathbf{w}_{j,r}(u, s) \in T} \mathbf{W}_{j,r}(u, s), \quad (7)$$

where \mathbf{D} is a $o_r \times o_r$ distance matrix of possible values, and T is the set of edges splicing the shortest path from v_r^s to v_r^u in $\mathcal{T}_{j,r}$. Since this is computed based on the weights defined based on the intra-cluster statistics by **Eq. (4)**, it is referred to as the clustering-customized subspace distance.

Remark 2 (Generalized Attribute Weighting through Relation Tree). *The category relationship represented by the relation tree implicitly captures the importance of each attribute through a generalized attribute weighting mechanism. Specifically, the clustering-customized subspace distance model adapts to variations in intra-attribute distances, allowing for different magnitudes of distances between attribute values across different attributes. As indicated by Eqs (4) and (7), a larger difference in attribute values leads to a more concentrated distribution of the r -th attribute within the cluster, making it more likely to be considered significant. This variability in the distribution of attribute values ensures that each attribute contributes differently when calculating attribute-level distances, thereby reflecting its importance within the overall category relationship. This approach can be viewed as a form of generalized attribute weighting, where the learned distances effectively decompose the attribute importance into a more detailed and dynamic layer of distance learning.*

Accordingly, the sample-cluster distance $\phi(x_{i,r}, m_{j,r})$ reflected by $\mathcal{T}_{j,r}$ for the r -th attribute can be defined as:

$$\phi(x_{i,r}, m_{j,r}; \mathcal{T}_{j,r}) = \mathbf{D}_{j,r}(u, s), \quad (8)$$

where sample value $x_{i,r}$ and center value $m_{j,r}$ are assumed be equal to the possible values v_r^u and v_r^s , respectively. The overall sample-cluster distance $\Phi(\mathbf{x}_i, \mathbf{m}_j)$ is defined as:

$$\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j) = \sum_{r=1}^l \phi(x_{i,r}, m_{j,r}; \mathcal{T}_{j,r}). \quad (9)$$

Theorem 2. *The sample-cluster distance $\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j)$ defined in the context of \mathcal{T}_j represents a valid distance metric.*

Proof. It is proved that the distance measure $\mathbf{D}_{j,r}(u, s)$ defined within the context of the relation tree $\mathcal{T}_{j,r}$ constitutes a valid distance metric. Then, the sample-cluster distance $\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j)$ is shown to be a weighted sum of $\mathbf{D}_{j,r}(u, s)$ with non-negative weights, also satisfies the conditions of non-negativity, symmetry, and triangle inequality, thereby forming a valid distance metric. The detailed proof can be found in **Appendix A.2**. \square

It can be seen that the cluster-customized distance satisfies the definition of a valid distance metric, which ensures consistent and reliable distance calculations across different clusters, thereby improving clustering accuracy.

3.4 Joint Learning of Distance and Clusters

To jointly optimize the distance metric and sample partition for better mutual adaptation, the subspace category relationship inference is integrated into the clustering objective, which can be written based on Eqs. (1), (3), and (8):

$$z(\mathbf{H}, M, \mathcal{T}) = \sum_{j=1}^k \sum_{i=1}^n h_{i,j} \cdot \sum_{r=1}^l \mathbf{D}_{j,r}(u, s), \quad (10)$$

where we assume that $x_{i,r} = v_r^u$ and $m_{j,r} = v_r^s$. Then the problem can be formulated as computing \mathbf{H} , M , and \mathcal{T} to minimize z , which can be summarized into the following three steps: 1) Fix M and \mathcal{T} , compute \mathbf{H} ; 2) Fix \mathcal{T} and \mathbf{H} , compute M ; 3) Repeat 1) and 2) until convergence, fix \mathbf{H} and M and reinfer \mathcal{T} according to Eqs. (4), (5), and (7). These three steps are repeated until the \mathbf{H} no longer changes. The algorithm is summarized as **Algorithm 1** in Extended Version, which can be proven to guarantee convergence and is with complexity of $O(nlkIE)$, linear to n , l , and k . (Proof in **Appendix A.3** and **Appendix A.4**).

Theorem 3. *DISC algorithm converges to a local minimum in a finite number of iterations.*

Proof. The inner loop is proved to attain a local minimum of $z(\mathbf{H}, M, \mathcal{T})$ within a finite number of iterations, when \mathcal{T} is fixed. Then, the outer loop E is proved to converges to a local minimum of the objective function $z(\mathbf{H}, M, \mathcal{T})$ in a finite number of iterations. The detailed proof can be found in **Appendix A.3**. \square

4 Experiment

The experimental setup is first introduced in § 4.1. To comprehensively validate the effectiveness and superiority of our method, four experiments have been conducted. § 4.2: Clustering performance evaluation; § 4.3: Ablation studies; § 4.4: Convergence and efficiency evaluation; § 4.5: Qualitative study of the clustering effect. The experiments are conducted on an Intel i7-13500H@2.60GHz workstation with 16G RAM and MATLAB R2022b.

4.1 Experimental Setup

10 counterparts are compared: KMD (Huang 1998) is a traditional method, CBDM (Ienco, Pensa, and Meo 2009), WOCIL (Jia and Cheung 2017), CURE (Jian et al. 2018b), CDCDR (Bai and Liang 2022), and HDC (Zhang and Cheung 2022a) are subspace clustering baselines, while CoForest (Zhao et al. 2024), MCDC (Cai et al. 2024), SigDT (Hu et al. 2025c), and HARR (Zhang et al. 2025b) are State-Of-The-Art (SOTA) distance metric learning and statistical-based methods. Each method is implemented 10 times and the average performance is reported.

12 Datasets from various domains are utilized for the experiments. All the datasets are real public datasets collected from the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017), and their statistical information is shown in **Table 1**. For all the counterparts, we set k at k^* , i.e., the true number of clusters given by the dataset labels.

Table 1: Dataset statistics. l , n , and k^* are the numbers of attributes, samples, and true number of clusters, respectively.

No.	Dataset	Abbrev.	l	n	k^*
1	Car Evaluation	CA	6	1728	4
2	Soybean (large)	SB	35	266	15
3	Nursery School	NU	8	12960	4
4	Contraceptive Choice	CC	7	1473	3
5	Amphibians	AP	12	189	2
6	Dermatology	DT	33	366	6
7	Auction Verifications	AV	6	2043	2
8	Obesity Levels	OB	8	2111	7
9	Teaching Assistant	TA	4	151	3
10	Bank Marketing	BM	9	45211	2
11	Congressional Voting	CV	16	435	2
12	Zoo	ZO	16	101	7

Three validity indices, including clustering accuracy (ACC) (He, Cai, and Niyogi 2005) and ARI coefficients (Rand 1971; Gates and Ahn 2017). To ensure a fair comparison, an entropy-based clustering CoMPactness (CMP) measure is introduced to eliminate the influence of different metric magnitudes, which is computed:

$$\text{CMP} = \frac{1}{lk} \sum_{j=1}^k \sum_{r=1}^l \sum_{u=1}^{o_r} \frac{-p(v_r^u | C_j) \log p(v_r^u | C_j)}{\log o_r}. \quad (11)$$

The CMP quantifies attribute value concentration within clusters and approaches zero for highly concentrated attribute values, indicating cluster compactness.

4.2 Clustering Performance Evaluation

The clustering performance of various methods is evaluated using both external (ACC and ARI) and internal (CMP) validation indices, as shown in **Table 2**. External validation measures how well the clustering matches the true labels, while internal validation assesses the compactness and distinguishability of the clusters.

External (label-based) validation: Overall, DISC outperforms most methods, as shown in Table 2. The observations include: 1) DISC consistently performs well, demonstrating its clustering superiority across all datasets. 2) On the ACC index, DISC is slightly behind the best on SB, NU, and CV datasets (differences under 0.015), still proves the competitiveness of DISC. 3) DISC performs similarly to the best methods on the ARI index, with small differences on AV, BM, and CV datasets. 4) SigDT performs well in ACC but poorly in ARI on CA and NU datasets. The reason is that it over-aggregates data into a single large cluster, inflating ACC while failing to capture inter-cluster differences, thus lowering ARI. In contrast, DISC maintains a higher ARI, demonstrating its ability to preserve true cluster structures.

Internal validation: As can be seen in **Table 2**, DISC achieves the best rank in the CMP index, indicating superior cluster compactness and distinguishability. It can be observed that DISC performs worse than SigDT in some cases (e.g., on AP, AV, BM and CV datasets). This is because on these datasets, SigDT tends to partition samples into excessively small clusters, which may surely demonstrate higher compactness but is relatively incompetent in

Table 2: Clustering performance evaluated by the external ACC and ARI, and internal CMP. Larger values of ACC (ranging from 0 to 1) and ARI (ranging from -1 to 1) and smaller values of CMP (ranging from 0 to 1) indicate better clustering performance. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively.

Indices	Data	KMD [DMKD'98]	CBDM [TKDD'12]	WOCIL [TNNLS'17]	CURE [TKDE'18]	CDCDR [PR'22]	HDC [TPAMI'22]	CoForest [ECAI'24]	MCDC [ICDCS'24]	SigDT [In Sci'25]	HARR [ESWA'25]	DISC (ours)
ACC	CA	0.3795±0.04	-	0.3665±0.06	0.4174±0.05	0.3376±0.06	0.3800±0.04	0.4024±0.05	0.3258±0.07	0.5336±0.00	0.3678±0.09	0.5826±0.06
	SB	0.4970±0.05	0.5248±0.05	0.5778±0.03	0.5711±0.05	0.5805±0.06	0.5226±0.06	0.5996±0.02	0.4846±0.04	0.5301±0.00	0.5421±0.04	0.5865±0.06
	NU	0.3315±0.03	-	0.3106±0.12	0.3110±0.02	0.2977±0.00	0.3315±0.03	0.2974±0.06	0.2547±0.01	0.4020±0.00	0.3127±0.04	<u>0.3908±0.05</u>
	CC	0.3800±0.02	0.4087±0.02	0.3854±0.02	0.3999±0.03	0.4144±0.02	0.4269±0.01	0.4016±0.02	<u>0.4304±0.01</u>	0.3374±0.00	0.4284±0.01	0.4432±0.01
	AP	0.5354±0.03	0.5619±0.04	0.6095±0.01	<u>0.6164±0.01</u>	0.6026±0.01	0.5667±0.03	0.5497±0.03	0.5545±0.03	0.5132±0.00	0.5630±0.02	0.6413±0.04
	DT	0.5972±0.12	0.6201±0.15	0.6000±0.11	0.7433±0.15	0.7050±0.11	0.6648±0.14	0.6757±0.09	0.5712±0.11	0.8436±0.00	0.7109±0.11	0.8592±0.06
	AV	0.6251±0.08	0.6695±0.10	0.6610±0.13	0.6564±0.09	<u>0.7068±0.10</u>	0.6364±0.10	0.6669±0.10	0.6348±0.12	0.6383±0.00	0.6441±0.14	0.7258±0.11
	OB	0.3403±0.05	0.3590±0.04	0.3622±0.03	0.3794±0.03	0.2923±0.04	0.3651±0.04	0.3828±0.02	0.3596±0.03	0.3813±0.00	0.3732±0.01	0.4251±0.02
	TA	0.4073±0.03	0.4272±0.02	0.4245±0.01	0.4225±0.03	0.3907±0.03	0.4159±0.02	0.4132±0.03	<u>0.4298±0.03</u>	0.4172±0.00	0.4252±0.01	0.4391±0.05
	BM	0.6395±0.10	0.6403±0.10	0.5997±0.07	0.5804±0.06	0.6977±0.11	0.5569±0.03	0.6087±0.07	0.6052±0.08	0.2492±0.00	0.6168±0.09	0.7290±0.10
	CV	0.8628±0.01	0.8754±0.00	0.8671±0.00	0.8455±0.08	0.8462±0.08	0.8736±0.00	0.8768±0.00	0.8414±0.08	0.5816±0.00	0.8736±0.00	0.8759±0.00
	ZO	0.6970±0.09	0.7287±0.07	0.6881±0.08	<u>0.7693±0.07</u>	0.7574±0.10	0.7287±0.08	0.7238±0.13	0.6257±0.07	0.7228±0.00	0.7257±0.03	0.8050±0.08
Average Rank		8.38	6.38	6.75	5.25	5.92	6.29	5.42	8.33	6.83	5.21	1.25
ARI	CA	0.0341±0.04	-	0.0433±0.05	0.0548±0.05	0.0027±0.01	0.0347±0.03	0.0668±0.05	0.0043±0.01	-0.0678±0.00	0.0433±0.05	0.0964±0.07
	SB	0.3245±0.04	0.3540±0.03	0.3755±0.02	0.4071±0.05	<u>0.4104±0.07</u>	0.3502±0.04	0.4057±0.02	0.2941±0.03	0.3811±0.00	0.3868±0.03	0.4135±0.04
	NU	0.0539±0.01	-	0.0679±0.15	0.0350±0.02	<u>0.0730±0.00</u>	0.0539±0.01	0.0451±0.07	0.0027±0.00	0.0310±0.00	0.0327±0.05	0.0761±0.07
	CC	0.0043±0.01	0.0289±0.01	0.0072±0.01	0.0164±0.02	0.0218±0.01	0.0313±0.01	0.0211±0.01	0.0311±0.01	0.0231±0.00	<u>0.0330±0.00</u>	0.0419±0.01
	AP	0.0002±0.01	0.0167±0.02	0.0410±0.01	<u>0.0481±0.01</u>	0.0342±0.01	0.0151±0.02	0.0055±0.02	-0.0007±0.02	0.0470±0.00	0.0022±0.01	0.0783±0.04
	DT	0.4342±0.15	0.5450±0.21	0.4774±0.16	0.6905±0.18	0.6752±0.13	0.5813±0.19	0.6163±0.11	0.4296±0.16	0.7821±0.00	0.6520±0.11	0.8122±0.07
	AV	0.0073±0.02	0.0365±0.02	0.0427±0.09	0.0319±0.03	0.0578±0.07	0.0199±0.02	0.0200±0.04	0.0094±0.02	0.0752±0.00	0.0214±0.03	0.0390±0.05
	OB	0.1286±0.06	0.1389±0.04	0.1374±0.03	0.1669±0.05	0.0958±0.03	0.1497±0.05	0.1652±0.03	0.1435±0.03	0.1674±0.00	<u>0.2032±0.02</u>	0.2255±0.01
	TA	0.0144±0.01	0.0285±0.01	0.0310±0.02	0.0188±0.01	0.0120±0.01	0.0166±0.01	0.0264±0.02	<u>0.0377±0.02</u>	0.0197±0.00	0.0113±0.00	0.0394±0.03
	BM	0.0281±0.02	0.0094±0.06	0.0118±0.02	-0.0163±0.03	0.0693±0.11	0.0012±0.02	0.0217±0.01	0.0087±0.03	-0.0177±0.00	0.0209±0.01	<u>0.0472±0.05</u>
	CV	0.5254±0.02	0.5627±0.00	0.5381±0.00	0.4948±0.17	0.4989±0.18	0.5572±0.00	0.5668±0.00	0.4831±0.17	0.3651±0.00	0.5572±0.00	<u>0.5633±0.00</u>
	ZO	0.6456±0.12	0.6822±0.10	0.6353±0.12	<u>0.7350±0.11</u>	0.7072±0.12	0.6847±0.12	0.6912±0.17	0.4950±0.16	0.5920±0.00	0.6765±0.05	0.7736±0.11
Average Rank		8.38	6.75	6.13	5.42	5.33	6.50	5.08	8.58	6.58	5.83	1.42
CMP	CA	0.8852±0.00	-	0.7960±0.02	0.8026±0.02	0.8271±0.01	0.8852±0.00	0.7932±0.02	0.8352±0.02	0.8792±0.00	0.8100±0.04	0.6557±0.06
	SB	0.3101±0.02	0.2936±0.02	0.2641±0.01	0.2799±0.01	0.2824±0.01	0.2862±0.01	0.2738±0.01	0.3560±0.02	0.3899±0.00	0.2949±0.01	0.2714±0.01
	NU	0.8908±0.00	-	0.7948±0.04	0.8162±0.02	0.8750±0.00	0.8908±0.00	0.7688±0.01	0.8416±0.02	0.9144±0.00	0.7735±0.02	0.6610±0.06
	CC	0.6270±0.03	0.5626±0.03	0.5988±0.02	0.5942±0.02	0.5828±0.02	0.5892±0.02	0.5603±0.02	0.5799±0.02	0.5085±0.00	0.5740±0.01	0.4222±0.04
	AP	0.6083±0.01	0.6044±0.02	0.5738±0.01	0.5814±0.00	0.5730±0.00	0.5850±0.02	0.5771±0.02	<u>0.5378±0.05</u>	0.5285±0.00	0.6211±0.05	0.5692±0.03
	DT	0.4240±0.02	0.4065±0.03	0.3942±0.01	0.3886±0.02	0.3804±0.02	0.3830±0.03	0.3954±0.01	0.4250±0.03	<u>0.3673±0.01</u>	0.3947±0.01	0.3314±0.02
	AV	0.7039±0.02	0.6623±0.05	0.6246±0.06	0.6194±0.08	0.6798±0.05	0.7011±0.02	0.6260±0.03	0.6317±0.06	0.4809±0.00	0.7073±0.05	0.6042±0.10
	OB	0.3324±0.04	0.3617±0.05	0.3128±0.03	0.3289±0.04	0.4115±0.03	0.3646±0.05	0.2870±0.03	0.2779±0.03	0.3824±0.00	<u>0.2445±0.02</u>	0.1793±0.05
	TA	0.4537±0.04	0.4393±0.06	<u>0.3701±0.02</u>	0.4993±0.04	0.3273±0.07	0.5097±0.04	0.3941±0.04	0.4701±0.08	0.6161±0.00	0.5332±0.01	0.3701±0.06
	BM	0.6293±0.02	0.6302±0.04	0.6016±0.03	0.6170±0.03	0.6264±0.04	0.6265±0.01	0.5806±0.01	0.6067±0.01	0.3462±0.00	0.5883±0.04	0.5530±0.06
	CV	0.5514±0.00	0.5465±0.00	0.5457±0.00	0.5643±0.06	0.5616±0.06	0.5450±0.00	0.5435±0.00	0.5669±0.07	0.5099±0.00	0.5454±0.00	0.5328±0.00
	ZO	0.2582±0.01	0.2615±0.02	0.2616±0.01	<u>0.2550±0.02</u>	0.2640±0.02	0.2714±0.01	0.2566±0.01	0.3288±0.05	0.3442±0.00	0.2563±0.00	0.2498±0.01
Average Rank		8.58	7.92	4.83	5.83	6.42	7.67	3.92	7.08	5.75	6.17	1.58

Table 3: ACC performance of ablated DISC variants formed by successively ablating the two core components: RL, CPDs. Red arrows indicate performance degradation over the former more complete variant.

DISC Components		Variants	Datasets											Average Rank	
RL	CPDs		CA	SB	NU	CC	AP	DT	AV	OB	TA	BM	CV		ZO
✓	✓	DISC	0.5826	0.5759	0.3918	0.4432	0.6413	0.8592	0.7258	0.4251	0.4391	0.7290	0.8759	0.8050	1.08
	✓	I	0.5524↓	0.5568↓	0.3876↓	0.4193↓	0.6423	0.7003↓	0.7018↓	0.3963↓	0.4377↓	0.6893↓	0.8722↓	0.8050↓	1.83
		II	0.3795↓	0.4959↓	0.3315↓	0.3800↓	0.5354↓	0.5972↓	0.6251↓	0.3478↓	0.4073↓	0.6395↓	0.8628↓	0.6970↓	3.00

exploring a proper number of prominent clusters, as confirmed in its weaker external validation results. It is also worth noting that the entropy-based evaluation index we use still has some limitations. That is, it actually measures the value inconsistency within different clusters by only distinguishing whether the values are identical or different. So it is relatively incompetent in precisely quantifying the effectiveness of the adopted distance measure in grouping similar samples to form compact clusters. In addition, the performance of CBDM is not reported for the NS and CE datasets, as the attributes are independent, making CBDM ineffective.

4.3 Ablation Studies

To demonstrate the effectiveness of DISC’s core components and verify the Euclidean compatibility of our distance metric, several ablated variants are compared from both algorithmic and data perspectives, as shown in Tables 3 and 4.

From the algorithmic perspective, we evaluate the pro-

posed subspace category relationship learning mechanism by modifying DISC to DISC^I, which infers the category Relationship once without iterative Learning (RL). Additionally, DISC^{II} adopts the traditional Hamming distance and uses the subspace Condition Probability Distribution-based measure (CPDs) for weight computation. **Table 3** compares the clustering accuracy of these variants. It can be seen that DISC outperforms the others on 11 out of 12 datasets, confirming the effectiveness of the proposed metric learning mechanism. DISC^I with CPD-based weights consistently outperforms DISC^{II} with Hamming distance, demonstrating the superiority of the probability distribution in distinguishing dissimilarity and improving clustering accuracy.

From the data perspective, to validate that our metric is compatible with the Euclidean distance, we combine the learned subspace distance metric for categorical attributes with the Euclidean distance for k-prototypes clustering, resulting in the DISC (mixed) method. **Table 4** compares

Table 4: Clustering Accuracy on mixed datasets, formed by adding the omitted numerical attributes back to the categorical datasets in Table 1. **Bold** indicates the best result.

Data	KPT	WOCIL	HARR	DISC	DISC (mixed)
CC	0.3983	0.3858	0.4303	0.4432	0.4494
AP	0.5158	0.5723	0.5757	0.6413	0.6556
DT	0.5313	0.6901	0.6145	0.8592	0.8615
AV	0.539	0.6733	0.6805	0.7258	0.7764
BM	0.5865	0.5724	0.5222	0.7290	0.7423

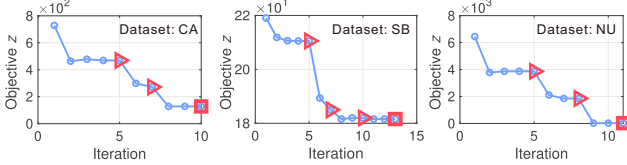


Figure 2: Convergence curves of DISC on different datasets.

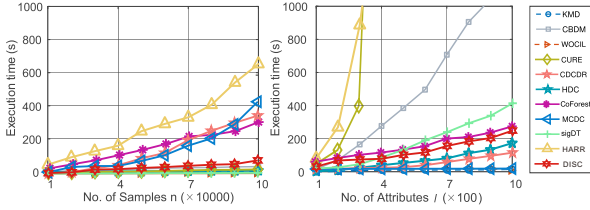


Figure 3: Execution time (y-axis) on synthetic datasets with different numbers of samples n and attributes l (x-axis).

it with k-prototypes (KPT) (Huang 1997) and two mixed data clustering methods, WOCIL (Jia and Cheung 2017) and HARR (Zhang et al. 2025b). DISC (mixed) consistently outperforms all competing methods across the five mixed datasets. Compared to the original DISC, which processes only categorical attributes, the mixed datasets extension achieves further performance gains, demonstrating seamless integration with Euclidean distance and yielding excellent results without complex operations. This simplicity validates its compatibility with Euclidean space, as discussed in **Theorem 4**, enabling DISC to handle heterogeneous data types effectively and making it a promising solution for real-world clustering.

4.4 Converge and Efficiency Analysis

To assess the convergence of DISC, we execute it on CA, SB, and NU datasets and present the convergence curves in **Figure 2**. The horizontal axis represents the number of iterations, while the vertical axis shows the objective function value z . The red triangles and squares correspond to updates of the relation trees and the iterations where DISC converges, respectively. After each relation tree update, z consistently decreases, indicating alignment with the optimization objective. DISC converges within 20 iterations, with a maximum of 10 updates, as expected in **Theorem 6**.

To evaluate the efficiency of DISC, large synthetic

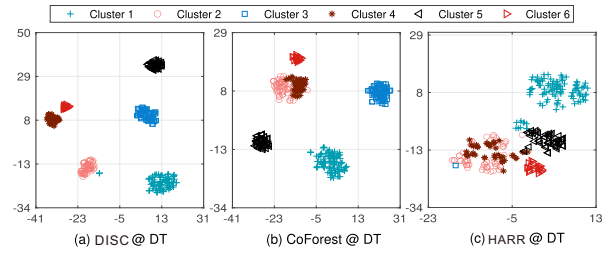


Figure 4: t-SNE visualization of the DT dataset.

datasets are generated with varying attributes and sample sizes. Two experiments are conducted: 1) Fixing the number of attributes at $l = 20$ and varying the sample size n from 10,000 to 100,000; 2) Fixing the sample size at $n = 2,000$ and varying the number of attributes l from 100 to 1000. Each attribute has five values, and $k = 5$. Execution times for all 11 methods are shown in **Figure 3**. DISC outperforms HARR, MCDC, CDCDR, and CoForest on large datasets and HARR, CURE, CBDM, sigDT, and CoForest on high-dimensional datasets. The execution time of DISC increases linearly with the number of samples or attributes, consistent with the time complexity analysis. In summary, DISC is efficient compared to State-Of-Art (SOTA) methods and incurs low computational cost.

4.5 Qualitative Results

To demonstrate the cluster discrimination ability of the DISC method, the attribute distances learned by DISC, CoForest, and HARR are utilized to encode the attributes of the DT dataset for 2-dimensional t-SNE (Maaten and Hinton 2008) visualization. It can be seen from **Figure 4** that DISC has significantly better cluster discrimination ability, as it performs clustering task-oriented distance learning to better suit the exploration of the k clusters of the DT dataset.

5 Concluding Remarks

This paper addresses a key challenge in categorical data clustering, where existing distance learning methods are limited by uniform category relationships across different attributes and clusters. A novel learning framework named DISC is introduced to integrate subspace category relationships into clustering and jointly optimize them. During the learning process, DISC infers more appropriate subspace category relationships based on current partition, leading to more compact clusters and superior accuracy. Theoretical analysis shows that the designed optimization algorithm of DISC converges quickly with guarantee, and also proves that the learned category relationships are both deterministic and Euclidean-compatible. Extensive experiments on 12 benchmark datasets with statistical evidence show that the proposed DISC significantly outperforms various state-of-the-art methods. By bridging the gap in distance metric formulation between categorical data and clustering analysis, DISC lays a solid foundation for addressing critical subsequent challenges in categorical and mixed data clustering, including the handling of high-dimensional and noisy data.

as well as the automatic determination of the optimal number of clusters, which have yet to be fully explored both in the literature and in our present study.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62476063, 62376233 and 62306181, the NSFC/Research Grants Council (RGC) Joint Research Scheme under the grant N_HKBU214/21, the Natural Science Foundation of Guangdong Province under grant: 2025A1515011293, the Natural Science Foundation of Fujian Province under grant: 2024J09001, the National Key Laboratory of Radar Signal Processing under grant: JKW202403, the General Research Fund of RGC under grants: 12201321, 12202622, and 12201323, the RGC Senior Research Fellow Scheme under grant: SRF52324-2S02, Guangdong and Hong Kong Universities “1+1+1” Joint Research Collaboration Scheme with the grant: 2025A0505000004, the grant for Faculty Niche Research Areas of Hong Kong Baptist University with the code: RC-FNRA-IG/23-24/SCI/02, the Guangdong Higher Education Upgrading Plan (2021–2025), the Shenzhen Science and Technology Program under grant: RCBS20231211090659101 and the Xiaomi Young Talents Program.

References

- Agresti, A. 2003. *Categorical data analysis*. John Wiley & Sons.
- Agresti, A. 2010. *Analysis of ordinal categorical data*. John Wiley & Sons.
- Ahmad, A.; and Dey, L. 2007. 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.
- Bai, L.; and Liang, J. 2022. A categorical data clustering framework on graph representation. *Pattern Recognition*, 128: 108694.
- Bai, L.; Liang, J.; Dang, C.; and Cao, F. 2011. A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognition*, 44(12): 2843–2861.
- Bondy, J. A.; and Murty, U. S. R. 2008. *Graph theory*. Springer.
- Cai, S.; Zhang, Y.; Luo, X.; Cheung, Y.-m.; Jia, H.; and Liu, P. 2024. Robust categorical data clustering guided by multi-granular competitive learning. In *Proceedings of the IEEE 44th International Conference on Distributed Computing Systems*, 288–299.
- Cao, F.; Liang, J.; Li, D.; and Zhao, X. 2013. A weighting k-modes algorithm for subspace clustering of categorical data. *Neurocomputing*, 108: 23–30.
- Chen, J.; Ji, Y.; Zou, R.; Zhang, Y.; and Cheung, Y.-m. 2024. QGRL: Quaternion graph representation learning for heterogeneous feature data clustering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 297–306.
- Cheung, Y.-m.; and Jia, H. 2013. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8): 2228–2238.
- dos Santos, T. R.; and Zárate, L. E. 2015. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3): 1247–1260.
- Dua, D.; and Karra Taniskidou, E. 2017. UCI Machine Learning Repository.
- Feng, S.; Zhao, M.; Huang, Z.; Ji, Y.; Zhang, Y.; and Cheung, Y.-M. 2025. Robust qualitative data clustering via learnable multi-metric space fusion. In *Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.
- Fernandez-Navarro, F.; Campoy-Munoz, P.; Hervás-Martínez, C.; and Yao, X. 2013. Addressing the EU sovereign ratings using an ordinal regression approach. *IEEE Transactions on Cybernetics*, 43(6): 2228–2240.
- Gates, A. J.; and Ahn, Y.-Y. 2017. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(1): 3049–3076.
- Hamming, R. W. 1950. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2): 147–160.
- He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 507–514.
- Hu, L.; Dong, J.; Jiang, M.; Liu, Y.; and He, Z. 2025a. Clusterability test for categorical data. *Knowledge and Information Systems*, 67(5): 4113–4138.
- Hu, L.; Jiang, M.; Dong, J.; Liu, X.; and He, Z. 2025b. Interpretable categorical data clustering via hypothesis testing. *Pattern Recognition*, 162: 111364.
- Hu, L.; Jiang, M.; Liu, X.; and He, Z. 2025c. Significance-based decision tree for interpretable categorical data clustering. *Information Sciences*, 690: 121588.
- Hu, L.; Jiang, M.; Liu, Y.; Zou, Q.; and He, Z. 2025d. Clustering Categorical Data via Multiple Hypothesis Testing. *ACM Transactions on Knowledge Discovery from Data*, 19(5): 1556–4681.
- Hu, L.; Li, Z.; Dong, J.; Jiang, M.; and He, Z. 2025e. Statistical significance of cluster membership for categorical data. *Engineering Applications of Artificial Intelligence*, 160: 111861.
- Huang, Z. 1997. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 21–34.
- Huang, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3): 283–304.
- Ienco, D.; Pensa, R. G.; and Meo, R. 2009. Context-based distance learning for categorical data clustering. In *Proceedings of the 8th International Symposium on Intelligent Data Analysis*, 83–94.

Ienco, D.; Pensa, R. G.; and Meo, R. 2012. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data*, 6(1): 1–25.

Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3): 264–323.

Jia, H.; and Cheung, Y.-m. 2017. 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.

Jian, S.; Hu, L.; Cao, L.; and Lu, K. 2018a. Metric-based auto-instructor for learning mixed data representation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 3318–3325.

Jian, S.; Pang, G.; Cao, L.; Lu, K.; and Gao, H. 2018b. CURE: Flexible categorical data representation by hierarchical coupling learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(5): 853–866.

Johnson, V. E.; and Albert, J. H. 2006. *Ordinal data modeling*. Springer Science & Business Media.

Le, S. Q.; and Ho, T. B. 2005. An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 26(16): 2549–2557.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 296–304.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11): 2579–2605.

Munkres, J. R. 2000. *Topology*. Prentice Hall.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Zhang, C.; Chen, X.; Tan, Z.; Gu, F.; Ji, Y.; and Zhang, Y. 2025a. Towards Clustering of Incomplete Mixed-Attribute Data. *Expert Systems*, 42(7): e70074.

Zhang, Y.; and Cheung, Y.-m. 2020. An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 6869–6876.

Zhang, Y.; and Cheung, Y.-m. 2022a. 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.

Zhang, Y.; and Cheung, Y.-m. 2022b. A new distance metric exploiting heterogeneous inter-attribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Cybernetics*, 52(2): 758–771.

Zhang, Y.; and Cheung, Y.-m. 2023. Graph-based dissimilarity measurement for cluster analysis of any-type-attributed Data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 6530–6544.

Zhang, Y.; Cheung, Y.-m.; and Tan, K. 2020. 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.

Zhang, Y.; Cheung, Y.-m.; and Zeng, A. 2022. 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*, 3758–3765.

Zhang, Y.; Zhao, M.; Chen, Y.; Lu, Y.; and Cheung, Y.-m. 2025b. Learning unified distance metric for heterogeneous attribute data clustering. *Expert Systems with Applications*, 273: 126738.

Zhang, Y.; Zhao, M.; Jia, H.; Lu, Y.; Li, M.; and Cheung, Y.-m. 2024. Order is all you need for categorical data clustering. *arXiv preprint arXiv:2411.15189*.

Zhang, Y.; Zou, R.; Zhang, Y.; Zhang, Y.; Cheung, Y.-m.; and Li, K. 2025c. Adaptive micro partition and hierarchical merging for accurate mixed data clustering. *Complex & Intelligent Systems*, 11(1): 84.

Zhao, L.; Zhang, Y.; Ji, Y.; Zeng, A.; Gu, F.; and Luo, X. 2022. Heterogeneous drift learning: classification of mix-attribute data with concept drifts. In *Proceedings of the 2022 IEEE 9th International Conference on Data Science and Advanced Analytics*, 1–10. IEEE.

Zhao, M.; Feng, S.; Zhang, Y.; Li, M.; Lu, Y.; and Cheung, Y.-m. 2024. Learning order forest for qualitative-attribute data clustering. In *Proceedings of the 27th European Conference on Artificial Intelligence*, 1943–1950.

Zhu, C.; Cao, L.; and Yin, J. 2022. Unsupervised heterogeneous coupling learning for categorical representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 533–549.

Zhu, C.; Zhang, Q.; Cao, L.; and Abrahamyan, A. 2020. Mix2Vec: Unsupervised mixed data representation. In *Proceedings of the 7th International Conference on Data Science and Advanced Analytics*, 118–127.

A Lemmas, Theromes and Algorithm

A.1 Deterministic and Euclidean-compatible Distance Metric

Lemma 1. *The relation tree $\mathcal{T}_{j,r}$ inferred from the fully connected graph $G_{j,r}$ retains all pairwise distances among nodes is deterministic and unique.*

Proof. Given a fully connected graph $G_{j,r}$, for the convenience of analysis without loss of generality, we first reindex the nodes such that their conditional probabilities satisfy:

$$p(v_r^1|C_j) < p(v_r^2|C_j) < \dots < p(v_r^{o_r}|C_j), \quad (12)$$

where the superscript indices $1, 2, \dots, o_r$, now explicitly represent the ranking order of probabilities. The edge with the smallest weight is selected at each step to incrementally form $\mathcal{T}_{j,r}$ according to Eq. (6). Since $\mathbf{W}_{j,r}(u, u+1)$ between adjacent nodes v_r^u and v_r^{u+1} exhibits the local minimum weight, the edge $b_r^{u,u+1}$ is included in the edge set of $\mathcal{T}_{j,r}$. Consequently, all pairs of adjacent nodes are sequentially connected, yielding a tree structure. Assuming that

$s > u$, the distance between any two nodes v_r^u and v_r^s is then given by:

$$\begin{aligned} W_{j,r}(u, s) &= \sum_{\eta=u}^{s-1} (p(v_r^{\eta+1}|C_j) - p(v_r^\eta|C_j)) \\ &= |p(v_r^u|C_j) - p(v_r^s|C_j)| \end{aligned} \quad (13)$$

This matches the direct edge weight in the original graph $G_{j,r}$, demonstrating that the relation tree introduces no information loss. This preservation implies that no alternative tree structure can yield the same set of pairwise distances unless it is identical to $\mathcal{T}_{j,r}$. Therefore, both the structure and the induced metric are uniquely determined by the conditional probability distribution, confirming the determinism of the construction. \square

Lemma 2. *The distance metric induced by the relation tree $\mathcal{T}_{j,r}$ always forms a linear structure, which is naturally compatible with the one-dimensional Euclidean geometry.*

Proof. To rigorously prove that the inferred tree $\mathcal{T}_{j,r}$ is linear, we have the following three steps:

- 1) Base Case: When $o_r = 2$, the statement trivially holds.
- 2) Inductive Hypothesis: Assume the proposition holds for some arbitrary natural number $o_r = \omega$, and reindex the nodes such that:

$$p(v_r^1|C_j) < p(v_r^2|C_j) < \dots < p(v_r^\omega|C_j), \quad (14)$$

the nodes $v_r^1, v_r^2, \dots, v_r^\omega$ form a linear structure, i.e., v_r^u connects v_r^{u+1} for $1 < u < \omega$.

3) Inductive Step: We now prove that the proposition holds for $o_r = \omega + 1$. Without loss of generality, assume that the probability of the newly added node $v_r^{\omega+1}$, denoted as $p(v_r^{\omega+1}|C_j)$, satisfies $p(v_r^\omega|C_j) < p(v_r^{\omega+1}|C_j)$. In this case, the edge $w_{j,r}^{\omega, \omega+1}$ is the smallest, and thus, according to Eq. (6), v_r^ω and $v_r^{\omega+1}$ are connected, leading to a linear structure represented as $v_r^1, v_r^2, \dots, v_r^\omega, v_r^{\omega+1}$.

Thus, the tree consisting of $\omega + 1$ nodes remains linear, which is consistent with Euclidean distance in \mathbb{R}^1 , allowing for seamless integration between the two. As a result, the metric induced by the relation tree is Euclidean-compatible. \square

According to Lemmas 1 and 2, we have Theorem 4.

Theorem 4. *The relation tree $\mathcal{T}_{j,r}$ inferred from the fully connected graph $G_{j,r}$ defines a deterministic, Euclidean-compatible distance metric over the attribute values.*

Proof. It follows from Lemmas 1 and 2 that $\mathcal{T}_{j,r}$ can be simplified into a linear structure with $o_r - 1$ edges connecting o_r nodes without loss of information, and the resulting distance metric is both deterministic and Euclidean compatible. \square

A.2 Valid Distance Metric

Lemma 3. *The distance measure $\mathbf{D}_{j,r}(u, s)$ defined within the context of the relation tree $\mathcal{T}_{j,r}$ indeed constitutes a valid distance metric.*

Proof. The distance measure $\mathbf{D}_{j,r}(u, s)$ follows non-negativity, symmetry, and triangle inequality for any $r \in \{1, 2, \dots, l\}$ and $u, s, g \in \{1, 2, \dots, o_r\}$ as shown below.

1) **Non-negativity:** $\mathbf{D}_{j,r}(u, s) \geq 0$. $\mathbf{D}_{j,r}(u, s)$ denotes a distance, which is always a non-negative weight according to Eqs. (4) and (7) in the submitted paper;

2) **Symmetry:** $\mathbf{D}_{j,r}(u, s) = \mathbf{D}_{j,r}(s, u)$. Since the relation tree is an undirected graph, the weights extracted from the undirected graph obey the commutative law for their summation;

3) **Triangle inequality:** $\mathbf{D}_{j,r}(u, s) \leq \mathbf{D}_{j,r}(u, g) + \mathbf{D}_{j,r}(g, s)$. The shortest path between two values with length $\mathbf{D}_{j,r}(u, s)$ on the relation tree is the unique path. If we detour from v_r^u to v_r^s via another node v_r^g that is not on the shortest path between nodes v_r^u and v_r^s , it necessarily involves extra weight(s) from the other paths. Given that each weight is non-negative, the result follows. \square

According to Lemma 3, we have the Theorem 5.

Theorem 5. *The sample-cluster distance $\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j)$ defined in the context of \mathcal{T}_j represents a valid distance metric.*

Proof. The computation of $\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j)$ can be viewed as the weighted sum of a series of $\mathbf{D}_{j,r}(u, s)$ with non-negative weights represented by the probabilities in $p(v_r^u|C_j)$ according to Eqs. (8) and (9) in the submitted paper. Since distance $\mathbf{D}_{j,r}(u, s)$ is a metric according to Lemma 3, $\Phi(\mathbf{x}_i, \mathbf{m}_j; \mathcal{T}_j)$ is also a metric following non-negativity, symmetry, and triangle inequality. The result follows. \square

A.3 Converge Guarantee

Lemma 4. *The inner loop attains a local minimum of $z(\mathbf{H}, M, \mathcal{T})$ within a finite number of iterations when \mathcal{T} is held constant.*

Proof. Given X with finite numbers of samples n and clusters k , the possible configurations of the partition \mathbf{H} and center set M are also finite. Consequently, each distinct pair (\mathbf{H}, M) can appear no more than once during the iterative process. Consider two distinct iterations \mathcal{I}_1 and \mathcal{I}_2 such that $\mathbf{H}^{\{\mathcal{I}_1\}} = \mathbf{H}^{\{\mathcal{I}_2\}}$ and $M^{\{\mathcal{I}_1\}} = M^{\{\mathcal{I}_2\}}$. It follows that $z(\mathbf{H}^{\{\mathcal{I}_1\}}, M^{\{\mathcal{I}_1\}}, \mathcal{T}^{\{E\}}) = z(\mathbf{H}^{\{\mathcal{I}_2\}}, M^{\{\mathcal{I}_2\}}, \mathcal{T}^{\{E\}})$. However, the values of the objective function $\{z(\mathbf{H}^{\{1\}}, M^{\{1\}}, \mathcal{T}^{\{E\}}), z(\mathbf{H}^{\{2\}}, M^{\{2\}}, \mathcal{T}^{\{E\}}), \dots, z(\mathbf{H}^{\{I\}}, M^{\{I\}}, \mathcal{T}^{\{E\}})\}$ produced by the inner loop exhibit strictly decreasing. Thus, the inner loop is guaranteed to converge to a local minimum in a finite number of steps. \square

Lemma 5. *The outer loop E converges to a local minimum of $z(\mathbf{H}, M, \mathcal{T})$ in a finite number of iterations.*

Proof. Note that the subspace category relationships \mathcal{T} reside in a DIScrete space, implying that the number of possible configurations for \mathcal{T} is finite. Additionally, the number of possible cluster partition \mathbf{H} and center set M is finite for X consisting of n samples and k clusters. Therefore, the number of distinct combinations of \mathbf{H} , M , and \mathcal{T} is also finite. Each possible combination of \mathbf{H} , M , and \mathcal{T} appears at most once during the execution of the algorithm. Suppose that there exist two

distinct iterations E_1 and E_2 , where $\mathbf{H}^{\{E_1\}} = \mathbf{H}^{\{E_2\}}$, $M^{\{E_1\}} = M^{\{E_2\}}$, and $\mathcal{T}^{\{E_1\}} = \mathcal{T}^{\{E_2\}}$. The corresponding updates $\mathcal{T}^{\{E_1+1\}}$ and $\mathcal{T}^{\{E_2+1\}}$ can be computed deterministically, and $\mathcal{T}^{\{E_1+1\}} = \mathcal{T}^{\{E_2+1\}}$. Upon convergence of the subsequent inner loop \mathcal{I} , the resulting assignments $\mathbf{H}^{\{E_1+1\}}$, $M^{\{E_1+1\}}$ and $\mathbf{H}^{\{E_2+1\}}$, $M^{\{E_2+1\}}$ must also satisfy $\mathbf{H}^{\{E_1+1\}} = \mathbf{H}^{\{E_2+1\}}$ and $M^{\{E_1+1\}} = M^{\{E_2+1\}}$. Consequently, the objective function at the next iteration satisfy $z(\mathbf{H}^{\{E_1+1\}}, M^{\{E_1+1\}}, \mathcal{T}^{\{E_1+1\}}) = z(\mathbf{H}^{\{E_2+1\}}, M^{\{E_2+1\}}, \mathcal{T}^{\{E_2+1\}})$. Since the sequence $\{z(\mathbf{H}^{\{1\}}, M^{\{1\}}, \mathcal{T}^{\{1\}}), z(\mathbf{H}^{\{2\}}, M^{\{2\}}, \mathcal{T}^{\{2\}}), \dots, z(\mathbf{H}^{\{E\}}, M^{\{E\}}, \mathcal{T}^{\{E\}})\}$ is monotonically decreasing, the outer loop E converges to a local minimum in a finite number of steps. \square

According to Lemmas 4 and 5 we have the Theorem 6.

Theorem 6. *DISC algorithm converges to a local minimum in a finite number of iterations.*

Proof. Based on Lemma 4 and Lemma 5, the convergence of the inner and outer loop, i.e., lines 5 - 9 and lines 3 - 16 of Algorithm 1 is proven, respectively. Consequently, the convergence of Algorithm DISC can be firmly validated. \square

A.4 Time Complexity of DISC

Theorem 7. *Time complexity of DISC is $O(nlk\mathcal{I}E)$.*

Proof. Assume solving problem $z(\mathbf{H}, \hat{M}, \hat{\mathcal{T}})$ involves \mathcal{I} iterations to compute \mathbf{H} , and $z(\hat{\mathbf{H}}, M, \hat{\mathcal{T}})$ involves \mathcal{I} iterations to compute M . The whole algorithm involves E iterations to infer \mathcal{T} for solving $z(\hat{\mathbf{H}}, \hat{M}, \mathcal{T})$. For worst-case analysis, we adopt σ to indicate the maximum number of possible values of a dataset by $\sigma = \max(o_1, o_2, \dots, o_l)$.

For each iteration of \mathcal{I} , \mathbf{H} should cluster n samples to k clusters by considering σ values of l attributes according to Eq. (2) in the submitted paper, with time complexity $O(nlk\sigma)$. M computes the mode for each cluster by traversing l attributes of all samples. The complexity is $O(nlk)$. Therefore, the time complexity of solving $z(\mathbf{H}, \hat{M}, \hat{\mathcal{T}})$ and $z(\hat{\mathbf{H}}, M, \hat{\mathcal{T}})$ in a total of \mathcal{I} iterations is $O(\mathcal{I}nlk\sigma)$.

For each iteration of E , since \mathbf{H} and M have been prepared, the fully connected graphs G should be prepared by going through all the samples within each cluster. The complexity is $O(nlk\sigma)$. The inference process from G to \mathcal{T} should consider l attributes with σ possible within each cluster. The time complexity is $O(kl\sigma^2 \log \sigma)$. For E iterations of the whole DISC algorithm, considering the \mathcal{I} inner iterations, the overall time complexity of DISC is $O(E(\mathcal{I}nlk\sigma + nlk\sigma + kl\sigma^2 \log \sigma))$.

Since σ is a small integer ranging from 2 to 8 in most cases, it is conventionally regarded as a constant term in asymptotic complexity analysis. Consequently, the dominant computational complexity of DISC is $O(nlk\mathcal{I}E)$. \square

A.5 Optimization Algorithm of DISC

The problem of optimization DISC can be formulated as computing \mathbf{H} , M , and \mathcal{T} to minimize z , which can be summarized into the following three steps: 1) Fix M and \mathcal{T} ,

compute \mathbf{H} ; 2) Fix \mathcal{T} and \mathbf{H} , compute M ; 3) Repeat 1) and 2) until convergence, fix \mathbf{H} and M , and reinfer \mathcal{T} according to Eqs. (4), (5), and (7). These steps are repeated until \mathbf{H} no longer changes. The whole algorithm is summarized as Algorithm 1, which can be proven to guarantee convergence and has a complexity of $O(nlk\mathcal{I}E)$, linear to n , l , and k .

Algorithm 1: Optimization algorithm of DISC.

Input : Dataset X , number of sought clusters k .
Output: Partition \mathbf{H} , subspace structure \mathcal{T} .

- 1 Initialization: Set outer and inner loop counters by $E \leftarrow 0$ and $\mathcal{I} \leftarrow 0$; Execute k -modes (Huang 1998) to obtain initial $\mathbf{H}^{\{E\}}$ and M ; Construct initial $\mathcal{T}^{\{E\}}$ according to $\mathbf{H}^{\{E\}}$; Set convergence flag for outer loop by $Flag_E \leftarrow False$.
- 2 **while** $Flag_E = False$ **do**
- 3 Set convergence flag for inner loop by $Flag_I \leftarrow False$;
- 4 **while** $Flag_I = False$ **do**
- 5 $\mathcal{I} \leftarrow \mathcal{I} + 1$; Compute $\mathbf{H}^{\{\mathcal{I}\}}$ by Eq. (2);
- 6 **if** $\mathbf{H}^{\{\mathcal{I}\}} = \mathbf{H}^{\{\mathcal{I}-1\}}$ **then**
- 7 $Flag_I \leftarrow True$;
- 8 **end**
- 9 **end**
- 10 **if** $\mathbf{H}^{\{E\}} = \mathbf{H}^{\{\mathcal{I}\}}$ **then**
- 11 $Flag_E \leftarrow True$;
- 12 **else**
- 13 $E \leftarrow E + 1$; $\mathbf{H}^{\{E\}} \leftarrow \mathbf{H}^{\{\mathcal{I}\}}$; Reinfer $\mathcal{T}^{\{E\}}$;
- 14 **end**
- 15 **end**
