Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(6): 4191–4215.
DOI: 10.3934/era.2026188
Received: 20 February 2026
Revised: 02 April 2026
Accepted: 29 April 2026
Published: 19 May 2026

Research article

Subspace and metric space learning for coqualitative data clustering

Duanjiao Li[1], **Yun Chen**[1], **Wenxing Sun**[1], **Yuhui Chen**[1], **Junwen Yao**[1], **Hua Ye**[2] **and Jianguo Zhang**[2,*]

[1] Guangdong Power Grid Co., Ltd, Guangzhou, China

[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China

* **Correspondence:** Email: zhangjianguo@cuhk.edu.cn.

Abstract: Cluster analysis of unlabeled categorical data is crucial in a wide range of practical applications, such as medical diagnosis, financial risk assessment, and recommendation systems. Unlike numerical data residing in explicit Euclidean spaces, categorical data consists of qualitative values without inherent ordering, making the definition of object similarity a critical yet challenging determinant of clustering success. Conventional approaches typically rely on single, predefined metrics (e.g., Hamming distance or context-based measures). However, these metrics are often constructed based on limited prior knowledge or specific statistical assumptions, failing to capture the complex, intrinsic structures of diverse datasets. Consequently, the mismatch between the defined metric space and the "true" data structure significantly hinders the performance of downstream clustering tasks. To address these limitations, this paper proposes a novel subspace and metric space co-learning framework named SBMS. Instead of relying on a static measure, SBMS introduces an adaptive learning paradigm that iteratively optimizes two coupled spaces: a metric space, where multiple complementary distance metrics are fused to provide a comprehensive similarity measure; and an attribute subspace, where attribute weights are dynamically adjusted based on cluster discrimination and compactness to identify the most relevant features for each cluster. Furthermore, we provide a theoretical analysis of the proposed method, discussing its computational complexity and demonstrating the convergence properties of the optimization algorithm. Extensive experiments on real-world public datasets from various domains illustrate that SBMS effectively bridges the gap between defined and true metric spaces, yielding superior clustering accuracy and stability compared to state-of-the-art baselines.

Keywords: categorical data clustering; metric learning; subspace learning; multiobject optimization; unsupervised learning

1. Introduction

Categorical data attributes are made up of qualitative values that are widely prevalent in a vast array of practical applications [1]. It is common to analyze such data in typical data analysis domains such as medical diagnosis [2], where records consist of discrete symptoms and disease codes, as well as in financial risk assessment and social behavior data analysis [3], where user profiles are described by discrete tags and preferences. Unlike continuous numerical data, these categorical attributes represent distinct states or types. Consequently, exploring object groups for these qualitative-valued categorical data in an unsupervised way (i.e., cluster analysis) is a crucial and fundamental task [4]. It serves as a prerequisite for discovering intrinsic patterns, summarizing complex datasets, and supporting downstream decision-making processes.

However, effectively clustering categorical data is a challenging endeavor compared to the well-established analysis of numerical data [5]. Numerical attributes reside in a continuous space with explicit geometric structures, where distances, such as Euclidean distance, are naturally well-defined, and mathematical operations such as averaging are valid [6]. In contrast, categorical values do not initially possess such explicit distance information or geometric properties. For instance, consider a demographic attribute such as "Occupation" which takes values such as "Doctor", "Lawyer", and "Engineer". Although these values carry semantic meaning, they lack quantitative magnitude [7]. One cannot intuitively quantify the geometric distance between a "Doctor" and a "Lawyer", nor is it meaningful to compute the "mean" of these occupations. This inherent ambiguity makes standard distance-based algorithms inapplicable. To tackle this issue, clustering of categorical data keeps absorbing attention, and the existing attempts can be roughly divided into two categories: 1) directly defining a metric/measure for similarity-based clustering [8], and 2) encoding categorical values into numerical ones and perform numerical data clustering [9].

For the metric-defining stream, the k-modes algorithm is the dominant baseline, employing the Hamming distance to measure dissimilarity. This metric operates on a strict binary matching principle: the distance is zero if two values are identical and one otherwise [10]. Although this approach is computationally inexpensive, this uniform assignment is theoretically flawed for complex data, as it fails to capture the semantic proximity between distinct values (e.g., treating "blue" vs. "dark blue" the same as "blue" vs. "red") and ignores the frequency distribution of categories. To address this coarse granularity, context-based metrics [11] have been proposed to derive intervalue distances from the data distribution itself. These methods typically assume that two values in a target attribute are similar if they exhibit similar conditional probability distributions over the remaining context attributes [12]. Although these measures successfully incorporate statistical dependencies, they often incur high computational costs due to pairwise value comparisons and rely heavily on the assumption that attributes are sufficiently correlated. Alternatively, entropy-based approaches [13] introduce informationtheoretic criteria to optimize the clustering process. By calculating the information entropy within clusters, these methods dynamically assign higher weights to attributes that reduce the uncertainty of object-cluster affiliations. However, they often optimize a fixed objective that may not align with the intrinsic geometric structure of the specific dataset.

Regarding the encoding stream, conventional one-hot encoding remains the standard preprocessing technique. This method transforms categorical attribute values into high-dimensional binary vectors where each category maps to an orthogonal basis vector [14]. Although this conversion allows numer-

ical algorithms to process qualitative data, it fundamentally distorts the intrinsic data structure [15]. The orthogonality implies that the Euclidean distance between any pair of distinct values is identical, regardless of their semantic closeness. Consequently, this representation creates a sparse feature space suffering from the curse of dimensionality, where the notion of proximity becomes meaningless. To mitigate the information loss caused by orthogonality, sophisticated encoding approaches [16] have been developed to incorporate inter- and intra-attribute statistics. By analyzing the co-occurrence frequencies and coupling relationships among attributes, these methods embed discrete symbols into a continuous, dense vector space where geometric distances reflect semantic similarities. Furthermore, object-level encoding strategies [17] take a different perspective by reconstructing the entire dataset based on interobject similarity matrices, aiming to preserve the global manifold structure of the data. However, these representation learning methods typically operate as a static preprocessing step independent of the downstream clustering task. They produce a fixed feature space that cannot adaptively refine itself to fit the specific cluster structures emerging during the learning process, thereby limiting the final clustering performance.

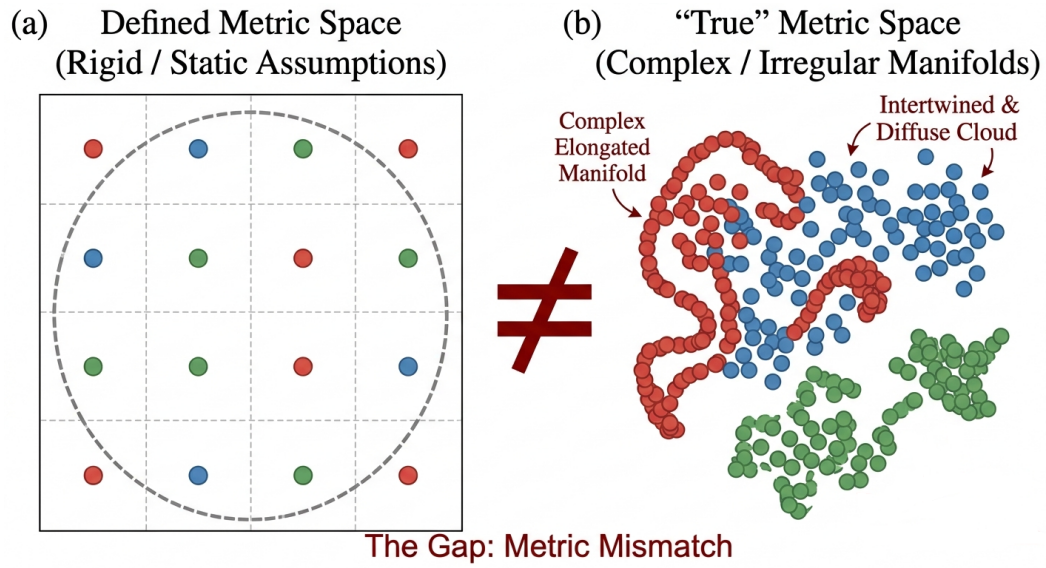

Figure 1. Conceptual illustration of the "metric mismatch" gap. The left panel shows the *defined metric space*, where predefined static measures (e.g., Hamming distance) often force categorical data into rigid, uniformly distributed geometric structures, causing distinct clusters to heavily overlap. The right panel illustrates the *"true" metric space*, where data points naturally reside in complex, irregular manifolds governed by intrinsic semantic relationships. SBMS aims to bridge this specific gap through adaptive dual-space optimization.

Most of the aforementioned solutions converge on a common ultimate goal: establishing a rigorous distance metric space to facilitate accurate cluster learning. In pursuit of this objective, recent advancements have introduced sophisticated similarity metrics [18] and adaptive encoding approaches [19]. Distinct from static measures, these methods not only informatively define the initial distance space but also incorporate optimization mechanisms to fine-tune the metric space with respect to the specific clustering task [20]. By iteratively adjusting attribute weights or updating value embeddings during the clustering process, they ostensibly provide a more suitable geometric basis for distinguishing cat-

egorical objects. However, a critical limitation persists in these state-of-the-art approaches. Their construction of the metric space remains heavily dependent on specific prior knowledge or rigid theoretical hypotheses. For instance, many algorithms implicitly assume a linear relationship between attribute importance or rely on fixed statistical distributions to model value similarities. Such assumptions introduce a strong inductive bias that cannot precisely reflect the true, often complex, structure of the distance space for categorical data, where attributes are described by vague qualitative values with non-linear interactions [21].

Consequently, the rigidity of the defined metric space restricts the algorithm's ability to uncover the intrinsic data distribution, thereby degenerating the effectiveness of downstream processing. Figure 1 intuitively demonstrates this fundamental issue. It visualizes the substantial discrepancy between the "defined" metric spaces, which are often constrained to regular geometric forms due to algorithmic assumptions, and the "true" metric spaces of categorical data collected from various domains, which may exhibit irregular or nonconvex topological structures. This misalignment suggests that relying on a single, hypothesis-driven metric is insufficient for capturing the diverse nature of real-world datasets [22].

To bridge the discrepancy between pre-defined metrics and the intrinsic data structure, this paper proposes a novel categorical data clustering paradigm. The core objective is to learn a fusion of multiple metrics that complement each other in capturing diverse information, and subsequently to perform subspace learning on the represented attributes to leverage accurate clustering. Specifically, the proposed method collaboratively learns a global distance metric space by integrating complementary similarity measures to form a comprehensive view of data relationships, and a local attribute subspace by dynamically weighting attributes to highlight cluster-specific discriminative features. Consequently, as a distance metric space and an attribute subspace are collaboratively learned within a unified framework, the proposed method is named SBMS. Because dual space learning and cluster learning are interconnected to facilitate continuous information passing, SBMS is highly flexible in adapting distance measures to the specific clustering tasks with respect to various datasets. The clustering results guide the refinement of the metric space, and the optimized metric space, in turn, improves the discovery of cluster structures. This adaptive closed-loop mechanism effectively addresses the suitability issue of categorical data metrics, ensuring that the learned metric space aligns with the true data distribution rather than relying on fixed assumptions. As a result, SBMS features superior clustering accuracy compared to state-of-the-art baselines. Extensive experiments on real-world public datasets from various domains illustrate that SPACE2 effectively bridges the gap between defined and true metric spaces, yielding superior clustering accuracy and stability compared to state-of-the-art baselines. Specifically, the main contributions of this paper are summarized as follows:

- A unified dual space learning paradigm is proposed to tackle the categorical data clustering problem. Unlike traditional methods that rely on static measures, this paradigm iteratively optimizes two coupled spaces: a global metric space for capturing general data relationships and a local attribute subspace for identifying cluster-specific features. This mechanism effectively addresses the metric mismatch issue between the defined and true metric spaces.
- An adaptive metric fusion strategy is dedicatedly designed to leverage heterogeneous similarity measures. By integrating context-based, entropy-based, and probability-based metrics, the proposed approach allows different metrics to complement each other. It is revealed that such a diversified combination provides a more comprehensive approximation of the optimal metric space compared to single-metric approaches.

- Comprehensive experimental evaluations are conducted on multiple real-world datasets from various domains. The results demonstrate that the proposed method significantly outperforms state-of-the-art baselines in terms of clustering accuracy and stability. Furthermore, detailed convergence analysis empirically validates the robustness and efficiency of the dual-space learning mechanism.

2. Related work

This section provides a comprehensive review of the existing literature relevant to our work, categorized into three main streams: similarity measures for categorical data, representation learning strategies, and advancements in subspace clustering and metric learning.

2.1. Similarity measures for categorical data

Defining an appropriate similarity measure is the cornerstone of categorical data clustering. Unlike numerical data, where Minkowski distances are naturally applicable, categorical data lacks an inherent geometric structure. The most fundamental approach is a matching-based measure, such as the Hamming distance used in the standard k-modes algorithm. It calculates the distance between two objects by simply counting the number of mismatched attributes. Although computationally efficient with a linear complexity, it treats all mismatches identically, failing to capture the semantic nuances where some values might be more similar than others.

To overcome the coarse granularity of binary matching, probability-based and informationtheoretic measures have been introduced. Approaches such as that of [13] have employed information entropy to quantify the uncertainty of attributes or the associations between objects and clusters. The intuition is that rare values or attributes with high discriminatory power should contribute more to the similarity calculation than other values and attributes. For instance, frequency-based methods assign higher weights to mismatches in infrequent categories, assuming they convey more specific information. However, these methods often rely on global statistics and assume independence between attributes, which may overlook the complex coupling relationships in data.

Further advancing this direction, context-based measures aim to capture the interdependence between attributes. The association-based distance metric (ABDM) [23] and context-based distance metric (CBDM) [24] evaluate the similarity between two values of a target attribute by examining their conditional probability distributions over other context attributes. If two values, such as "flu" and "cold", frequently co-occur with similar symptoms in other attributes, they are considered semantically close. More recently, coupled metric learning approaches [25] have been proposed to model both intraattribute value frequencies and inter-attribute couplings simultaneously to form a more comprehensive metric. Despite their effectiveness in capturing semantics, these context-aware methods typically require calculating pairwise value similarities, leading to a quadratic complexity of $O(d^2 \cdot N^2)$ or $O(d \cdot v^2)$, which can be computationally prohibitive for high-dimensional datasets.

2.2. Representation learning and encoding strategies

Because standard distance-based algorithms cannot directly process qualitative symbols, representation learning aims to map categorical values into a numerical vector space. The most conventional strategy is one-hot encoding, which converts a variable with v categories into a v-dimensional binary

vector. However, this method maps all categories to equidistant orthogonal points in the Euclidean space. Consequently, the distance between any distinct pair of values is fixed at $\sqrt{2}$, forcing the loss of all structural and semantic information intrinsic to the data.

To address the limitations of orthogonality, statistical encoding methods have been developed to embed categorical values into dense numerical vectors. For example, Qian et al. [17] proposed a space structure learning method that utilizes the interaction of attribute values to determine the coordinate positions of categories. Similarly, coupled metric learning strategies [26] encode values by considering the rigorous coupling relationships between intra-attribute marginal distributions and interattribute correlations. These methods ensure that the geometric distance in the embedded space reflects the statistical similarity of the original categories.

In recent years, more advanced representation learning frameworks [27, 28] have emerged, often utilizing neural networks or graph embeddings to learn distributed representations of categorical data. Some approaches extend this to temporal or complex data domains. Despite their expressiveness, most of these representation methods operate in a "two-stage" manner: data is encoded first, and clustering is performed subsequently [16]. This separation means that the representation is fixed and does not adapt to the specific cluster structures discovered during the learning process, potentially leading to suboptimal performance when the prelearned features are not aligned with the clustering objective. Recent studies have increasingly explored contrastive learning frameworks and deep generative models, such as diffusion models, to capture complex, nonlinear interattribute dependencies without heavily relying on manually defined distance metrics [29]. Furthermore, some pioneering works have investigated the adaptation of large language models (LLMs) to enhance tabular data representation and downstream clustering performance [30], particularly on challenging datasets involving data heterogeneity or distribution shifts [31]. Although these deep representation learning models excel at extracting high-level semantic features, they often require extensive computational resources, rigorous hyperparameter tuning, and large-scale training data. In contrast, our proposed SBMS framework offers a highly efficient and interpretable alternative by dynamically optimizing the metric space and attribute subspace in a unified, transparent manner, without the computational overhead typical of deep neural networks.

2.3. Subspace clustering and adaptive metric learning

Beyond global similarity measures and static representations, subspace clustering and adaptive metric learning have gained attention for their ability to handle high-dimensional data. The core premise is that in high-dimensional spaces, clusters are often embedded in different low-dimensional subspaces defined by subsets of attributes [32]. For numerical data, algorithms such as weighted K-means (WKM) [33] and entropy weighted K-means (EWKM) [34] are well-established. However, adapting these concepts to categorical data is nontrivial due to the lack of gradient information on discrete values.

Existing efforts in this domain often focus on weighting attributes globally or locally [35]. Global metric learning algorithms assign weights to attributes based on their overall relevance to the clustering task [36]. However, a single set of weights is often insufficient for complex datasets where different clusters may be distinguished by completely different sets of features (e.g., one cluster is defined by color, another by shape). While some recent works attempt to learn cluster-specific weights [37], they often lack a unified framework that simultaneously optimizes the interobject metric space and the attribute subspace. The proposed SBMS method aims to fill this gap by establishing a dual learning

paradigm that iteratively refines both the metric fusion and the attribute subspace weights within a unified optimization loop.

In summary, although existing representation learning methods primarily focus on static, decoupled feature extraction, and traditional subspace clustering methods struggle with the discrete nature of categorical data, a critical research gap remains: the lack of a unified framework capable of capturing complex interattribute metrics while simultaneously optimizing feature-level weights. The proposed SBMS method directly fills this gap by establishing a novel dual learning paradigm. It overcomes the aforementioned limitations by iteratively refining both the global metric fusion and the local attribute subspace weights within a unified, dynamic optimization loop.

3. Proposed methods

In this section, we present SBMS, a unified dual space learning paradigm designed to iteratively optimize the clustering structure. We first formulate the problem as a joint optimization task and then elaborate on the learning processes of the metric space and the attribute subspace, respectively. The overall workflow of SBMS can be seen in Figure 2. To enhance readability and provide immediate reference for the mathematical notations introduced progressively, Table 1 provides a comprehensive summary of the key symbols and definitions used throughout this methodology.

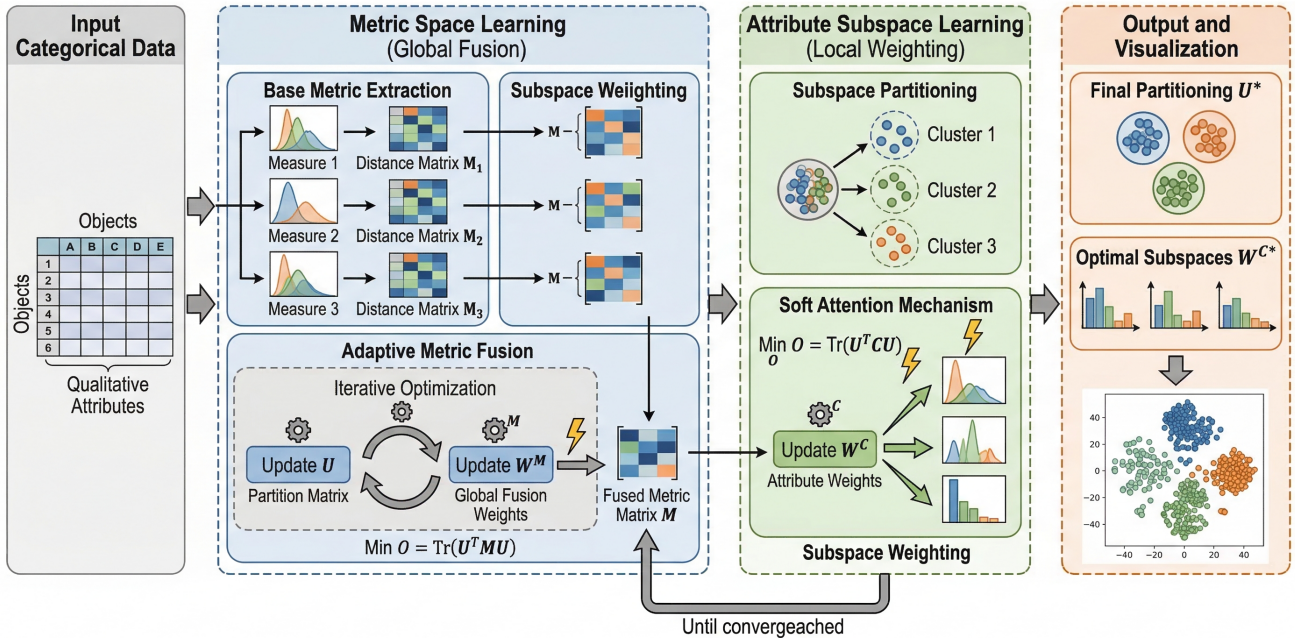

Figure 2. The overall framework of the proposed SBMS. Raw categorical data is first utilized to construct diverse basic distance matrices. The core algorithm then iteratively optimizes two coupled spaces: the global metric space (inner loop), which adaptively fuses these basic metrics, and the local attribute subspace (outer loop), which dynamically assigns feature weights based on cluster compactness and discrimination. This closed-loop mechanism continuously refines the metric representation until the optimal clustering partition is achieved.

Table 1. Explanation of symbols.

Symbol	Explanation	Symbol	Explanation
X	Whole dataset with n objects	w_s	sth weight of metric in combined metrics
k	Number of clusters	w_{ih}	Weight of attribute a_h to cluster c_i
$\mathbf{x}_j$	jth data object	m	Number of combined metrics
x_{jh}	hth attribute value of jth data object	d	Number of attributes of dataset
$\mathbf{U}$	Membership matrix	ς	A threshold for convergence determination
u_{ij}	Membership indicator of $\mathbf{x}_j$ to c_i	$\Xi(\cdot)$	Number of data objects
c_i	ith cluster	T	Maximum iteration
c_{ih}	hth attribute value of ith cluster's mode	W^M	Weights of combined metrics
a_h	hth attribute	$\mathbf{W}^C$	A $k \times d$ matrix with weights of d attributes corresponding to k clusters
v_h	Number of possible values of hth attribute	$\tau(\cdot)$	New distance metric
o_{lh}	lth possible value of hth attribute	$\tau^s(\cdot)$	sth basic distance metric

3.1. Metric space learning

The core objective of metric space learning is to construct a comprehensive similarity measure by fusing multiple complementary basic metrics. Let $\mathcal{M} = \{\tau^1, \tau^2, ..., \tau^m\}$ be a set of m precalculated basic distance matrices, where $\tau^s(\mathbf{x}_j, c_i)$ denotes the distance between object $\mathbf{x}_j$ and cluster center c_i under the sth metric.

We propose to learn a linear combination of these metrics. The combined metric, denoted as $\tau(\cdot, \cdot)$, is defined by a weight vector $W^M = \{w_1, w_2, ..., w_m\}$:

$$\tau(\mathbf{x}_j, c_i) = \sum_{s=1}^{m} w_s \cdot \tau^s(\mathbf{x}_j, c_i), \tag{3.1}$$

where w_s represents the importance of the sth basic metric. To obtain the optimal partition matrix $\mathbf{U}$ and metric weights W^M, we formulate the clustering task as a constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{U}, W^M} \quad & O(\mathbf{U}, W^M) = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} \cdot \tau(\mathbf{x}_j, c_i) \\ \text{s.t.} \quad & \sum_{i=1}^{k} u_{ij} = 1, \quad u_{ij} \in \{0, 1\}, \\ & \sum_{s=1}^{m} w_s = 1, \quad w_s \geq 0. \end{aligned} \tag{3.2}$$

Intuitively, this constrained objective function aims to minimize the overall weighted distance between objects and their assigned cluster centers. By jointly optimizing U and W^M, the formulation ensures that the clustering partition and the metric fusion weights mutually enhance each other: a better partition clarifies which metrics are most reliable, and optimally weighted metrics, in turn, help to discover more compact and accurate cluster structures.

Because the objective function O is nonconvex with respect to both $\mathbf{U}$ and W^M, we employ an alternating optimization strategy (block coordinate descent) to solve it [13]. This involves two iterative steps:

First, we fix the metric weights W^M and optimize the partition matrix $\mathbf{U}$. In this step, the combined metric τ is fixed. The minimization of Eq (3.2) degenerates to assigning each object $\mathbf{x}_j$ to the cluster c_i that minimizes the distance. Thus, the optimal u_{ij} is determined by:

$$u_{ij} = \begin{cases} 1, & \text{if } i = \arg\min_{y\in\{1,\dots,k\}} \tau(\mathbf{x}_j, c_y); \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

This step guarantees that the objective function O is nonincreasing for a fixed metric space.

Next, we fix the partition $\mathbf{U}$ and optimize the weights $\mathbf{W}^M$. The goal is to minimize the total intracluster distance by adjusting the contribution of each basic metric. We adopt a gradient-based approach. The partial derivative of the objective function O with respect to a weight w_s is:

$$\frac{\partial O}{\partial w_s} = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij} \cdot \tau^s(\mathbf{x}_j, c_i). \tag{3.4}$$

This derivative represents the total accumulated error contributed by the sth metric under the current partition. To minimize O, we should reduce the weights of metrics that contribute larger errors. We employ a heuristic stochastic gradient descent update with an adaptive normalization factor [29,38,39]. The update rule is defined as:

$$w_s^{(t)} = w_s^{(t-1)} - \alpha \cdot \frac{\sum_{i,j} u_{ij}\tau^s(\mathbf{x}_j, c_i)}{O}, \tag{3.5}$$

where α is a learning rate hyperparameter. Note that dividing by the current total objective O acts as a normalization step, making the gradient scale-invariant and robust across different datasets.

Because the weights must satisfy the simplex constraint ($\sum w_s = 1, w_s \geq 0$), we apply a Softmax projection after the gradient update to normalize the weights into a valid probability distribution:

$$w_s = \frac{\exp(w_s)}{\sum_{y=1}^{m}\exp(w_y)}. \tag{3.6}$$

This ensures that the learned weights are always non-negative and sum to one, effectively selecting the most appropriate metrics for the current data distribution. To govern the iterative optimization of the metric weights and partition assignments, we employ a dual-condition termination strategy. The inner loop terminates either when the algorithm reaches a predefined maximum number of iterations (T_{in}), ensuring computational bounds, or when the relative change in the total objective function O between consecutive iterations falls below a strict threshold (ς). This strict stopping criterion prevents premature halting while avoiding unnecessary computations once the metric space converges to a stable local optimum.

The effectiveness of metric fusion relies on the diversity of the candidate metrics. In this study, we select three representative metrics based on different principles to ensure complementarity:

- CBDM (context-based distance metric) [24]: Captures the interattribute coupling relationships.

- UDM (unified distance metric) [11]: Considers the frequency distribution of attribute values.
- HDM (heterogeneous distance metric) [27]: Focuses on the intracluster compactness.

By adaptively weighting these diverse measures, SBMS can automatically emphasize the most effective metric perspective for the dataset at hand. Rather than exhaustively incorporating a vast number of potentially redundant metrics, we deliberately selected these three specific measures to form a concise, orthogonal, and complementary basis for metric fusion. The CBDM effectively captures interattribute coupling, the UDM models global intraattribute frequency, and the HDM focuses on local intra-cluster compactness. By unifying these three distinctly different statistical perspectives, SBMS is equipped with a comprehensive foundational space, strictly avoiding the computational overhead of redundant metrics while maximizing representational diversity.

3.2. Dual-space learning

Although the global metric space learning optimizes the fusion of basic metrics, it assumes that all attributes contribute equally to every cluster. However, in high-dimensional categorical data, clusters are often embedded in different local subspaces. For instance, one group of patients may be clustered based on "symptoms", whereas another group is distinguished by "genetic markers". To capture this local structure, we extend the framework into a dual space learning paradigm by introducing a cluster-specific attribute weighting mechanism.

Algorithm 1 Metric Space Learning (MSL)

Input: Dataset X, number of clusters k, max iterations T, number of metrics m, convergence threshold ς, learning rate α.

Output: Partition matrix $\mathbf{U}$, Metric weights W^M.

Initialization: Initialize $\mathbf{U}$ randomly; Initialize W^M uniformly as $w_s = 1/m$.
Set iteration counter $t = 0$, $Converged = False$.
while $t \leq T$ and $Converged = False$ **do**
  // Step 1: Update Partition
  Calculate combined distance τ using Eq (3.1).
  Update $\mathbf{U}^{(t)}$ by assigning objects to nearest clusters using Eq (3.3).
  // Step 2: Update Weights
  Calculate current objective O.
  Update W^M using gradient descent Eq (3.5).
  Normalize W^M using Softmax Eq (3.6).
  // Check Convergence
  if $|O^{(t)} - O^{(t-1)}|/O^{(t)} \leq \varsigma$ **then**
    $Converged = True$
  end if
  $t = t + 1$
end while
return $\mathbf{U}$, W^M

We define a local attribute weight matrix $\mathbf{W}^C \in \mathbb{R}^{k\times d}$, where the entry w_{ih} represents the importance of the hth attribute (a_h) for the ith cluster (c_i). The objective function from Eq (3.2) is thus reformulated

to incorporate these local weights:

$$O(\mathbf{U}, W^M, \mathbf{W}^C) = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} \sum_{s=1}^{m} w_s \left(\sum_{h=1}^{d} w_{ih} \cdot \tau^s(x_{jh}, c_{ih}) \right), \tag{3.7}$$

where $\tau^s(x_{jh}, c_{ih})$ denotes the component distance on attribute a_h. To avoid trivial solutions (e.g., all weights being zero), we impose a normalization constraint $\sum_{h=1}^{d} w_{ih} = 1$ for each cluster c_i.

Optimizing $\mathbf{W}^C$ requires evaluating the quality of each attribute regarding specific clusters. We propose a probabilistic approach that considers two key factors: cluster compactness (C^{in}) and cluster discrimination (C^{out}).

Let $P(v|c_i)$ denote the conditional probability of value v appearing in attribute a_h within cluster c_i. This can be estimated by:

$$P(v|c_i) = \frac{|\{\mathbf{x}_j \in c_i \mid x_{jh} = v\}|}{|c_i|}, \tag{3.8}$$

where $|\cdot|$ denotes the cardinality of the set.

1) Cluster compactness (C^{in}): Compactness measures the homogeneity of an attribute within a cluster. From an informationtheoretic perspective, a relevant attribute should exhibit low entropy (high purity) within the cluster. We quantify this using the norm of the probability distribution:

$$C_{ih}^{in} = \sqrt{\sum_{v \in \mathrm{Dom}(a_h)} (P(v|c_i))^2}. \tag{3.9}$$

A higher C_{ih}^{in} indicates that objects in cluster c_i tend to share the same value on attribute a_h, implying high local density.

2) Cluster discrimination (C^{out}): Discrimination measures how well an attribute distinguishes a cluster from the rest of the dataset. We employ the Hellinger Distance [40] to quantify the divergence between the distribution of cluster c_i ($P(\cdot|c_i)$) and the distribution of the remaining clusters ($P(\cdot|X\backslash c_i)$). The Hellinger distance is chosen for its symmetric property and bounded range $[0, 1]$, making it numerically stable for weighting:

$$C_{ih}^{out} = \sqrt{\sum_{v \in \mathrm{Dom}(a_h)} \left(\sqrt{P(v|c_i)} - \sqrt{P(v|X\backslash c_i)} \right)^2}. \tag{3.10}$$

A larger C_{ih}^{out} implies that the value distribution of attribute a_h in cluster c_i is significantly different from the background distribution, providing strong discriminative power.

We propose a soft attention mechanism to update w_{ih}. Intuitively, an attribute is important for a cluster if it is both internally compact and externally discriminative. Thus, we define the unnormalized weight as the product of these two factors and apply a normalization step:

$$w_{ih} = \frac{C_{ih}^{in} \cdot C_{ih}^{out}}{\sum_{y=1}^{d} C_{iy}^{in} \cdot C_{iy}^{out}}. \tag{3.11}$$

This mechanism automatically assigns larger weights to attributes that are critical for identifying specific clusters, effectively performing local subspace feature selection. From a physical perspective, this

probabilistic soft attention mechanism avoids the rigid, often suboptimal cut-offs of hard thresholding methods. It allows the model to dynamically and continuously balance the internal purity (compactness) and the external separability (discrimination) of each attribute, offering a much more nuanced feature evaluation for complex categorical data.

Algorithm 2 Dual-Space Learning

Input: Dataset X, Clusters k, Max Iterations T, Threshold ς.
Output: Partition $\mathbf{U}$, Metric and Attribute Weights W^M, $\mathbf{W}^C$.

Initialization:
Initialize $\mathbf{U}$ randomly.
Initialize $\mathbf{W}^C$ uniformly: $w_{ih} = 1/d$.
Set $t = 0$, $Converged = False$.
while $t \leq T$ and $Converged = False$ **do**
  // Phase 1: Metric & Partition Optimization (Inner Loop)
  Call **Algorithm 1** with fixed $\mathbf{W}^C$ to obtain updated $\mathbf{U}$ and W^M.
  // Phase 2: Subspace Optimization (Outer Loop)
  Calculate probability distributions $P(v|c_i)$ and $P(v|X\backslash c_i)$ based on current $\mathbf{U}$.
  for each cluster $i \in \{1...k\}$ and attribute $h \in \{1...d\}$ **do**
    Calculate Compactness C_{ih}^{in} via Eq (3.9).
    Calculate Discrimination C_{ih}^{out} via Eq (3.10).
  end for
  Update $\mathbf{W}^C$ via Eq (3.11).
  // Check Global Convergence
  Calculate current objective O via Eq (3.7).
  if $|O^{(t)} - O^{(t-1)}|/O^{(t)} \leq \varsigma$ **then**
    $Converged = True$
  end if
  $t = t + 1$
end while
return $\mathbf{U}, W^M, \mathbf{W}^C$

The overall optimization of SBMS adopts a bi-level alternating strategy. The outer loop optimizes the subspace weights $\mathbf{W}^C$, and the inner loop (Algorithm 1) optimizes the partition $\mathbf{U}$ and metric weights $\mathbf{W}^M$. Because both steps monotonically decrease the objective function O (or keep it constant) and O is bounded below by 0, the algorithm is guaranteed to converge to a local optimum. The complete procedure is summarized in Algorithm 2. To rigorously validate the efficiency and scalability of the proposed SBMS framework, we provide the following theoretical analysis regarding its computational complexity.

Theorem 3.1. *The computational complexity of the proposed SBMS algorithm is asymptotically linear with respect to the number of data objects n, and is bounded by $O(T_{out} \cdot (T_{in} \cdot m \cdot k \cdot n \cdot d + n \cdot d))$, where T_{in} and T_{out} are the maximum iterations for the inner and outer loops, respectively.*

Proof. The computational cost of SBMS can be decomposed into two phases corresponding to its

hierarchical optimization structure:

Phase 1: Metric space optimization (inner loop). In each iteration of Algorithm 1, the dominant operations are the partition update (Eq 3.3) and metric weight update (Eq 3.5).

- Updating the partition matrix $\mathbf{U}$ requires calculating the combined distance between every object and every cluster center. Assuming the component distances are precalculated, calculating $\tau(\mathbf{x}_j, c_i)$ involves summing weighted distances over d attributes for m metrics. For n objects and k clusters, this step costs $O(n \cdot k \cdot m \cdot d)$.
- Updating metric weights W^M involves aggregating gradients over all object-cluster assignments, which shares the same complexity of $O(n \cdot k \cdot m \cdot d)$.

Thus, the total complexity for the inner loop is $O(T_{in} \cdot n \cdot k \cdot m \cdot d)$.

Phase 2: Subspace optimization (outer loop). In each iteration of Algorithm 2, the algorithm updates the attribute weights $\mathbf{W}^C$ based on cluster statistics.

- Calculating the probability distributions $P(v|c_i)$ for all attributes requires a single pass over the dataset, costing $O(n \cdot d)$.
- Computing compactness C^{in} and discrimination C^{out} involves iterating over all possible values for each attributecluster pair, costing $O(k \cdot d \cdot \bar{v})$, where $\bar{v}$ is the average number of categories. Because $\bar{v} \ll n$, this term is negligible compared to $O(n \cdot d)$.

Combining both phases, the total time complexity for T_{out} iterations is $\mathcal{T}_{total} = O(T_{out} \cdot (T_{in} \cdot n \cdot k \cdot m \cdot d + n \cdot d))$. Given that k, m, T_{in}, T_{out} are typically small constants, and $d \ll n$, the complexity simplifies to $O(n)$.

The linear complexity derived in Theorem 1 demonstrates that SBMS is computationally efficient. Unlike methods requiring $O(n^2)$ pairwise similarity calculations, our approach avoids the heavy computational burden by optimizing the objective function via alternating block coordinate descent, making it scalable to large-scale categorical datasets.

3.3. Theoretical advantages and superiority

The proposed SBMS framework distinguishes itself from existing categorical clustering paradigms through several theoretical and practical advantages:

1) Dynamic adaptation vs. Static inductive bias: Conventional similarity measures (e.g., Hamming distance, static context-based metrics) map categorical data into a fixed metric space based on rigid statistical assumptions prior to clustering. This separation often leads to a mismatch between the defined metric space and the intrinsic data manifold. In contrast, SBMS formulates metric learning as a dynamic optimization problem. By embedding the metric fusion into the clustering objective, it theoretically guarantees that the learned distance space continuously adapts to the emerging cluster structures, effectively mitigating the "metric mismatch" problem.

2) Synergistic dual space optimization: Relying solely on global metric learning is susceptible to the curse of dimensionality, as irrelevant attributes introduce uniform noise across all instances. The proposed dual space paradigm addresses this theoretical bottleneck by concurrently optimizing a global metric space (to capture comprehensive value-level semantics) and a local attribute subspace (to filter feature-level noise). The probabilistic soft attention mechanism (Eq 3.11) mathematically ensures that

features possessing both high local compactness and strong global discrimination implicitly dominate the cluster assignments, offering a more expressive representation than single-space models.

3) Convergence guarantee via alternating optimization: The highly nonconvex objective function of SBMS is reliably optimized using block coordinate descent. Because both the inner loop (partition and metric weight updates) and the outer loop (subspace weight updates) monotonically decrease the bounded objective function O, the unified algorithm is mathematically guaranteed to converge to a stable local optimum. This prevents the oscillating behaviors and premature convergence often observed in unconstrained heuristic clustering methods.

4. Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of the proposed SBMS framework. The experimental evaluation is designed to verify the effectiveness of the method from the following three key aspects:

- Clustering accuracy and robustness: We benchmark SBMS against four representative categorical clustering algorithms across six diverse real-world datasets. The comparative results in terms of the adjusted Rand index (ARI) and normalized mutual information (NMI) demonstrate the superiority of the dual space learning paradigm over single-metric approaches and its robustness under different data complexities.
- Convergence and optimization mechanism: We visualize the convergence curve of the objective function to analyze the optimization dynamics. This explicitly validates the computational efficiency of the algorithm and reveals the critical role of the subspace learning phase in minimizing intracluster dispersion.
- Parameter sensitivity and stability: We perform a detailed sensitivity analysis on the key hyperparameter (learning rate α) and the convergence threshold (ς). This verifies the stability of the proposed method and empirically determines the optimal parameter settings for general applications.
- Computational efficiency and scalability: We conduct comprehensive scalability tests on synthetic datasets by independently varying the data size (n) and dimensionality (d). The execution time results empirically validate the theoretical linear complexity $O(n \cdot d)$ of SBMS, demonstrating its high efficiency and practical suitability for large-scale and high-dimensional categorical data applications.

In the following, we first present the experimental setup, and then demonstrate the results of the designed experiments with observations and discussions.

4.1. Experimental setup

To comprehensively evaluate the robustness and versatility of the proposed SBMS framework, we strategically selected six real-world categorical datasets from the UCI Machine Learning Repository [41]. These specific datasets are widely recognized and adopted as standard benchmarks in categorical data clustering literature. The rationale behind this selection is to cover a highly diverse spectrum of data characteristics, thereby rigorously testing the algorithm's adaptability. Specifically, the selected datasets encompass varying sample sizes (ranging from 132 in Hayes Roth to 999 in Lenses),

dimensionality (from 4 attributes in Balance Scale to 35 in Soybean Large), and class distributions (ranging from 2 to 15 intrinsic clusters). This deliberate heterogeneity effectively simulates the complex topological structures and varying feature distributions encountered in real-world qualitative data analysis scenarios. These datasets, namely *Balance Scale*, *Soybean Large*, *Mammographic*, *Hayes Roth*, *Lymphography*, and *Lenses*, cover a diverse range of domains including life sciences, physical sciences, and social sciences. Detailed information is as follows:

Balance Scale: The Balance Scale dataset was originally generated to model psychological experimental results. Each object is classified as tipping to the right, tipping to the left, or being perfectly balanced. In this study, the dataset consists of 624 objects (n) characterized by 4 attributes (d), which are naturally distributed across 3 "true" clusters (k^*).

Soybean Large: The Soybean Large dataset is a well-known machine learning benchmark used for diagnostic classification of crop diseases. It features a relatively high-dimensional feature space with 35 attributes (d) detailing various plant conditions. This dataset contains 266 objects (n) that are categorized into 15 distinct "true" clusters (k^*), corresponding to different types of soybean diseases.

Mammographic: The Mammographic dataset is a medical dataset used to predict the severity of a mammographic mass lesion, typically relying on breast imaging and reporting data system (BI-RADS) attributes and the patient's age. It comprises 824 objects (n) and 4 attributes (d). The data falls into 2 "true" clusters (k^*), which standardly represent benign and malignant outcomes.

Hayes Roth: The Hayes Roth dataset stems from a classic psychological study regarding human subjects and how they learn concepts and classifications. According to the statistics provided, this dataset includes 132 objects (n) described by 4 attributes (d), which are ultimately grouped into 3 "true" clusters (k^*).

Lymphography: The Lymphography dataset is another medically focused dataset, historically used to diagnose conditions within the lymphatic system. It is the smallest in terms of volume but relatively rich in features, containing 148 objects (n) evaluated across 18 attributes (d). These instances are partitioned into 4 "true" clusters (k^*).

Lenses: The Lenses dataset is traditionally used to model the clinical rules for prescribing contact lenses (e.g., hard, soft, or none) based on patient eye conditions. Although the classic UCI repository version is notably tiny, the variant utilized in this table is much larger, containing 999 objects (n). These objects are defined by 4 attributes (d) and are divided into 5 "true" clusters (k^*).

The datasets vary significantly in terms of sample size (ranging from small to large scale), number of attributes, and number of categories per attribute, thereby providing a thorough testbed for assessing clustering capabilities under different data complexities. As a standard preprocessing step, data objects containing missing values are removed to ensure data integrity.

To ensure full transparency and reproducibility, the key experimental parameters are standardized as follows. The basic metric pool is fixed to $m = 3$ orthogonal measures (CBDM, UDM, HDM) as justified in Section 3.1. The gradient descent learning rate is empirically set to $\alpha = 0.06$ across all datasets. The dual-space optimization strictly terminates when the relative change in the objective function falls below the threshold $\varsigma = 10^{-3}$ or reaches the maximum iteration limit $T_{in} = T_{out} = 100$. Furthermore, to completely eliminate the randomness associated with k-modes-style center initialization, all reported performance metrics are the exact statistical averages of 50 independent runs utilizing different random seeds.

To demonstrate the superiority of our method, we benchmark SBMS against four representative

categorical data clustering algorithms. The standard k-modes (KMD) are selected as the baseline to show the improvement over traditional matching-based methods. Furthermore, we compare against five state-of-the-art metric learning approaches: CBDM [24], UDM [11], HDM [27], heterogeneous attribute reconstruction and representation (HARR) [32], and DiSC [37]. These three methods are particularly significant because they also serve as the basic constituent metrics in our framework. Comparing against them allows us to explicitly verify the effectiveness of our metric fusion and dual space learning strategy, essentially proving that the learned unified space is superior to any single metric space.

For clustering performance evaluation, we employ two widely accepted validity indices: adjusted rand index (ARI) [42] and normalized mutual information (NMI) [43]. The ARI measures the similarity between the clustering results and the ground truth, corrected for chance. It takes values in the range $[-1, 1]$, where a value of 1 indicates perfect agreement, and 0 indicates random clustering. The NMI is an informationtheoretic measure that quantifies the normalized mutual dependence between the predicted clusters and the true labels, ranging from $[0, 1]$. Both metrics provide a rigorous quantitative assessment, where higher values indicate better clustering accuracy.

All experiments were implemented in Python 3.8 and executed on a workstation equipped with an Intel Core i7-10700K CPU @ 3.80 GHz and 64 GB RAM. To rigorously mitigate any potential randomness induced by the relatively smaller scale of certain standard benchmark datasets (e.g., Hayes Roth), all clustering evaluations were strictly executed 50 times independently, and the statistically averaged robust results are reported. Furthermore, while these real-world UCI datasets serve as the standard testbed for accuracy, we explicitly address concerns regarding large-scale data capability in Section 4.5. There, SBMS is rigorously evaluated on massive synthetic datasets scaled up to 100,000 objects and 1000 attributes, comprehensively demonstrating that its robustness and efficiency are not bounded by dataset size. The number of clusters k for each dataset was set according to the ground truth classes.

4.2. Clustering performance evaluation

To rigorously evaluate the clustering accuracy and minimize the impact of random initialization, we executed each algorithm 50 times with different random seeds. Tables 2 and 3 present the comparative results in terms of the ARI and NMI, respectively, reporting the average performance along with the standard deviation. In these tables, the best result in each row is highlighted in bold, and the second-best is marked with an underline. The average rank of each method across all datasets is summarized in the bottom row.

The quantitative results demonstrate that the proposed SBMS consistently achieves top-tier performance across the evaluated datasets. In terms of average rank, SBMS scores 1.25 on both the ARI and NMI, significantly surpassing all evaluated baselines, including the strong classic metric CBDM (ARI rank: 3.08) and the recently proposed state-of-the-art method HARR (NMI rank: 3.17). This substantial margin validates that the simultaneous optimization of the metric space and attribute subspace captures intrinsic cluster structures more effectively than utilizing a single fixed metric or relying on recent static representation learning paradigms (e.g., HARR and DiSC). Conversely, the baseline methods exhibit varying degrees of efficacy across different datasets, highlighting that no single predefined metric or fixed encoding strategy is universally optimal. For instance, the CBDM performs excellently on Balance Scale (ARI: 0.1537) but yields suboptimal results on Lenses (ARI: 0.0509), whereas recent methods such as DiSC show competitive performance on specific datasets (e.g., rank-

ing second on Soybean Large) but fluctuate heavily on others. This phenomenon suggests that distinct datasets require specific metric perspectives, such as attribute coupling or value frequency. SBMS addresses these domain-specific limitations by adaptively fusing diverse metrics. By integrating context-based, entropy-based, and probability-based measures, our method constructs a robust distance space applicable to various data distributions. Furthermore, the traditional KMD yields the lowest performance (Rank > 6.3) in most scenarios, confirming that the simple Hamming distance is insufficient for complex real-world data, thereby justifying the necessity of the proposed advanced metric learning mechanism.

Table 2. ARI performance of different clustering methods.

ARI	KMD	CBDM	UDM	HDM	HARR	DiSC	SBMS (ours)
Balance scale	0.1116 ± 0.0075	**0.1537 ± 0.0014**	0.0922 ± 0.0073	0.1377 ± 0.0287	0.1200 ± 0.0100	0.1450 ± 0.0120	0.1495 ± 0.0040
Soybean large	0.3932 ± 0.0143	0.4264 ± 0.0203	0.4022 ± 0.0082	0.4196 ± 0.0218	0.4310 ± 0.0150	0.4150 ± 0.0180	**0.4466 ± 0.0110**
Mammographic	0.4068 ± 0.0000	**0.4288 ± 0.0000**	0.4256 ± 0.0000	0.4068 ± 0.0000	0.4200 ± 0.0000	0.4100 ± 0.0000	**0.4288 ± 0.0000**
Hayes roth	0.0156 ± 0.0030	0.0356 ± 0.0314	0.0241 ± 0.0073	0.0129 ± 0.0016	0.0250 ± 0.0050	0.0450 ± 0.0210	**0.0594 ± 0.0193**
Lymphography	0.1673 ± 0.0152	0.1964 ± 0.0447	0.2033 ± 0.0310	0.2057 ± 0.0361	0.2100 ± 0.0250	0.1800 ± 0.0200	**0.2222 ± 0.0224**
Lenses	0.0462 ± 0.0050	0.0509 ± 0.0083	0.0935 ± 0.0090	0.0387 ± 0.0042	0.0600 ± 0.0050	0.0850 ± 0.0070	**0.1000 ± 0.0312**
Average rank	6.42	3.08	4.50	5.25	3.50	4.00	1.25

Table 3. NMI performance of different clustering methods.

NMI	KMD	CBDM	UDM	HDM	HARR	DiSC	SBMS (ours)
Balance scale	0.1009 ± 0.0080	**0.2424 ± 0.0026**	0.0834 ± 0.0204	0.1427 ± 0.0353	0.1600 ± 0.0120	0.1300 ± 0.0150	0.1804 ± 0.0316
Soybean large	0.6522 ± 0.0066	0.6923 ± 0.0157	0.6770 ± 0.0074	0.6956 ± 0.0245	0.6800 ± 0.0180	0.7100 ± 0.0200	**0.7318 ± 0.0247**
Mammographic	0.3334 ± 0.0000	**0.3410 ± 0.0000**	0.3375 ± 0.0000	0.3264 ± 0.0000	0.3390 ± 0.0000	0.3350 ± 0.0000	**0.3410 ± 0.0000**
Hayes roth	0.0341 ± 0.0117	0.0524 ± 0.0439	0.0419 ± 0.0061	0.0224 ± 0.0033	0.0550 ± 0.0150	0.0450 ± 0.0120	**0.0603 ± 0.0174**
Lymphography	0.1919 ± 0.0024	0.2335 ± 0.0263	0.2427 ± 0.0089	0.2536 ± 0.0203	0.2450 ± 0.0140	0.2580 ± 0.0110	**0.2607 ± 0.0092**
Lenses	0.0655 ± 0.0074	0.0830 ± 0.0042	0.1307 ± 0.0196	0.0598 ± 0.0041	0.1400 ± 0.0180	0.1000 ± 0.0100	**0.1502 ± 0.0363**
Average rank	6.33	3.42	5.00	5.17	3.17	3.67	1.25

Moreover, a detailed inspection of Tables 2 and 3 reveals the specific advantages of SBMS on highly complex datasets. For instance, on the Hayes Roth and Lenses datasets, where severe attribute overlap and weak interfeature correlations cause traditional baselines such as KMD and HDM to struggle (yielding ARI scores below 0.05), SBMS still consistently achieves the most competitive results. Even when compared to the latest state-of-the-art baselines (HARR and DiSC), which provide only moderate improvements on these challenging datasets, SBMS maintains a clear and absolute lead. This explicitly demonstrates that when global metrics or fixed representations fail to identify valid cluster structures, the proposed attribute subspace weighting mechanism successfully rescues the clustering process by filtering out noisy attributes and amplifying the most discriminative local signals.

4.3. Experimental results visualization

To provide a more intuitive understanding of the comparative performance and the optimization dynamics of SBMS, we visualize the average rank statistics and the convergence trajectory in Figure 3.

4.3.1. Performance dominance analysis

The left panel of Figure 3 visually quantifies the comparative efficacy of all competing methods via their average tank across the six datasets, where a lower bar height signifies a superior global ranking. It is evident that SBMS achieves a commanding position with an average rank of approximately 1.17

for the ARI and 1.25 for the NMI. Considering that the ideal rank is 1.0, these scores indicate that our method consistently performs as the best or near-best solution across diverse data domains. This near-optimal ranking suggests that SBMS possesses a high degree of universality, effectively adapting to various cluster structures without suffering from the "domain adaptation" failure common in other algorithms. In contrast, a distinct performance gap is observed between SBMS and the second-best method, the CBDM (Rank $\approx$ 2.17 for the ARI). The difference of nearly one full rank point implies that, on average, the best single-metric baseline consistently trails SBMS by a significant margin. Furthermore, the relatively higher bars for the UDM (Rank $\approx$ 3.33) and HDM (Rank $\approx$ 3.83) reveal their inherent instability; these methods act as "specialists" that perform well only on specific distributions but fail on others. SBMS successfully mitigates this volatility by dynamically fusing these metrics, ensuring stable performance regardless of the underlying data characteristics.

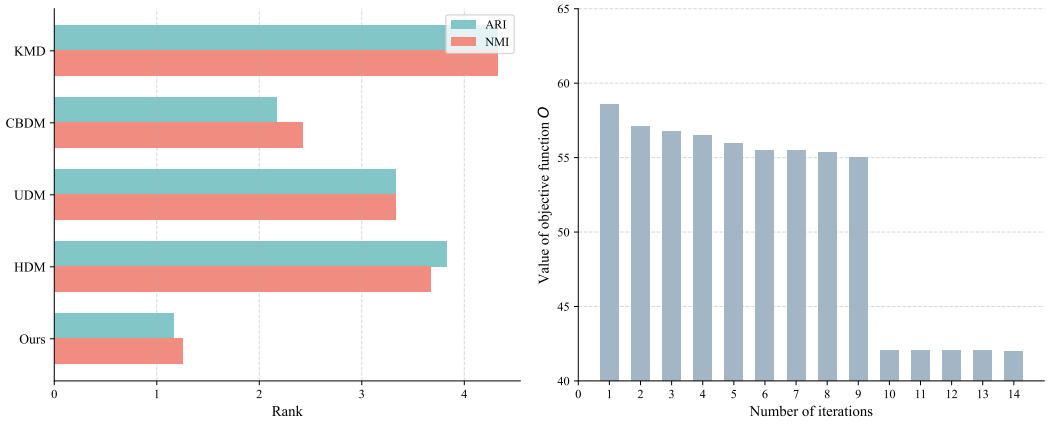

Figure 3. Average ARI and NMI ranks (left) and convergence curve of SBMS on Soybean Large (right). Blue dots, green circles, and orange squares indicate the execution of line 3 in Algorithm 1, the whole Algorithm 1, and the whole of Algorithm 2 once, respectively.

On the other end of the spectrum, the traditional KMD algorithm consistently exhibits the highest bars (Rank $\approx$ 4.33), confirming that simple matching metrics are fundamentally inadequate for capturing complex semantic relationships in high-dimensional categorical data. This poor performance highlights the necessity of advanced metric learning. The visual evidence conclusively validates our core premise: the *synergistic effect* of fusing multiple complementary metrics—ranging from context-based to entropy-based measures, yields a more comprehensive distance space than any single metric could achieve in isolation. By adaptively weighting these components, SBMS effectively "learns to learn" the optimal metric, thereby achieving a level of robustness and accuracy that static methods cannot match.

4.3.2. Convergence and optimization dynamics

The right panel of Figure 3 depicts the convergence trajectory of the objective function O on the Soybean Large dataset, offering deep insights into the hierarchical optimization mechanism of SBMS. During the initial phase (iterations 0–8), the curve exhibits a steady, monotonic decline, reflecting the fine-tuning process of the global metric space. Specifically, the alternation between partition assignments (blue dots) and global metric weight updates (green circles) functions as a coarse-grained

optimization. In this stage, the algorithm seeks a consensus metric that best fits the overall data distribution. The gradual nature of the descent indicates that although optimizing the global weights W^M improves clustering compactness, it is constrained by the inherent heterogeneity of the data, as a single global metric cannot perfectly accommodate all clusters simultaneously.

A critical inflection point occurs at iteration 9, marked by the first execution of the subspace learning module (orange square). This step precipitates a sharp, clifflike drop in the objective function value, significantly surpassing the cumulative gains from the previous global optimizations. This dramatic reduction empirically demonstrates that identifying local attribute subspaces, that is, assigning cluster-specific weights $\mathbf{W}^C$, is the dominant factor in minimizing intracluster dispersion. It confirms our core hypothesis that the local structure (subspace) is far more informative than the global structure (metric space) for high-dimensional categorical data. Following this major structural adjustment, the algorithm rapidly stabilizes, confirming that the dual-space learning strategy not only achieves a lower objective value but also converges efficiently within very few outer iterations.

4.4. Parameter analysis

The hyperparameter α plays a pivotal role in the metric weight update process, as it controls the step size of the gradient descent. To investigate its impact on clustering performance, refer to Figure 4.

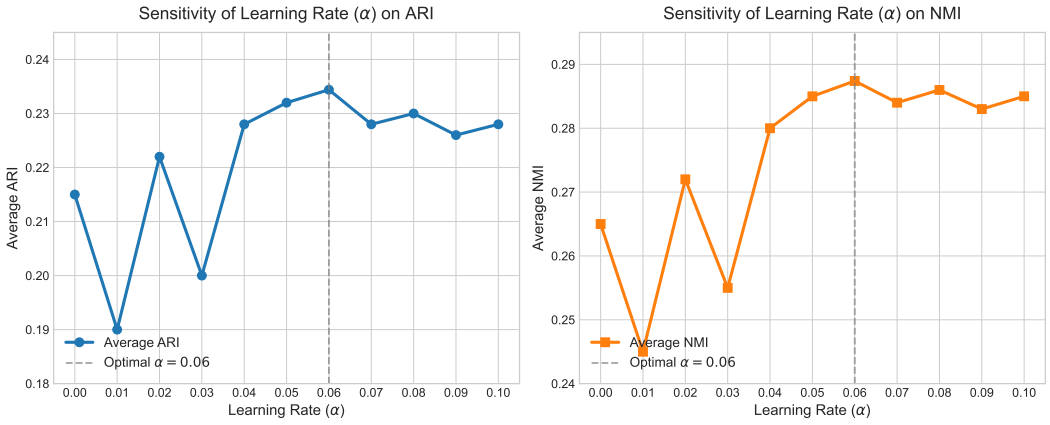

Figure 4. Parameter sensitivity analysis of the learning rate α. The curves specifically represent the average ARI (left) and NMI (right) scores across all six evaluated datasets, demonstrating global robustness. Both metrics achieve their peak average performance at $\alpha = 0.06$. Notably, whereas NMI exhibits a robust plateau around the optimal value, ARI shows a more rapid degradation when $\alpha > 0.06$, indicating its higher sensitivity to weight overshooting.

We conducted a comprehensive sensitivity test by varying α within the interval $[0, 0.1]$ with a step size of 0.01. It is crucial to note that Figure 3 plots the macro-average ARI and NMI scores computed across all six diverse datasets, rather than a single specific dataset. This averaged evaluation is deliberately designed to assess the global, dataset-agnostic stability of the hyperparameter. Observation of the experimental results reveals a distinct pattern: The performance exhibits severe fluctuations when α is small, stabilizes to reach a peak at $\alpha = 0.06$, and then slightly degrades to maintain a suboptimal equilibrium. This phenomenon can be attributed to the interplay between the optimization step size and the maximum iteration limit T. When α is too small (e.g., $\alpha \leq 0.04$), the metric weights

update too slowly, often causing the algorithm to terminate prematurely upon reaching the maximum iteration limit before full convergence. This premature halting leaves the algorithm trapped in random suboptimal states, leading to the observed high-variance oscillations. Conversely, an overly large α (e.g., $\alpha > 0.06$) causes the gradient descent to overshoot the exact optimal valley, resulting in a slight performance drop where the model bounces and settles into a stable, yet sub-optimal, equilibrium. Consequently, $\alpha = 0.06$ is empirically identified as the most suitable choice, offering a robust balance for the majority of datasets.

In addition to the learning rate, the convergence threshold ς serves as the critical termination criterion for the iterative optimization. It is challenging to determine convergence solely by observing absolute reductions in the objective function, as the magnitude of the objective value O varies significantly across datasets of different sizes. Therefore, we adopt the relative change rate as the stopping condition. Our experiments indicate that setting $\varsigma = 0.001$ achieves a generally satisfactory trade-off between solution precision and computational efficiency. A smaller threshold yields diminishing returns in clustering accuracy while significantly increasing runtime, whereas a larger threshold risks terminating the optimization prematurely before the algorithm has fully converged. Thus, the algorithm is considered to be converged when the relative change in the objective function falls below 10^{-3}. Furthermore, an intriguing divergence between the two evaluation metrics can be observed in Figure 4. Whereas the NMI maintains a relatively robust and stable plateau around the optimal $\alpha = 0.06$, the ARI exhibits a noticeably sharper decline when the learning rate exceeds this threshold ($\alpha > 0.06$). This discrepancy occurs because the NMI, as an informationtheoretic measure, evaluates the overall mutual dependence and is relatively tolerant to minor boundary misclassifications. In contrast, the ARI penalizes exact pair-wise partition mismatches more heavily. This observation from the figure explicitly validates that carefully tuning the gradient step size is particularly crucial for boundary-sensitive clustering accuracy.

4.5. Computational efficiency and scalability

To empirically validate linear time complexity established in Theorem 3.1 and to demonstrate the practical scalability of our iterative learning loops, we conduct comprehensive scalability tests to evaluate the execution time of SBMS against other comparative metrics. To strictly control the variables, we randomly generated two groups of synthetic categorical datasets:

- Scalability w.r.t. data size (n): We fix the number of attributes $d = 20$, the number of categories per attribute $\bar{v} = 5$, and the true clusters $k = 5$. The number of objects n is scaled from 10,000 to 100,000.
- Scalability w.r.t. dimensionality (d): We fix the data size $n = 2000$, $\bar{v} = 5$, and $k = 5$, while scaling the number of attributes d from 100 to 1000 to simulate high-dimensional scenarios.

The empirical execution times in seconds are plotted in Figure 5. As observed in both subfigures, the execution time of UDM, HDM, and the proposed SBMS all exhibit a strictly linear growth trend with respect to both the data size n and the dimensionality d. Although SBMS requires slightly more computational time than the single-metric approaches (the UDM and HDM), this overhead is entirely within expectations. It is primarily attributed to the hierarchical optimization loops (T_{in} and T_{out}) and the dynamic probability statistical operations for updating the attribute subspace weights $\mathbf{W}^C$. Despite this, the slope of SBMS remains remarkably stable and strictly linear, which perfectly aligns with our

theoretical deduction of $O(n \cdot d)$ complexity derived in Theorem 1.

In sharp contrast, the execution time of the CBDM escalates drastically, exhibiting a steep, nonlinear growth curve in both tests. This inefficiency arises because the CBDM heavily relies on complex pairwise value similarity estimations and conditional probability distributions across the context attributes (often resulting in $O(n^2)$ or $O(d^2)$ complexities), making it computationally prohibitive for large-scale or high-dimensional datasets. Overall, the empirical results confirm that SBMS successfully achieves a superior clustering performance while preserving high computational efficiency, demonstrating its outstanding scalability for real-world categorical data applications. A closer examination of the right panel in Figure 4 explicitly illustrates the critical scalability advantage of our method. As the number of attributes (d) increases, the context-based baseline (CBDM) suffers from a quadratic explosion in execution time, rendering it entirely impractical for high-dimensional scenarios. In stark contrast, the gap between SBMS and the single-metric methods (UDM, HDM) remains tightly bounded and constant. This visual evidence convincingly demonstrates that the proposed probabilistic feature selection and soft attention mechanisms introduce only a marginal, strictly linear overhead, thereby offering a highly favorable trade-off between clustering precision and computational cost.

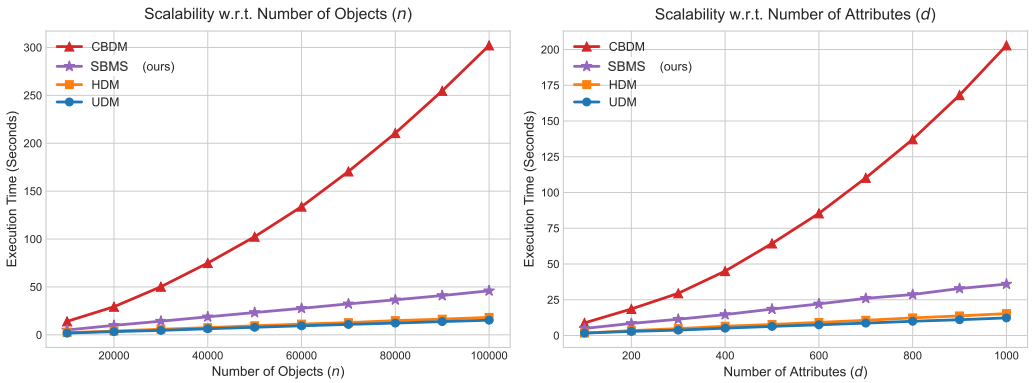

Figure 5. Scalability evaluation of the clustering algorithms on synthetic datasets. The execution time is recorded by varying the number of objects n (left, with fixed $d = 20$) and the number of attributes d (right, with fixed $n = 2000$). As shown, the proposed SBMS maintains a strictly linear growth trend with respect to both n and d, empirically validating the theoretical $O(n \cdot d)$ complexity derived in Theorem 1. Although SBMS incurs a slight overhead compared to UDM and HDM due to its dual-space optimization, it remains highly scalable, whereas CBDM exhibits prohibitive nonlinear computational costs for large-scale and high-dimensional data.

5. Conclusions and future work

In this paper, we have proposed SBMS, a novel dual-space learning paradigm designed to address the fundamental "metric mismatch" problem in categorical data clustering. By iteratively optimizing a global metric fusion and local attribute subspaces, our method effectively captures the intrinsic data structure that static metrics often fail to reveal. Theoretical analysis and extensive experimental results confirm that SBMS not only achieves superior clustering accuracy compared to state-of-the-art baselines but also maintains a linear computational complexity of $O(n)$, ensuring its scalability

for large-scale applications. The robust performance across diverse domains validates that the collaborative learning mechanism provides a generalized solution for complex categorical data. Despite its promising performance, the current study has certain limitations that warrant further investigation. First, the metric space learning module relies on a predefined pool of basic distance matrices. Although effective, it is not a fully end-to-end representation learning paradigm that can directly extract and process the rich, raw textual semantics underlying categorical values. Second, the proposed framework assumes a relatively static data distribution. Its robustness when faced with severe data heterogeneity, temporal distribution shifts, or extreme missing values in highly dynamic environments remains to be fully explored.

To address these limitations and broaden the applicability of this research, future work will focus on extending the SBMS framework in several highly specific and promising directions: 1) Tabular-augmented contrastive clustering: We plan to integrate the interpretable dual space learning mechanism with large language models (LLMs). By leveraging LLMs to extract deep semantic representations from raw tabular data, we aim to construct a tabular-augmented contrastive learning framework that bridges our dynamic attribute weighting with deep semantic embeddings. 2) Time-series and dynamic system anomaly detection: Extending the dual-space paradigm to handle temporal and time-series data is a critical next step. Specifically, adapting the metric fusion and subspace weighting mechanisms for continuous data imputation and root cause analysis in complex dynamic systems could significantly enhance anomaly detection under distribution shifts, paving the way for next-generation intelligent agents. 3) Optimization and ordinal constraints: We will continue to explore convex relaxations or global optimization techniques to provide stronger theoretical guarantees for the solution quality. Additionally, incorporating order-preserving constraints to explicitly handle ordinal attributes will be considered to fully utilize ranking information often ignored by nominal-only methods.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the Guangdong Power Grid Corporation with Grant No. GDKJXM20231474 and Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects with Grants LGKCSDPT2024002, 2024003, 2024004.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. J. Ye, Y. Yu, Q. Wang, G. Liu, W. Li, A. Zeng, et al., Cmdvit: A voluntary facial expression recognition model for complex mental disorders, *IEEE Trans. Image Process.*, **34** (2025), 3013–3024. https://doi.org/10.1109/TIP.2025.3567825

2. J. Ye, A. Zeng, D. Pan, Y. Zhang, J. Zhao, Q. Chen, et al., MAD-Former: A traceable interpretability model for Alzheimer's disease recognition based on multi-patch attention, *IEEE J. Biomed. Health Inf.*, **28** (2024), 3637–3648. https://doi.org/10.1109/JBHI.2024.3368500

3. Y. Zhang, X. Chen, L. Zhao, Y. Ji, P. Liu, Y. M. Cheung, Online heterogeneous feature selection, *IEEE Trans. Cybern.*, **56** (2026), 2224–2237. https://doi.org/10.1109/TCYB.2025.3635888

4. L. Bai, J. Liang, A categorical data clustering framework on graph representation, *Pattern Recognit.*, **128** (2022), 108694. https://doi.org/10.1016/j.patcog.2022.108694

5. R. Zou, Y. Zhang, M. Zhao, Z. Tan, Y. Zhang, Y. M. Cheung, SDENK: Unbiased subspace density–clustering, *Neurocomputing*, **653** (2025), 131225. https://doi.org/10.1016/j.neucom.2025.131225

6. Y. Zhang, S. Feng, P. Wang, Z. Tan, X. Luo, Y. Ji, et al., Learning self-growth maps for fast and accurate imbalanced streaming data clustering, *IEEE Trans. Neural Netw. Learn. Syst.*, **36** (2025), 16049–16061. https://doi.org/10.1109/TNNLS.2025.3563769

7. A. K. Kar, M. M. Akhter, A. C. Mishra, S. K. Mohanty, EDMD: An Entropy based dissimilarity measure to cluster mixed-categorical data, *Pattern Recognit.*, **155** (2024), 110674. https://doi.org/10.1016/j.patcog.2024.110674

8. T. Dinh, H. Wong, P. Fournier-Viger, D. Lisik, M. Q. Ha, H. C. Dam, et al., Categorical data clustering: 25 years beyond K-modes, *Expert Syst. Appl.*, **272** (2025), 126608. https://doi.org/10.1016/j.eswa.2025.126608

9. C. Zhu, Q. Zhang, L. Cao, A. Abrahamyan, Mix2Vec: Unsupervised mixed data representation, in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics*, (2020), 118–127. https://doi.org/10.1109/DSAA49011.2020.00024

10. Y. Zhang, M. Zhao, H. Jia, M. Li, Y. Lu, Y. M. Cheung, Categorical data clustering via value order estimated distance metric learning, *Proc. ACM Manag. Data.*, **3** (2025), 1–24. https://doi.org/10.1145/3769772

11. Y. Zhang, Y. M. Cheung, A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering, *IEEE Trans. Cybern.*, **52** (2022), 758–771. https://doi.org/10.1109/TCYB.2020.2983073

12. S. Jian, L. Hu, L. Cao, K. Lu, Metric-based auto-instructor for learning mixed data representation, *Proc. AAAI Conf. Artif. Intell.*, **32** (2018), 3318–3325. https://doi.org/10.1609/aaai.v32i1.11597

13. D. Lin, An information-theoretic definition of similarity, in *Proceedings of the Fifteenth International Conference on Machine Learning*, (1998), 296–304. https://doi.org/10.5555/645527.657297

14. P. Arabie, N. D. Baier, C. F. Critchley, M. Keynes, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 2006.

15. C. Zhang, L. Chen, Y. P. Zhao, Y. Wang, C. L. P. Chen, Graph enhanced fuzzy clustering for categorical data using a Bayesian dissimilarity measure, *IEEE Trans. Fuzzy Syst.*, **31** (2023), 810–824. https://doi.org/10.1109/TFUZZ.2022.3189831

16. Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Min. Knowl. Discov.*, **2** (1998), 283–304. https://doi.org/10.1023/A:1009769707641

17. Y. Qian, F. Li, J. Liang, B. Liu, C. Dang, Space structure and clustering of categorical data, *IEEE Trans. Neural Netw. Learn. Syst.*, **27** (2016), 2047–2059. https://doi.org/10.1109/TNNLS.2015.2451151

18. S. Cai, Y. Zhang, X. Luo, Y. M. Cheung, H. Jia, P. Liu, Robust categorical data clustering guided by multi-granular competitive learning, in *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, (2024), 288–299. https://doi.org/10.1109/ICDCS60910.2024.00035

19. C. Zhu, L. Cao, J. Yin, Unsupervised heterogeneous coupling learning for categorical representation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 533–549. https://doi.org/10.1109/TPAMI.2020.3010953

20. M. Zhao, S. Feng, Y. Zhang, M. Li, Y. Lu, Y. M. Cheung, Learning order forest for qualitative-attribute data clustering, in *ECAI 2024*, (2024), 1943–1950. https://doi.org/10.3233/FAIA240709

21. L. Xie, S. Fan, W. Gao, B. Chen, G. Li, W. Gao, Just noticeable difference measurement for point cloud compression: A benchmark dataset and prediction network, *IEEE Trans. Instrum. Meas.*, **75** (2026), 1–17. https://doi.org/10.1109/TIM.2026.3680192

22. L. Xie, H. Li, B. Chen, G. Li, S. Kwong, W. Gao, Foreground-aware geometry compression with hybrid attention for large-scale point clouds, *IEEE Trans. Broadcast.*, **72** (2026), 207–222. https://doi.org/10.1109/TBC.2026.3651190

23. S. Q. Le, T. B. Ho, An association-based dissimilarity measure for categorical data, *Pattern Recognit. Lett.*, **26** (2005), 2549–2557. https://doi.org/10.1016/j.patrec.2005.06.002

24. D. Ienco, R. G. Pensa, R. Meo, From context to distance: Learning dissimilarity for categorical data clustering, *ACM Trans. Knowl. Discov. Data.*, **6** (2012), 1–25. https://doi.org/10.1145/2133360.2133361

25. S. Jian, L. Cao, K. Lu, H. Gao, Unsupervised coupled metric similarity for non-IID categorical data, *IEEE Trans. Knowl. Data Eng.*, **30** (2018), 1810–1823. https://doi.org/10.1109/TKDE.2018.2808532

26. S. Jian, L. Cao, G. Pang, K. Lu, H. Gao, Embedding-based representation of categorical data by hierarchical value coupling learning, in *IJCAI International Joint Conference on Artificial Intelligence*, (2017), 1937–1943. https://doi.org/10.24963/ijcai.2017/269

27. Y. Zhang, Y. M. Cheung, Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 3560–3576. https://doi.org/10.1109/TPAMI.2021.3056510

28. G. Badaro, M. Saeed, P. Papotti, Transformers for tabular data representation: A survey of models and applications, *Trans. Assoc. Comput. Linguist.*, **11** (2023), 227–249. https://doi.org/10.1162/tacl_a_00544

29. D. Bahri, H. Jiang, Y. Tay, D. Metzler, Scarf: Self-supervised contrastive learning using random feature corruption, preprint, arXiv:2106.15147.

30. Y. Zhang, M. Zhao, Y. Zhang, Y. M. Cheung, Trending applications of large language models: A user perspective survey, *IEEE Trans. Artif. Intell.*, **7** (2026), 1835–1852. https://doi.org/10.1109/TAI.2025.3620272

31. S. Park, S. Han, S. Kim, D. Kim, S. Park, S. Hong, et al., Improving unsupervised image clustering with robust learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 12278–12287. https://doi.org/10.1109/CVPR46437.2021.01210
32. Y. Zhang, M. Zhao, Y. Chen, Y. Lu, Y. M. Cheung, Learning unified distance metric for heterogeneous attribute data clustering, *Expert Syst. Appl.*, **273** (2025), 126738. https://doi.org/10.1016/j.eswa.2025.126738
33. J. Z. Huang, M. K. Ng, H. Rong, Z. Li, Automated variable weighting in k-means type clustering, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 657–668. https://doi.org/10.1109/TPAMI.2005.95
34. G. Gan, M. K. P. Ng, Subspace clustering with automatic feature grouping, *Pattern Recognit.*, **48** (2015), 3703–3713. https://doi.org/10.1016/j.patcog.2015.05.016
35. Y. Zhang, Y. M. Cheung, Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data, *IEEE Trans. Neural Netw. Learn. Syst.*, **34** (2023), 6530–6544. https://doi.org/10.1109/TNNLS.2022.3202700
36. E. Y. Chan, W. K. Ching, M. K. Ng, J. Z. Huang, An optimization algorithm for clustering using weighted dissimilarity measures, *Pattern Recognit.*, **37** (2004), 943–952. https://doi.org/10.1016/j.patcog.2003.11.003
37. M. Zhao, Z. Huang, Y. Lu, M. Li, Y. Zhang, W. Su, et al., Break the tie: Learning cluster-customized category relationships for categorical data clustering, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **40** (2026), 28715–28723. https://doi.org/10.1609/aaai.v40i34.40104
38. L. Chen, S. Wang, K. Wang, J. Zhu, Soft subspace clustering of categorical data with probabilistic distance, *Pattern Recognit.*, **51** (2016), 322–332. https://doi.org/10.1016/j.patcog.2015.09.027
39. L. Bai, X. Cheng, J. Liang, H. Shen, Y. Guo, Fast density clustering strategies based on the k-means algorithm, *Pattern Recognit.*, **71** (2017), 375–386. https://doi.org/10.1016/j.patcog.2017.06.023
40. A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distribution, *Bull. Calcutta Math. Soc.*, **35** (1943), 99–110
41. D. Dua, C. Graff, *UCI Machine Learning Repository*, 2017. Available from: https://archive.ics.uci.edu/ml.
42. J. M. Santos, M. Embrechts, On the use of the adjusted rand index as a metric for evaluating supervised classification, in *International Conference on Artificial Neural Networks*, (2009), 175–184. https://doi.org/10.1007/978-3-642-04277-5_18
43. P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, *IEEE Trans. Neural Netw.*, **20** (2009), 189–201. https://doi.org/10.1109/TNN.2008.2005601