Towards Unbiased Minimal Cluster Analysis of Categorical-and-Numerical Attribute Data

Yunfan Zhang[1], Xiaopeng Luo[2], Qingsheng Chen[1], Rong Zou[3], Yiqun Zhang[1,3(✉)], and Yiu-ming Cheung[3]

[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
{3121008002,2112205080}@mail2.gdut.edu.cn
[2] School of Computer Engineering, Guangzhou Huali College, Guangzhou, China
[3] Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China
{rongzou,ymc}@comp.hkbu.edu.hk, yqzhang@gdut.edu.cn

Abstract. Categorical and numerical attributes occur frequently in cluster analysis tasks. To bridge the information gap between the heterogeneous categorical and numerical attributes in cluster analysis, the existing approaches usually adopt prior assumptions to distance definition and cluster distribution, which unavoidably introduce bias to the clustering process. To address this issue, we propose to analyze mixed data comprising both categorical and numerical attributes by forming minimal clusters through neighborhood set theory. As the minimal clusters are the smallest cluster units that can be obtained without relying on prior assumptions, unbiased cluster analysis can be facilitated accordingly. To avoid information loss, distance and density metrics that are unified on both numerical and categorical attributes are also proposed and utilized to merge the minimal clusters hierarchically. It turns out that our proposed approach is highly interpretable, and is capable of accurately and robustly clustering data sets composed of any combination of numerical and categorical attributes. Extensive experimental evaluations demonstrate its efficacy.

Keywords: Cluster analysis · Categorical attribute · Neighborhood rough set · Mixed data · Unsupervised learning

1 Introduction

Cluster analysis is a common data analytic technique to identify cluster patterns from data sets. In real clustering tasks, numerical attributes with quantitative values and categorical attributes [1] with qualitative values are very common, where we call the data set composed of both numerical and categorical attributes mixed data. However, as shown in Fig. 1, the distance space of mixed data cannot

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15302, pp. 254–269, 2025.
https://doi.org/10.1007/978-3-031-78166-7_17

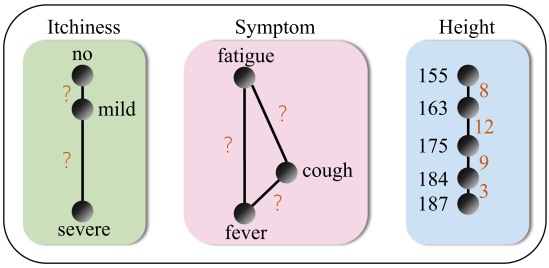


Fig. 1. Numerical attributes such as 'Height' can be effectively represented in Euclidean space, while quantifying dissimilarity between possible values within categorical attributes like 'Itchiness' and 'Symptoms' poses great challenges as the categorical values cannot directly participate in arithmetic operations.

be well-defined like the Euclidean distance due to the qualitative categorical data values. Additionally, the possible values of categorical attributes are usually divergent concepts in different domains with distinct implicit distance structures, which brings great challenges to the cluster analysis of mixed data.

Most existing attempts for mixed data clustering focus on the distance defining across heterogeneous attributes, and can be roughly divided into the following two streams: (i) k-ProtoType (KPT)-type methods: directly weight and combine different dissimilarity measures, e.g., Euclidean and Hamming metrics during clustering, and (ii) dedicated metric-based methods, which usually define a metric unified on the numerical and categorical attributes for distance measurement during clustering.

For KPT-type methods, the conventional KPT algorithm [11] combines Euclidean and Hamming distances [3] to cluster mixed data sets. A recent variant [17] improves the metric of categorical attributes by representing categorical values via inter-value and inter-attribute couplings, thus encoding relationships for better distance measurement. Context-based metric [13] considers attribute interdependence to form an informative categorical attribute metric. More advanced clustering methods like [16] measure the distance between possible values using Conditional Probability Distributions (CPDs) across attributes. However, these methods focus solely on proposing more advanced categorical attribute metrics and combine them with Euclidean distance for mixed data clustering, neglecting the heterogeneity of categorical and numerical attributes.

For metric-based methods, the work proposed in [7] quantifies inter-object-cluster similarity for numerical and categorical attributes within a unified probability framework, while an entropy-based approach [23] further considers the value order in categorical attributes and measures the dissimilarity degrees between different possible values from an information theory perspective. Nevertheless, these methods assume the independence of attributes, leading to information loss when applied to real-world data sets with interdependent attributes. Advanced distance definitions [14,21] take into account attribute interdependence and preserve the corresponding information by reflecting distances based on more relevant attributes. However, they are not robust to various data sets as their effectiveness highly relies on the consistency between inherent data char-

acteristics and their assumptions, e.g., the existence of inter-value order, and inter-attribute dependence, etc.

In general, almost all the existing mixed data clustering approaches rely on certain prior knowledge or assumptions of data sets. Specifically, context-based [13] methods adopt the prior knowledge that the similarities of their possible values are reflected by the CPDs corresponding to the values obtained from other attributes, while the information theory-based methods like [23] measure dissimilarity between two possible values according to the degree of information chaos jointly demonstrated by them. Moreover, the number of true clusters is usually assumed to be known in advance. However, direct searching for oversized clusters may hinder the exploration of locally compact smaller-sized clusters. The above issues will inevitably lead to various clustering biases and thus influence clustering accuracy.

To this end, this paper proposes a universal clustering algorithm robust to various mixed data, addressing the challenges of considering heterogeneous attributes and lifting the restriction of prior knowledge. It groups data objects with distinct boundaries according to neighborhood set search, where only intra-cluster objects are expected to be included in a compact group (also called micro-cluster). Then the micro-clusters are merged to form larger "true" clusters (macro-cluster), and thus the proposed algorithm is called Mic2Mac. More specifically, a novel neighborhood relation is first proposed, forming rational and compact micro-clusters by comprehensively considering the distance and density of data objects. Subsequently, a hierarchical merging mechanism is designed to merge the current most similar micro-clusters into macro-clusters progressively. As the hierarchical merging is performed at the cluster level, the computation cost is thus not obviously increased. Extensive experiments, including comparative results, ablation studies, and visualization, affirm the superiority of Mic2Mac across various clustering methods on real benchmark data sets. The main contributions of this paper are three-fold:

1) A new clustering method is proposed based on neighborhood relationship to accurately form clusters of arbitrary shapes, tackling the cluster distribution bias of existing mixed data clustering methods.
2) An adaptive neighborhood relationship is defined based on both distance and density, leading to the generation of compact and non-overlapping micro-clusters, which has been proven to be universal and practical in the exploration of complex real-world data distributions.
3) Clustering process of Mic2Mac conforms to the inference process from deterministic micro-clusters to uncertain macro-clusters, providing interpretable cluster nesting relationships for multi-granular distribution analysis.

2 Related Work

As our proposed mixed data clustering approach is based on the data object partition technique, this section makes an overview of mixed data measures. It focuses on mixed data clustering methods, and data object partition techniques including k-means-type partition techniques, and neighborhood rough sets.

2.1 Mixed Data Measures

Early mixed data clustering methods like k-prototypes [11], utilized one-hot encoding to transform categorical attribute values [2] into binary vectors. However, the Hamming distance has obvious limitations in discerning differences between various value pairs. Consequently, numerous advanced techniques have emerged to efficiently address the heterogeneous attribute data, including similarity-based and representation-based approaches.

For similarity-based measures, such as context-based distance measures [13], they evaluate the distance between related attribute CPDs to highlight their dissimilarity and identify attributes with weaker dependencies. Nonetheless, these methods do not fully account for the heterogeneity of the complex categorical attributes. Subsequently, the information theory-based metric [23] measures the distances for categorical attributes by incorporating attribute weighting. Most recently, a distance learning-based approach [19] has been proposed to learn the ordinal structure of the qualitative attributes and then cluster them, while AMPHM [24] is proposed to cluster mixed data based on the rough set theory.

For representation-based measures, an interpretable representation method [16] encodes original data and further performs k-means clustering and PCA for more accurate representation. However, it is designed for categorical data only. Recently, a deep learning clustering method [5] transforms both numerical and categorical attribute values into a unified space to enable more appropriate clustering. Most recently, an approach [25] constructs minimal spanning trees for possible values to tackle qualitative-attribute clustering tasks. Moreover, the competitive theory has been utilized to handle the qualitative categorical data [4] and clustering in a federated scenario [26]. Most existing methods for clustering mixed data typically have one or both following restrictions: 1) they are tailored to data sets with one specific attribute type, and 2) they often rely on prior knowledge or assumptions.

2.2 Data Object Partition Techniques

The early k-means-type approach [12] was widely used for partitioning numerical and categorical attributes data into k clusters, while it treats all categorical variables equally during the clustering process. Recently, the representative attribute weighting partition methods w-k-means [10] was proposed for reasonably selecting variables, thereby partitioning mixed data. Nevertheless, it unreasonably assigns identical distances to different pairs of adjacent categories that may have intrinsically unequal distances, thus showing unsatisfactory partition results. Most recently, The clustering approach in [6] is proposed for incomplete data, but designed for numerical data only.

Neighborhood rough set (also called neighborhood set interchangeably for simplicity) is commonly used to partition categorical or mixed data sets. Specifically, it lets each object $\mathbf{x}_i$ find a micro-cluster based on the neighborhood set, consisting of objects that are closer to $\mathbf{x}_i$. $D(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The common neighborhood relations are the k-nearest

$$M^k(\mathbf{x}_i) = \{\mathbf{x}_j | D_k(\mathbf{x}_i, \mathbf{x}_j) < D(\mathbf{x}_i, \mathbf{x}_g)\}, \tag{1}$$

and the δ-radius

$$M^\delta(\mathbf{x}_i) = \{\mathbf{x}_j | D(\mathbf{x}_i, \mathbf{x}_j) \leq \delta\}, \tag{2}$$

where $j, g \in \{1,2,...,n\}$, $g \neq j$, and the k in Eq. (1) represents the first k objects with the closest distance to $\mathbf{x}_i$. For simplification, we employ $M(\mathbf{x}_i)$ to denote the general neighborhood relation.

3 Proposed Method

In this section, we begin by formulating the problem in Sect. 3.1. Then, we present the micro partition based on the neighborhood set and the mixed data distance metric in Sect. 3.2. Finally, the hierarchical merging mechanism, and the whole clustering algorithm Mic2Mac are proposed in Sect. 3.3.

3.1 Problem Formulation

Given a mixed data set $S = \{\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n\}$ comprising n data objects, each data object $\mathbf{x}_i = [x_i^1, x_i^2,... ,x_i^d]^\top$ is a d-dimensional vector with values from the d attributes, which can also be denoted as a set $A = \{a^1, a^2,..., a^d\}$. The possible value set $V = \{V^1, V^2,..., V^d\}$ stores the value domains corresponding to each attribute. The goal of clustering is to assign the n objects to k suitable clusters $C = \{C_1, C_2,..., C_k\}$, where C_l is the set of data objects in the l-th cluster, with $S = \bigcup_{l=1}^{k} C_l$. To represent each cluster, a representative objects set $R = \{\mathbf{r}_1, \mathbf{r}_1,..., \mathbf{r}_k\}$ is maintained during clustering, and each representative object $\mathbf{r}_l$ of R is a data object selected from S. A common way is to use an $n \times k$ matrix $\mathbf{Q}$, indicating which cluster is an object assigned to. The (i, l)-th entry $q_{i,l}$ of $\mathbf{Q}$ is denoted as

$$q_{i,l} = \begin{cases} 1, \text{ if } & l = arg\min\limits_{g} D(\mathbf{x}_i, \mathbf{r}_g), \\ 0, \text{ if } & l \neq g. \end{cases} \tag{3}$$

According to Eq. (3), we have

$$\sum_{l=1}^{k} q_{i,l} = 1, \quad 1 \leq i \leq n, \tag{4}$$

and $q_{i,l} \in \{0, 1\}$. To appropriately cluster mixed data sets, we first need inter-object distances to be prepared where a common form can be

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{a^r \in A} D^r(x_i^r, x_j^r)^2}. \tag{5}$$

In Eq. (5), $x_i^r \in V^r$ represents the value of $\mathbf{x}_i$ on a^r, while $D^r(x_i^r, x_j^r)$ quantifies the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ w.r.t. a^r. In the following subsection, we present how to define $D^r(x_i^r, x_j^r)$ to form neighborhood sets.

3.2 Micro Partition Based on Neighborhood Set

To unify the distance metric on heterogeneous attributes, we use transformation cost that quantifies the effort required to transform one Conditional Probability Distribution (CPD) into another. We begin by defining the CPD and establishing the distance between possible values of a categorical attribute to explain the principles of transformation cost quantification more clearly. Subsequently, we illustrate how this approach unifies both categorical and numerical scenarios. Finally, we derive the object-level distance and propose the micro partition based on the neighborhood set. Given a possible value v_h^r from attribute a^r, the CPD of a^t with v^t possible values $V^t = \{v_1^t, v_2^t, ..., v_{\mathrm{v}^t}^t\}$ is computed accordingly

$$\Psi_h^{rt} = [p(v_1^t|v_h^r), p(v_2^t|v_h^r), ..., p(v_{\mathrm{v}^t}^t|v_h^r)]^\top, \tag{6}$$

where $p(v_o^t|v_h^r)$ is the conditional probability of v_o^t as given v_h^r. We represent the CPD as Ψ_h^{rt} where the superscript rt signifies that this CPD characterizes the h-th possible value of a^r concerning the values of a^t. The distinction between two CPDs, such as Ψ_h^{rt} and Ψ_o^{rt}, captures the dissimilarity between v_h^r and v_o^r, according to the possible values V^t.

To quantitatively measure this dissimilarity between the CPDs describing two possible values of a categorical attribute, we employ the Earth Mover's Distance (EMD) [20], which was designed to calculate the transformation costs between two histogram descriptors. Thus, the dissimilarity between two possible values v_h^r and v_o^r, reflected by a^t can be calculated using EMD by

$$D^{rt}(v_h^r, v_o^r) = \Gamma(\Psi_h^{rt} - \Psi_o^{rt}, \boldsymbol{O}) \cdot \mathbf{I}, \tag{7}$$

where $\Gamma(\cdot, \cdot)$ compares each pair of corresponding bits of two vectors and retains the maximum value, while $\boldsymbol{O}$ and $\mathbf{I}$ represent a v^t-dimensional vector with all values equal to 0 and 1, respectively.

Different attributes a^ts can have varying contributions to the distance $D^{rt}(v_h^r, v_o^r)$ due to variations in inter-attribute dependence. The overall $D^{rt}(v_h^r, v_o^r)$ reflected by its respective weight w^{rt} is computed by

$$D^r(v_h^r, v_o^r) = \sum_{a^t \in A} D^{rt}(v_h^r, v_o^r) \cdot w^{rt}. \tag{8}$$

The Eq. (7) is further extended to quantify the inter-attribute dependence as the weights w^{rt}, which can be expressed as

$$w^{rt} = \frac{\sum_{h=1}^{\mathrm{v}^r - 1} \sum_{o=h+1}^{\mathrm{v}^r} D^r(v_h^r, v_o^r)}{\mathrm{v}^r(\mathrm{v}^r - 1)/2}, \tag{9}$$

where v^r represents the number of possible values contained within a^r. More specifically, w^{rt} measures the average transformation cost of the $\mathrm{v}^r(\mathrm{v}^r - 1)/2$ pairs of possible values of attribute a^r reflected by a^t. According to Eqs. (7)–(9), the heterogeneous attributes are uniformly quantified as the transformation cost.

According to the work proposed in [22], the possible values of a categorical attribute are considered as concepts, so that the above process essentially quantifies the average inter-concept distances of a^r as influenced by a^t. To illustrate the principle of Eq. (8), we examine an extreme scenario. Assuming attributes a^r and a^t are identical, they will exhibit perfect interdependence, and thus their $D^{rt}(v_h^r, v_o^r)$ always reaches the maximal value, i.e., "1", for any combinations of h and o with $h \neq o$, according to Eq. (7). Consequently, w^{rt} also reaches the maximal value of "1", representing 100% dependence of two attributes. By applying Eqs. (7)–(9), we can obtain the distance between data objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

The defined dissimilarity measure applies to both categorical attributes and numerical attributes, as Eq. (8) provides a uniform treatment of heterogeneous attributes. Then we prove that our measure is a distance metric.

Theorem 1. *$D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a distance metric.*

Proof. As Eq. (7) satisfies the properties of a metric, it follows naturally Eq. (8), which is derived from Eq. (7), is also a metric. Moreover, the calculation of Eq. (5) involves finite arithmetic processes according to Eq. (8), guaranteeing that $D(\mathbf{x}_i, \mathbf{x}_j)$ adheres to all essential metric properties for any $i, j, h \in \{1, 2, ..., n\}$, which are listed as follows:

(1) $D(\mathbf{x}_i, \mathbf{x}_j) \geq 0$; $D(\mathbf{x}_i, \mathbf{x}_j) = 0$ iff $\mathbf{x}_i = \mathbf{x}_j$;
(2) $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i)$;
(3) $D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_h) + D(\mathbf{x}_h, \mathbf{x}_j)$. □

The conventional neighborhood sets $M^k(\mathbf{x}_i)$ and $M^\delta(\mathbf{x}_i)$ (i.e., Eqs. (1) and (2)) generate n neighborhood sets, which may partially overlap with surrounded ones, causing laborious computation with a large n. Additionally, these neighborhood relations may group dissimilar objects in the uneven distribution of data objects. To better partition objects and reduce computational costs, we have developed a new approach called micro partition based on the neighborhood set, which considers both distance and density. This approach creates non-overlapping neighborhood sets by selecting representative objects and grouping their corresponding neighbors based on *merging interval*.

Definition 1. *Merging interval: Given an object $\boldsymbol{x}_i$ with a density ρ_i, the merging interval ϕ_i signifies the minimum distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with a higher corresponding density ρ_j, which can be expressed as:*

$$\phi_i = \min D(\mathbf{x}_i, \mathbf{x}_j),\ s.t. \rho_i < \rho_j \text{ and } \mathbf{x}_j \in S \backslash \mathbf{x}_i, \tag{10}$$

where $S \backslash \mathbf{x}_i$ is the data set that excludes $\mathbf{x}_i$, while ρ_i and ρ_j denote the densities of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Furthermore, for the object with the maximum density, its merging interval is defined as $\max D(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_j \in S$.

In Definition 1, the density ρ_i can be computed as

$$\rho_i = \frac{D(\mathbf{x}_i, \mathbf{x}_{\langle i, q_i \rangle})}{q_i}. \tag{11}$$

Equation (11) computes the distance corresponding to the ranking q_i of the adjacent object $\mathbf{x}_{\langle i,q_i\rangle}$, which can be seen as the density of $\mathbf{x}_i$. Assuming the $\mathbf{x}_i$ is a center point, $\mathbf{x}_{\langle i,q_i\rangle}$ is the q_i-th closest object to $\mathbf{x}_i$ in n objects. Specifically, we initially create the neighbor set $AM_i = \{\mathbf{x}_{\langle i,0\rangle}, \mathbf{x}_{\langle i,1\rangle}, \mathbf{x}_{\langle i,2\rangle}, ..., \mathbf{x}_{\langle i,n-1\rangle}\}$ in *ascending* sequence relative to $\mathbf{x}_i$, where $\mathbf{x}_{\langle i,0\rangle} \equiv \mathbf{x}_i$ and $AM_i(y) = \mathbf{x}_{\langle i,y\rangle}$. Afterwards, when we iterate through AM_i from small to large, we choose the object $\mathbf{x}_{\langle i,g\rangle}$ that first satisfies the condition $D(\mathbf{x}_i, \mathbf{x}_{\langle i,g\rangle})/g < D(\mathbf{x}_{\langle i,g-1\rangle})/(g-1)$, which confirms the value of q_i as $q_i = g - 2$. The density calculation effectively selects neighboring objects, ensuring that objects beyond a noticeable interval boundary are not included in the neighborhood set corresponding to $\mathbf{x}_i$. Hence, it will partition objects into compact clusters, which contain the most similar objects.

To select the most suitable representative object for a micro-cluster, we prioritize objects with higher density than their neighbors and positioning far from other representative objects. According to Definition 1, objects with greater merging intervals are considered more suitable to be the representative objects. Thus, we rank data objects based on their merging intervals in *descending* sequence and form micro-clusters based on neighborhood set by

$$M^{\phi}(\mathbf{x}_i) = \left\{ \bigcup_{j=1}^{q_i} AM_i(j) \right\} \setminus \left\{ \bigcup_{\phi_p > \phi_i} M^{\phi}(\mathbf{x}_p) \right\}, \tag{12}$$

where q_i is the q_i-th closeness to $\mathbf{x}_i$ among all the n objects, as mentioned in Eq. (11), while the objects $\mathbf{x}_p$ with larger merging intervals than the $\mathbf{x}_i$ will be excluded from $M^{\phi}(\mathbf{x}_i)$ corresponding to $\mathbf{x}_i$. The process of forming micro-clusters will continue until all objects are contained by these micro-clusters. All the representative objects in each micro-cluster are stored in the micro representative objects set $MR = \{\mathbf{b}_1, \mathbf{b}_1, ..., \mathbf{b}_m\}$, where m is the number of representative objects.

To illustrate our calculation and merging processes more clearly, Fig. 2 provides a toy example shown in processes 1–4. The proposed micro partition based on the neighborhood set is outlined in Algorithm 1. The mechanism for merging $M^{\phi}(\mathbf{x}_i)$ is crucial and will be discussed in the next subsection.

3.3 Merge Micro-Clusters Into Macro-Clusters

Based on our proposed micro partition, a hierarchical merging mechanism is presented to merge micro-clusters.

Given the data set S and the number of clusters k, we iteratively compute micro-clusters $M^{\phi}(\mathbf{x}_i)$ and update data set S at each layer in the following two steps: 1) fix S, compute $M^{\phi}(\mathbf{x}_i)$ and micro representative object set MR by Algorithm 1 according to dissimilarity matrices D, and 2) fix MR, update S based on MR. Specifically, the hierarchical merging mechanism utilizes MR from the previous layer as the new local data set in the next layer. This process enables multiple partitioning and merging of objects while preserving the local micro-clusters. These two steps iterate until $m = k$, where m is the number of

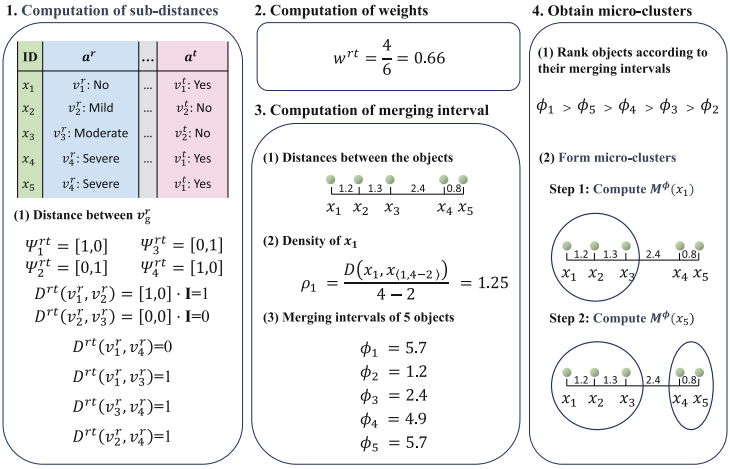


Fig. 2. A toy example illustrates the calculation processes. In processes 1 and 2, we compute the sub-distances in attribute a^r. Then, we compute the contribution of a^t to a^r. In processes 3 and 4, we confirm q_i, where $AM_1 = \{\mathbf{x}_{\langle 1,0 \rangle}, \mathbf{x}_{\langle 1,1 \rangle}, \mathbf{x}_{\langle 1,2 \rangle}, \mathbf{x}_{\langle 1,3 \rangle}, \mathbf{x}_{\langle 1,4 \rangle}\}$. After obtaining the merging intervals corresponding to each object, we then merge data objects into micro-clusters according to the descending order of the merging intervals.

Algorithm 1. MPNS: Micro Partition based on Neighborhood Set

Input: S, D.
Output: $M^{\phi}(\mathbf{x}_i)$, MR.

```
for i = 1 to n do
    Update the density ρ_i of x_i based on Eq. (11);
end for
for i = 1 to n do
    Update the merging interval φ_i of x_i based on Eq. (10);
end for
for i = 1 to n do
    if ρ_i > 0 then
        Select x_i as representative object b_i to MR;
    end if
    Update M^φ(x_i) based on Eq. (12);
end for
```

representative objects. The overall Mic2Mac clustering algorithm is outlined in Algorithm 2.

Theorem 2. *The time complexity of* Mic2Mac *is* $O(d^2n + n\log n)$ *for each iteration.*

Proof. In the worst-case scenario, all attributes are categorical, and V is equal to the maximum number of possible values across all the categorical attributes.

Algorithm 2. Mic2Mac: Merge Micro-Clusters into Macro-Clusters

Input: S, k.
Output: $\mathbf{Q}$.

1: Initialize the iteration counter by $\tau = 0$; Set each object as a micro-cluster;
2: **while** $|MR^{(\tau)}| > k$ **do**
3: Update D based on Eq. (5);
4: Update $M^{\phi,(\tau)}(\mathbf{x}_i)$ and $MR^{(\tau)}$ by Algorithm 1;
5: Update $S^{(\tau+1)}$ by $S^{(\tau+1)} = MR^{(\tau)}$;
6: Update the iteration counter by $\tau = \tau + 1$;
7: **end while**
8: Update $R = MR^{(\tau)}$;
9: Compute $\mathbf{Q}^{(\tau)}$ according to Eq. (3).

To analyze the overall complexity, we compute the complexity of $D^{(\tau)}$, $M^{\phi,(\tau)}$, and hierarchical merging once, respectively.

To compute the dissimilarity matrices $D^{(\tau)}$, we need to derive $d \times d$ pairs of CPDs by scanning n data objects in data set S. This results in a $O(d^2 n)$ complexity. For computing the distances between a pair of intra-attribute possible values, it takes $O(V)$ complexity for every attribute. Thus, obtaining $D^{(\tau)}$ incurs a complexity of $O(nd^2 + V)$.

Given $D^{(\tau)}$ obtained from Algorithm 2, to compute $M^{\phi,(\tau)}$, we need to sort an $n \times n$ matrix, taking $O(n + n \log n)$ complexity. Subsequently, we sort n merging intervals in $O(n \log n)$ complexity. Therefore, computing $M^{\phi,(\tau)}$ takes $O(n + 2n \log n)$.

To implement hierarchical merging, we need to update S in each iteration according to the micro representative objects set MR, which takes $O(n)$.

Therefore, the overall complexity of Mic2Mac at a given iteration τ can be simplified to $O(d^2 n + n \log n)$. □

4 Experiments

4.1 Experimental Settings

This section presents three types of experiments to comprehensively evaluate the clustering performance of our proposed Mic2Mac: (1) Clustering performance evaluation, (2) Ablation study, and (3) Visualization of cluster discrimination ability. Counterparts, validity indices, and data sets are introduced below.

Ten counterparts are compared, including Jia's Distance Metric (JDM) [14], Coupled Similarity Metric (CSM) [15], Entropy-based Distance Metric (EDM) [23], and Zhang's Distance Metric (ZDM) [21] incorporated with the conventional k-modes (KMD) [12] and k-prototypes (KPT) [11] approaches based on the attribute composition of data sets. Cheung's Iterative Learning (CIL) [7], designed for data sets with numerical and categorical attributes, is also selected. JDM, CSM, EDM, and ZDM represent state-of-the-art methods. Additionally, three conventional clustering algorithms, namely Attribute Weighting k-means

Table 1. Summary of nine utilized data sets. The columns "Categorical", "Numerical", "Objects", and "Clusters" are the numbers of categorical attributes, numerical attributes, data objects, and clusters, respectively.

No.	Data Set	Abbrev.	Categorical	Numerical	Objects	Clusters
1	Dermatology	Derm	33	1	366	6
2	Autism-Adolescent	Autism	2	7	104	2
3	Common Toad	Toad	12	2	189	2
4	Hayes-Roth	Hayes	4	0	132	3
5	Breast Cancer	Cancer	9	0	286	2
6	Lymphography	Lym	18	0	148	4
7	Congressional Voting	Vote	16	0	435	2
8	Employee Selection	Employee	4	0	488	9
9	Social Workers	Workers	10	0	1000	4

Table 2. Clustering performance evaluated by CA, where the best results are highlighted in **bold** and the second-best results are underlined.

Methods	Derm	Autism	Toad	Hayes	Cancer	Lym	Vote	Employee	Workers
KMD	0.554±0.10	0.545±0.11	0.548±0.03	0.364±0.01	0.519±0.02	0.453±0.04	0.864±0.00	0.367±0.03	0.392±0.03
KPT	0.554±0.10	0.530±0.03	0.530±0.02	0.364±0.01	0.519±0.02	0.453±0.04	0.864±0.00	0.367±0.03	0.392±0.03
WKM	0.623±0.09	0.525±0.02	0.523±0.03	0.408±0.05	0.584±0.09	0.439±0.05	0.857±0.07	0.368±0.03	0.375±0.03
CIL	0.675±0.10	0.519±0.03	0.506±0.00	0.376±0.04	0.541±0.06	0.500±0.04	**0.881±0.00**	0.384±0.04	0.373±0.03
JDM	0.665±0.10	0.579±0.05	0.522±0.02	0.375±0.02	0.582±0.10	0.473±0.04	0.868±0.00	0.351±0.03	0.334±0.03
CSM	0.602±0.14	0.524±0.03	0.526±0.02	0.405±0.04	0.528±0.04	0.419±0.05	0.865±0.01	**0.402±0.04**	0.331±0.03
EDM	0.587±0.10	0.558±0.03	0.537±0.03	0.407±0.03	0.530±0.02	0.452±0.04	0.832±0.10	0.366±0.02	0.332±0.01
ZDM	0.685±0.11	0.558±0.02	**0.578±0.02**	0.404±0.03	0.569±0.19	0.470±0.04	0.872±0.00	0.368±0.03	0.374±0.03
Mic2Mac	**0.768±0.00**	**0.596±0.00**	0.545±0.00	**0.417±0.00**	**0.766±0.00**	**0.561±0.00**	0.874±0.00	0.393±0.00	**0.435±0.00**

(WKM) clustering algorithm [10], the original KMD, and KPT adopting Hamming and Euclidean distance metrics, are also included in the comparison. Furthermore, two variations of Mic2Mac, named Mic2-Mac$^{\mathrm{I}}$ and Mic2Mac$^{\mathrm{II}}$, are introduced for ablation studies, and additional details about these two Mic2Mac variants are provided in Sect. 4.3.

Two validity indices have been chosen for comprehensively verifying the clustering performance, including CA [9] with a value range of [0, 1], and ARI [8] with a value range of [-1, 1]. A higher value for both these indices indicates better clustering performance.

Nine real-world data sets from various domains, including medicine, biology, sociology, etc., have been selected, which are shown in Table 1. Data sets 1–7 are public data sets collected from the UCI machine learning library[1]. Data sets 8 and 9 are obtained from the Weka website[2]. All data sets are pre-processed by removing objects with missing values.

[1] https://archive.ics.uci.edu/
[2] https://waikato.github.io/weka-wiki/datasets/

Table 3. Clustering performance evaluated by ARI, where the best results are highlighted in **bold** and the second-best results are underlined.

Methods	Derm	Autism	Toad	Hayes	Cancer	Lym	Vote	Employee	Workers
KMD	0.396±0.15	−0.003±0.01	−0.002±0.02	−0.012±0.00	−0.004±0.00	0.113±0.04	0.530±0.00	0.162±0.02	0.057±0.02
KPT	0.422±0.12	−0.003±0.01	−0.008±0.01	−0.012±0.00	−0.004±0.00	0.113±0.04	0.530±0.00	0.162±0.02	0.057±0.02
WKM	0.509±0.09	−0.006±0.01	−0.008±0.01	0.007±0.02	0.040±0.07	0.085±0.04	0.527±0.11	0.172±0.03	0.046±0.02
CIL	0.606±0.10	−0.007±0.01	−0.021±0.00	−0.004±0.02	0.011±0.04	**0.182±0.05**	**0.579±0.00**	0.193±0.02	0.052±0.02
JDM	0.614±0.13	0.018±0.03	−0.014±0.01	−0.006±0.01	0.041±0.07	0.123±0.04	0.541±0.01	0.167±0.02	0.052±0.01
CSM	0.518±0.17	−0.009±0.01	−0.008±0.01	0.008±0.02	0.003±0.02	0.089±0.04	0.532±0.03	**0.212±0.03**	0.051±0.02
EDM	0.439±0.12	0.006±0.01	0.002±0.01	0.008±0.02	0.007±0.01	0.089±0.03	0.478±0.17	0.163±0.04	0.059±0.01
ZDM	0.627±0.15	−0.015±0.01	**0.013±0.02**	0.007±0.02	0.062±0.02	0.132±0.05	0.553±0.01	0.211±0.02	0.076±0.01
Mic2Mac	**0.678±0.00**	**0.019±0.00**	−0.001±0.00	**0.009±0.00**	**0.109±0.00**	0.129±0.00	0.557±0.00	0.173±0.00	**0.085±0.00**

Table 4. Ave. Rank of CA and ARI rows report the average performance ranks, where the best results are highlighted in **bold**, while the second-best results are underline.

Ave. Rank	KMD	KPT	WKM	CIL	JDM	CSM	EDM	ZDM	Mic2Mac
Ave. Rank @ CA	5.944	6.389	5.278	4.889	5.222	6.111	6.167	3.556	**1.444**
Ave. Rank @ ARI	6.611	6.722	6.389	4.722	4.833	5.556	5.222	3.056	**1.889**

4.2 Clustering Performance Evaluation

The clustering performance is reported in Tables 2 and 3, which are accessed by CA and ARI, respectively. The average ranks of the CA and ARI performances across all data sets for the compared methods are presented in Table 4, based on the results in Tables 2 and 3.

The key observations are as follows: (1) Mic2Mac consistently performs the best on most data sets in terms of CA index. (2) On certain data sets, such as Toad, Vote, and Employee, Mic2Mac does not achieve the best result, but the performance gaps between Mic2Mac and the best-performing counterparts are tiny, also highlighting the superiority of Mic2Mac. (3) While Mic2Mac does not yield the best results in terms of the ARI on some data sets, e.g., Lym and Employee, it consistently performs the best and the second-best on most data sets, which still verifies its effectiveness. Intuitively, if a data set contains only numerical attributes, the performance of Mic2Mac downgrades to traditional k-means. The more categorical attributes a data set contains, the better the Mic2Mac can perform. Meanwhile, Mic2Mac also performs well on mixed data.

4.3 Ablation Study

In ablation studies, we focus on the clustering performance assessed by the ARI. Firstly, to assess the effectiveness of the dissimilarity metric proposed for heterogeneous attributes, we restrict Mic2Mac to utilize the combination of Hamming distance and Euclidean distance to tackle mixed data, forming Mic2Mac$^{\text{I}}$. Secondly, to evaluate the effectiveness of our proposed hierarchical merging mechanism, we compare Mic2Mac and Mic2Mac$^{\text{I}}$ with their variation Mic2Mac$^{\text{II}}$, which incorporates the partitioning strategy of KPT by partitioning the representative

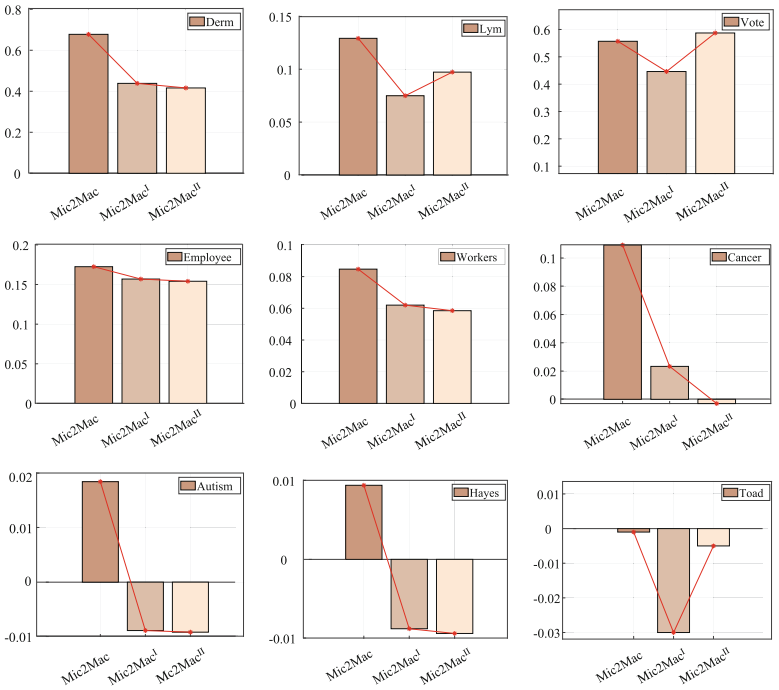
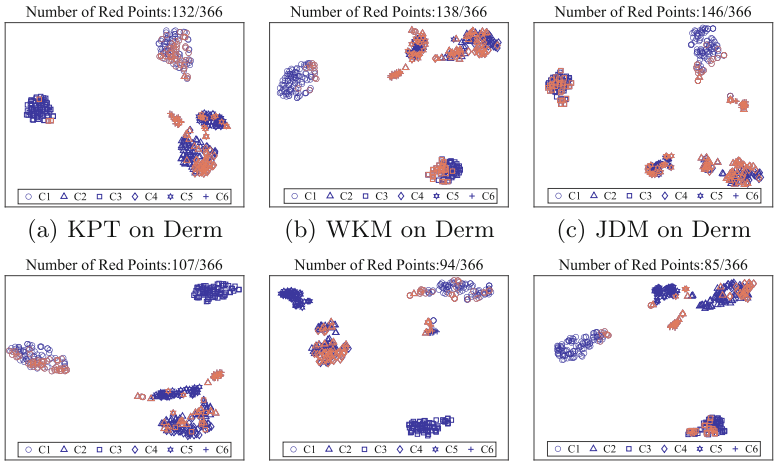


Fig. 3. Comparison of clustering performance among Mic2Mac, Mic2Mac$^{\text{I}}$, and Mic2Mac$^{\text{II}}$ on all the 9 data sets. A better measure has a higher value. The Ave. Rank @ ARI of Mic2Mac, Mic2Mac$^{\text{I}}$, and Mic2Mac$^{\text{II}}$ are 1.111, 2.333, and 2.556, respectively.

objects after the first formation of the micro-clusters. The clustering performance and the average rank of Mic2Mac with its two variations are illustrated in Fig. 3.

The overall result reveals that Mic2Mac consistently outperforms its two variations, demonstrating the effectiveness of Mic2Mac. Specifically, Mic2Mac surpasses Mic2Mac$^{\text{I}}$ on nine data sets, indicating that Mic2Mac can effectively measure the original heterogeneous attribute data information. Furthermore, Mic2Mac outperforms Mic2Mac$^{\text{II}}$ on eight data sets, and Mic2Mac$^{\text{I}}$ performs better than Mic2Mac$^{\text{II}}$ on six data sets. This emphasizes the effectiveness of the proposed merging mechanism. The reason why Mic2Mac$^{\text{I}}$ perform worse than Mic2Mac$^{\text{II}}$ on certain data sets (i.e. Toad, Lym, and Vote) would be that Mic2Mac$^{\text{I}}$ employs the simplest Euclidean and Hamming distance measures, which makes it hard to handle the complex issues in real-world data distribution, e.g., overlapping, and coupling categorical attributes.



(a) KPT on Derm (b) WKM on Derm (c) JDM on Derm

(d) CSM on Derm (e) ZDM on Derm (f) Mic2Mac on Derm

Fig. 4. t-SNE visualization of the Derm data sets, where data points marked in "C1" to "C6" indicate their "true" cluster labels, while objects marked in red indicate they were incorrectly clustered.

4.4 Visualization

In Fig. 4, t-SNE [18] is employed to showcase the cluster discrimination ability of Mic2Mac. The Derm data set is first clustered using KPT, WKM, JDM, CSM, ZDM, and Mic2Mac. Subsequently, the data is encoded according to the distance matrix of objects created by the distance metrics of the corresponding approaches, respectively. These distance matrices are treated as the representations of the data and are then projected into two-dimensional space using t-SNE for visualization. Data points are marked with different markers to indicate their "true" cluster labels. The red markers are utilized to indicate the objects that are incorrectly clustered. Intuitively, fewer red markers indicate a more accurate clustering performance and a more separable distribution of different markers indicates a more powerful cluster discrimination ability.

The visualization in Fig. 4 clearly shows that Mic2Mac exhibits fewer red markers and a more separable distribution of different markers, signifying its stronger cluster discrimination ability than the conventional and state-of-the-art methods.

5 Concluding Remarks

In this paper, a novel approach called Mic2Mac has been proposed for mixed data clustering, which simultaneously tackles two challenges inherent in clustering real-world mixed data sets, i.e., the information gap of heterogeneous attributes

and the bias brought by prior knowledge. To address these challenges, we have proposed: (1) A heterogeneous attribute metric for preserving and leveraging original data information; (2) A micro partition approach based on neighborhood set theory for forming unbiased micro-clusters; and (3) A merging mechanism for hierarchically merging micro-clusters into sought number of clusters. The superiority of Mic2Mac is evidenced through extensive experiments. Moreover, the clustering process of Mic2Mac is highly interpretable due to the nested relationship among multi-granular clusters extracted during the merging phase.

In the future, this research will be extended to address more challenging clustering analysis tasks, e.g., federated mixed data clustering, and exploring cluster patterns for unstructured multi-modal data. Moreover, the potential of the dendrogram formed by merging the micro-clusters will also be explored for understanding complex data sets.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62476063 and 62102097, the NSFC/Research Grants Council (RGC) Joint Research Scheme under the grant N_HKBU214/21, the Natural Science Foundation of Guangdong Province under grant 2023A15150-12855, the General Research Fund of RGC under grants: 12201321, 12202622, and 12201323, and the RGC Senior Research Fellow Scheme under grant SRFS2324-2S02.

References

1. Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley (2002)
2. Alamuri, M., Surampudi, B.R., Negi, A.: A survey of distance/similarity measures for categorical data. In: The 24th International Joint Conference on Neural Networks, pp. 1907–1914. IEEE (2014)
3. Arabie, P., Baier, N.D., Critchley, C.F., Keynes, M.: Studies in classification, data analysis, and knowledge organization. Springer (2006)
4. Cai, S., Zhang, Y., Luo, X., Cheung, Y.m., Jia, H., Liu, P.: Robust categorical data clustering guided by multi-granular competitive learning. In: The IEEE 44th International Conference on Distributed Computing Systems, pp. 288–299 (2024)
5. Chen, J., Ji, Y., Zou, R., Zhang, Y., Cheung, Y.m.: QGRL: Quaternion graph representation learning for heterogeneous feature data clustering. In: The 30th SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1–10 (2024)
6. Cheng, M., You, X.: Leachable component clustering. In: The 26th International Conference on Pattern Recognition, pp. 1858–1864 (2022)
7. Cheung, Y.m., Jia, H.: Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recogn. **46**(8), 2228–2238 (2013)
8. Gates, A.J., Ahn, Y.Y.: The impact of random models on clustering similarity. J. Mach. Learn. Res. **18**, 3049–3076 (2017)
9. He, X., Cai, D., Niyogi, P.: Laplacian Score for Feature Selection. In: The 17th Advances in Neural Information Processing Systems, pp. 507–514 (2005)
10. Huang, J., Ng, M., Rong, H., Li, Z.: Automated variable weighting in k-means type clustering. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 657–668 (2005)

11. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: The 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 21–34 (1997)
12. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. **2**(3), 283–304 (1998)
13. Ienco, D., Pensa, R.G., Meo, R.: From context to distance. ACM Trans. Knowl. Discov. Data **6**(1), 1–25 (2012)
14. Jia, H., Cheung, Y.m., Liu, J.: A new distance metric for unsupervised learning of categorical data. IEEE Trans. Neural Networks Learn. Syst. **27**(5), 1065–1079 (2016)
15. Jian, S., Cao, L., Lu, K., Gao, H.: Unsupervised coupled metric similarity for Non-IID categorical data. IEEE Trans. Knowl. Data Eng. **30**(9), 1810–1823 (2018)
16. Jian, S., Pang, G., Cao, L., Lu, K., Gao, H.: CURE: flexible categorical data representation by hierarchical coupling learning. IEEE Trans. Knowl. Data Eng. **31**(5), 853–866 (2019)
17. Qian, Y., Li, F., Liang, J., Liu, B., Dang, C.: Space structure and clustering of categorical data. IEEE Trans. Neural Networks Learn. Syst. **27**(10), 2047–2059 (2016)
18. Van Der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)
19. Wang, P., Zhang, Y., Zhang, Y., Lu, Y., Li, M., Cheung, Y.m.: Clustering by learning the ordinal relationships of qualitative attribute values. In: The 34th International Joint Conference on Neural Networks, pp. 1–8 (2024)
20. Xu, J., Lei, B., Gu, Y., Winslett, M., Yu, G., Zhang, Z.: Efficient similarity join based on earth mover's distance using MapReduce. IEEE Trans. Knowl. Data Eng. **27**(8), 2148–2162 (2015)
21. Zhang, Y., Cheung, Y.M.: A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. IEEE Trans. Cybern. **52**(2), 758–771 (2022)
22. Zhang, Y., Cheung, Y.M.: Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data. IEEE Trans. Neural Networks Learn. Syst. **34**(9), 6530–6544 (2023)
23. Zhang, Y., Cheung, Y.M., Tan, K.C.: A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering. IEEE Trans. Neural Networks Learn. Syst. **31**(1), 39–52 (2020)
24. Zhang, Y., Zou, R., Zhang, Y., Zhang, Y., Cheung, Y.M., Li, K.: Adaptive micro partition and hierarchical merging for accurate mixed data clustering. Complex Intell. Syst., 1–13 (2024)
25. Zhao, M., Feng, S., Zhang, Y., Li, M., Lu, Y., Cheung, Y.M.: Learning order forest for qualitative-attribute data clustering. In: The 27th European Conference on Artificial Intelligence, pp. 1–8 (2024)
26. Zou, R., Zhang, Y., Zhang, Y., Lu, Y., Li, M., Cheung, Y.M.: Federated clustering with unknown number of clusters. In: The 6th International Conference on Data-driven Optimization of Complex Systems. pp. 1–6 (2024)