Complex & Intelligent Systems (2025) 11:84
https://doi.org/10.1007/s40747-024-01695-7

ORIGINAL ARTICLE

Adaptive micro partition and hierarchical merging for accurate mixed data clustering

Yunfan Zhang[1] · Rong Zou[2] · Yiqun Zhang[1,2] · Yue Zhang[3] · Yiu-ming Cheung[2] · Kangshun Li[4]

Received: 5 November 2023 / Accepted: 23 November 2024 / Published online: 19 December 2024
© The Author(s) 2024

Abstract

Heterogeneous attribute data (also called mixed data), characterized by attributes with numerical and categorical values, occur frequently across various scenarios. Since the annotation cost is high, clustering has emerged as a favorable technique for analyzing unlabeled mixed data. To address the complex real-world clustering task, this paper proposes a new clustering method called Adaptive Micro Partition and Hierarchical Merging (AMPHM) based on neighborhood rough set theory and a novel hierarchical merging mechanism. Specifically, we present a distance metric unified on numerical and categorical attributes to leverage neighborhood rough sets in partitioning data objects into fine-grained compact clusters. Then, we gradually merge the current most similar clusters to avoid incorporating dissimilar objects into a similar cluster. It turns out that the proposed approach breaks through the clustering performance bottleneck brought by the pre-set number of sought clusters k and cluster distribution bias, and is thus capable of clustering datasets comprising various combinations of numerical and categorical attributes. Extensive experimental evaluations comparing the proposed AMPHM with state-of-the-art counterparts on various datasets demonstrate its superiority.

Keywords Cluster analysis · Heterogeneous attributes · Neighborhood rough set · Unsupervised learning

Introduction

Cluster analysis is a frequently employed real-world data analytics technique [1]. The primary aim of cluster analysis is to discover inherent patterns or structures within the data. In real scenarios, datasets often include a variety of attributes that may have diverse value types, such as numerical attributes and categorical attributes [2]. Unlike numerical attributes, which have well-defined distances in Euclidean space, categorical attributes cannot undergo arithmetic computations and lack a well-defined similarity space due to the absence of explicit numerical meanings and the presence of divergent concepts among different possible attribute values.

Table 1 demonstrates two fragments of such real datasets with all two types of attributes. The upper fragment is from a medical dataset, containing the categorical attributes "Scalp",[1] "Family",[2] numerical attribute "Age", and class attribute "Diagnosis". The lower part is a fragment of the Contraceptive Prevalence Survey, which includes categori-

Yunfan Zhang and Rong Zou are co-first authors.

✉ Yiqun Zhang
yqzhang@gdut.edu.cn

Yunfan Zhang
3121008002@mail2.gdut.edu.cn

Rong Zou
rongzou@comp.hkbu.edu.hk

Yue Zhang
zhangyue@gpnu.edu.cn

Yiu-ming Cheung
ymc@comp.hkbu.edu.hk

Kangshun Li
likangshun@scau.edu.cn

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

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

3 School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

4 College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China

[1] "Scalp" refers to the scalp involvement in the dataset.

[2] "Family" refers to the family history in the dataset.

Table 1 Fragments of mixed datasets "dermatology" (upper part) and "contraceptive method choice" (lower part)

ID	Scalp[1] (categorical)	Family[2] (categorical)	Age (numerical)	...	Diagnosis
$Patient_1$	None	No	55	...	seboreic dermatitis
$Patient_2$	Mild	Yes	8	...	psoriasis
$Patient_3$	None	No	26	...	lichen planus
$Patient_4$	Moderate	No	40	...	psoriasis
ID	Education (categorical)	Exposure[3] (categorical)	Children (numerical)	...	Contraceptive used
$Family_1$	Basic	Good	3	...	No-use
$Family_2$	Low	Not good	10	...	No-use
$Family_3$	High	Not good	2	...	Long-term
$Family_4$	Intermediate	Good	1	...	Short-term

cal attributes "Education", "Exposure",[3] numerical attribute "Children", and class attribute "Contraceptive used". Such mixed datasets are prevalent in real-world unsupervised learning tasks, while the heterogeneity of attributes introduces complexities into cluster analysis. Specifically, when conventional Hamming distance is employed and only yields "0" and "1" distances, the original dissimilarity degrees among categorical attribute values are thus wasted.

To cope with the mixed data cluster problem, most existing clustering methods can be roughly divided into two types: (i) approaches that directly weight and fuse different dissimilarity measures, e.g., Euclidean [3] and Hamming distances [4], and (ii) approaches that first define a metric for heterogeneous attributes and then perform based on it.

For the former type [5, 6], numerical and categorical attributes are typically measured using the Euclidean and Hamming distances, respectively, for combination. To enhance the effectiveness of combining metrics, metric-based approaches have been designed in recent years, which measure the similarity between categorical attribute values [7] based on the inter-object dissimilarities present in the dataset as a solution for capturing the relationships between values and attributes in categorical data. Recently, some metrics [8, 9] are introduced to measure the similarity of possible values as the Conditional Probability Distributions (CPDs) of their respective values across disparate attributes. However, all the aforementioned approaches are proposed for nominal data only, which is a subtype of categorical data.

For the latter type that defines metrics for heterogeneous attributes, some measures and metrics are defined in a unified way in the literature. The similarity metric [20] quantifies the similarity of data objects within a unified probability framework, addressing both numerical and categorical attributes, but ignores to further measuring the dissimilarity of divergent concepts among different possible attribute values. Lin's distance metric [10] and the entropy-based similarity metric [16] consider the value orders of categorical attributes and measure the dissimilarity between potential values from an information theory standpoint. However, these dissimilarity-based clustering methods often fail to account for the varying degrees of dissimilarity among heterogeneous attribute values, which may result in grouping less similar data objects or splitting similar ones. For defining similarities more finely, the state-of-the-art methods may still encounter challenges when it comes to clustering various heterogeneous attribute data. For example, context-based distance metrics [13–15] quantify interdependence among categorical attributes, select relevant attributes, and use CPDs to indicate similarity between potential attribute values. Recently, there has been an exploration of the interdependence between categorical attributes through the development of a unified distance metric [18]. Most recently, more advanced solutions, including Mix2Vec [21], Het2Hom [22] and homogeneous distance metric [19] have been proposed. The first two approaches effectively preserve the intrinsic structural information of heterogeneous attribute values. However, their effectiveness heavily depends on the chosen encoding strategies for categorical attributes, e.g., Het2Hom projects all attribute values into spaces created by different pairs of concepts related to that attribute. In contrast, the third approach [19] uniformly establishes distance metric to measure nominal [23] and ordinal categorical attributes but is not suitable for datasets containing typical numerical attributes.

In general, all the approaches mentioned for clustering mixed data also face a fundamental challenge, i.e., the limited application caused by the reliance on prior knowledge or assumptions. Even the most appropriately defined metrics for handling heterogeneous attributes rely on certain assumptions or models, such as probability, entropy, and so on. These assumptions and models are based on prior knowledge, which may not always be well-suited for the complexity of real datasets and clustering tasks. Table 2 provides a sum-

[3] "Exposure" refers to the media exposure in the dataset.

Table 2 The significant issues considered by existing clustering methods

Methods	Heterogeneity	Intra-attribute	Intter-attribute	Importance	Unification
Euclidean + hamming distance [5]	√	√			
Lin's similarity metric [10]		√			
Association-based dissimilarity measure [11]			√		
Attribute-weighting k-means [12]	√	√		√	
Context-based measure [13], [14]			√	√	
Jia's distance metric [15]		√	√	√	
Entropy-based distance metric [16]		√		√	
Distance learning-based clustering [17]		√		√	
Heterogeneous coupling learning [9]		√	√	√	
Unified distance metric [18]		√	√	√	
Homogeneous distance measure [19]		√		√	
AMPHM (proposed)	√	√	√	√	√

mary of various existing methods about their consideration of significant mixed data clustering issues, where "Heterogeneity", "Intra-attribute", "Inter-attribute", "Importance", "Unification" indicate the consideration of: (1) data heterogeneity (i.e., the existence of data made of both numerical and categorical attributes); (2) intra-attribute dependence; (3) inter-attribute dependence; (4) importance of different attributes; and (5) unification of distance metrics (i.e., unification of metrics of numerical and categorical attributes). It can be seen that none of the existing methods considers all the significant issues, which may result in an unsatisfactory clustering performance.

Therefore, we propose a novel clustering algorithm that is based on original data information. Specifically, only the data objects with a relatively sparse boundary, allowing them to be clearly distinguished from their surrounding objects, are included in the micro partition, thereby reducing the grouping of objects that are not very similar based on existing distance measures. We utilize the neighborhood rough set theories, which are designed to group local similar data objects into one neighborhood set, to form the spatially local compact clusters. Our proposed method can appropriately partition data objects based on distance and density while ensuring that the granularity of each small cluster is fine-tuned. These clusters can be called micro clusters. To merge these micro clusters appropriately, we further propose a merging mechanism that initiates the entire clustering process from the most granular level, i.e., micro clusters, and gradually merges to the more rough-grained state of larger clusters, ultimately forming rational clusters. Comprehensive experiments including clustering performance and ablation studies have been conducted on various real-world datasets to verify the effectiveness of our proposed method.

The main contributions of this article are three-fold:

(1) To the best of our knowledge, this makes the first attempt to specifically formulate a clustering approach based on neighborhood rough set theory and hierarchical merging for addressing the problem of reliance on prior knowledge and assumptions in existing methods. Thus, this innovative paradigm has the potential to stimulate further exploration of unsupervised learning tasks.
(2) A novel neighborhood relation is proposed to simultaneously take into account both distance and density, which partition heterogeneous attribute data into fine-grained clusters to ensure local compact partition while the similar objects may cooperate to reject to be merged into dissimilar objects. It turns out that it applies to analyzing a wide range of data with complex object distributions.
(3) With the proposed novel merging mechanism, the algorithm hierarchically merges micro clusters, so that the micro clusters are processed to a more rough-grained state of larger clusters. This process appropriately extracts and exploits information. Therefore, it achieves a more intuitively rational inference process in forming the clusters and boosts the performance.

The rest of this article is organized as follows. Section "Related work" provides an overview of existing related techniques. In Sect. "Proposed method", the heterogeneous attribute metric, the micro partition, the merging mechanism, and the proposed algorithm are presented. Then, experimental results are shown in Sect. "Experiments". Finally, we conclude in Sect. "Conclusion".

Related work

As our approaches mainly consist of two parts, i.e., dissimilarity measurement in the existing heterogeneous attribute

data clustering method, and neighborhood set, this section mainly makes an overview of the existing above-mentioned related works, focusing on mixed data clustering methods and neighborhood rough sets.

Existing mixed data clustering methods

Early attempts like the k-prototypes method and its variants [5, 6], used one-hot encoding to represent categorical attribute values [24] as binary vectors, which allowed these attributes to be treated as numerical data. Then, numerical attributes were typically measured using the Euclidean distance, while categorical attributes employed the Hamming distance [4], which has limitations in distinguishing the dissimilarities between different value pairs. Hence, the conventional k-prototypes method overlooks the fact that the categorical attributes do not contain a well-defined space.

Thus, many advanced similarity-based methods were designed to preserve the information and cluster the mixed data well. Using statistical information from related attributes to reflect similarities in a target categorical attribute is a viable method for measuring similarity more reasonably. Various measures in this field, such as association-based [11], Ahmad's [25], and context-based [13, 14] measures, share a common approach to calculating the distance between two conditional probability distributions derived from related attributes, which indicates their dissimilarity. Recent metrics [15, 26] can effectively utilize both intra-attribute and inter-attribute information. This prevents issues with measurement when attributes are independent of each other. Most recently, a method [27] employed PCA and density to merge data. However, the above-mentioned approaches still have not well solved the heterogeneity of various categorical attributes.

Lin's distance metric [10] was one of the initial efforts to define suitable measures for categorical attributes. It calculated entropy values for attribute values to express their similarity. The entropy-based distance metric [16, 28] unifies the similarity of categorical attributes and provides an attribute weighting scheme. Moreover, the distance learning-based approaches [17, 29] innovatively learned and adjusted the order structure of the ordinal attributes and then clustering them. After this, the unified distance metric [15] was introduced to unify the measurement of distances between nominal and ordinal attributes, consistently capturing valuable relationship information. Most recently, the homogeneous distance metric [19] was introduced, focusing on the graph structure of attribute values, while the Mic2Mac [30] has been proposed to cluster heterogeneous attribute data based on the neighborhood set.

For data representation methods, common methods include numerical coding (NC), which uses one-hot encoding for nominal attributes and ranking for ordinal ones. Nonetheless, NC has its drawbacks, such as the absence of robust theoretical foundations, resulting in the curse of dimensionality for nominal attributes. Therefore, to be interpretable and facilitate a more comprehensive exploration of information, an advanced representation method introduced in [9] uses multiple kernel spaces to learn different types of value couplings, which has demonstrated superior clustering performance on categorical data. Recently, an entropy-based hierarchical clustering approach [31] has been proposed to analyze the nominal attribute, while an order forest method [32] has been proposed to tackle qualitative-attribute clustering tasks. It's essential to emphasize that the representation methods mentioned above are tailored exclusively for categorical data. The above-mentioned researches have promising prospects in switched reaction–diffusion systems [33], switched nonlinear systems [34, 35], and nanodental internet of things systems [36]. Meanwhile, competitive learning also has been introduced into categorical data clustering [37] and federated clustering [38].

Most recently, for large-scale datasets, the deep learning model-based clustering approach Mix2Vec [21] and QGRL [39] converts heterogeneous attribute values into vectors while preserving their structural distribution information, which has many promising applications, e.g., agricultural data [40], financial decision aid [41]. However, it still treats numerical attributes and categorical attributes differently and does not fully account for the heterogeneity of categorical attributes.

Overall, most existing methods for clustering heterogeneous attributes typically face one or both of the following limitations: (1) they are designed for datasets with one specific type of attributes only, and (2) they rely heavily on prior knowledge or assumptions.

Neighborhood rough sets

Neighborhood rough set (also referred to interchangeably as the neighborhood set for simplicity) is commonly used for partitioning categorical or heterogeneous attribute biomedical data [42]. Specifically, it lets each data object $\mathbf{x}_i$ to find a neighborhood set $N(\mathbf{x}_i)$, consisting of closer data objects (i.e. with smaller distance) to $\mathbf{x}_i$. The distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, as reflected by attributes A, is denoted as $D(\mathbf{x}_i, \mathbf{x}_j)$. Conventionally, there are two types of neighborhood relations, i.e., the k-nearest

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

and the δ-radius

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

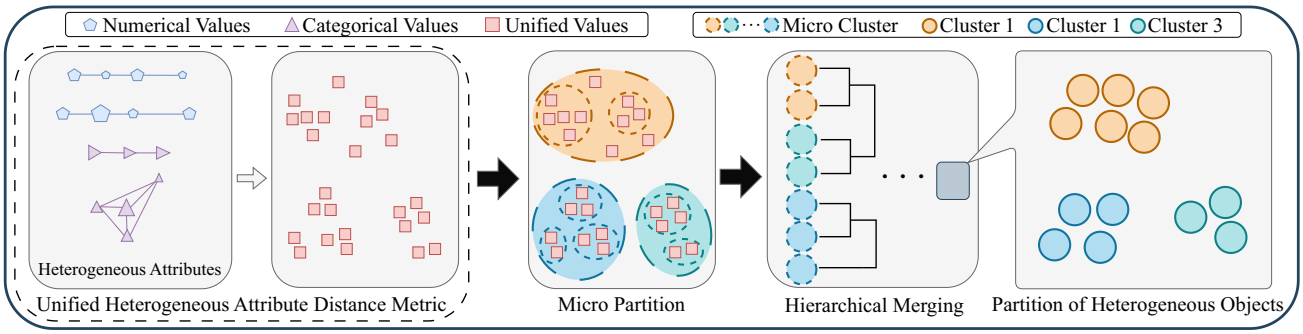

Fig. 1 Pipline of the proposed approach. Initially, we use the heterogeneous distance metric to measure different types of attributes in a unified way. Then, our proposed micro partition and hierarchical merging are used to form and merge the micro clusters, respectively. Finally, we can partition mixed attribute data

where $j, g \in \{1,2,...,n\}$, $g \neq j$, and the k in Eq. (1) represent the first k data objects with the nearest distance to $\mathbf{x}_i$. To simplify without causing any ambiguity, we use $N(\mathbf{x}_i)$ to refer to the neighborhood relation in a general sense. The classical rough set is to select categorical attribute subsets, which is applied in two Online Feature Selection (OFS) approaches [43, 44], which aim to select attribute subsets that effectively distinguish more data objects. However, these methods operate under the assumption that all attributes are categorical, which does not align with the characteristics of most real datasets.

Thus, many heterogeneous attribute selection methods [45] have been proposed. Recently, a more advanced [46] has been proposed, which incorporates entropy to improve the distinction of neighborhoods. Nevertheless, the challenging information gap between heterogeneous attributes cannot be effectively bridged through metric combination through these neighborhood sets approaches, potentially leading to significant information loss in subsequent analyses.

Proposed method

Our proposed clustering approach AMPHM primarily consists of three parts, i.e., heterogeneous attribute distance metric, neighborhood set-based micro partition, and hierarchical merging mechanism. Specifically, the second part is to partition objects appropriately based on the first proposed distance metric. Then the hierarchical merging mechanism merges neighborhood sets until the number of neighborhood sets is equal to the number of clusters. The whole technological process is shown in Fig. 1. Therefore, in this section, we start by defining the problem, then we introduce the heterogeneous attribute distance metric, neighborhood rough set-based micro partition, hierarchical merging mechanism, and the whole clustering algorithm in Section "Problem formulation", respectively.

Problem formulation

Table 3 provides a list of frequently used symbols in this paper, and their specific definitions will be provided upon their first appearance in the following text. A heterogeneous attribute dataset, denoted as $X = \{\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n\}$, is a set of n data objects. $A = \{A_1, A_2,..., A_d\}$ is the attribute set consisted of d attributes. The value set of attributes, denoted as $V = \{V_1, V_2,..., V_d\}$, stores the value domains corresponding to each attribute, with V_r representing the possible values of the r-th attribute. Clustering involves assigning n data objects in X to k appropriate clusters, denoted as $C = \{C_1, C_2,..., C_l,..., C_k\}$, where C_l is a cluster containing a set of data objects, and $X = \bigcup_{l=1}^{k} C_l$. The set of representative data objects in each cluster is denoted as $R = \{\mathbf{r}_1, \mathbf{r}_1,..., \mathbf{r}_g,..., \mathbf{r}_k\}$. A commonly used representation for clustering involves the maintenance of an $n \times k$ partition matrix $\mathbf{U}$, which indicates the assignment of n objects to k clusters. The value of the (i, l)-th entry in $\mathbf{U}$ as $u_{i,l}$. and its value is calculated by

$$u_{i,l} = \begin{cases} 1, & l = arg\min_{g} D(\mathbf{x}_i, \mathbf{r}_g) \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

according to (3) we have

$$\sum_{l=1}^{m} u_{i,l} = 1,\ 1 \leq i \leq n, \tag{4}$$

and $u_{i,l} \in \{0, 1\}$. Since the neighborhood sets for our hierarchical merging is based on $N(\mathbf{x}_i)$, while obtaining $N(\mathbf{x}_i)$ needs the inter-object distances, we adopt a common form of inter-object distance based on attributes A as

Table 3 Explanation of symbols

Symbol	Explanation
X	Dataset
$\mathbf{x}_i$	i-th data object
x_i^r	r-th value of $\mathbf{x}_i$
A	Attribute set
A_r	r-th attribute
V_r	Possible value of A_r
v_h^r	h-th value of V_r
$\mathrm{v_r}$	Number of possible values of A_r
R	Representative objects set
$\mathbf{r}_g$	g-th representative object R
d	Number of attributes
n	Number of objects
C_l	l-th cluster
$u_{i,l}$	A value indicating the affiliation between $\mathbf{x}_i$ and c_l
$N^{\xi}(\mathbf{x}_i)$	The micro partition of $\mathbf{x}_i$
ρ_i	i-th density
ξ_i	i-th density gap
$D(\cdot,\cdot)$	Data dissimilarity
w^{rt}	Weight indicating the importance of A_t to A_r

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

where $x_i^r \in V_r$ is the value of $\mathbf{x}_i$ on attribute A_r, while $D^r(x_i^r, x_j^r)$ measures the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in terms of A_r. In the following subsections, we present how to define $D^r(x_i^r, x_j^r)$, $N(\mathbf{x}_i)$, and the hierarchical merging mechanism, respectively.

HADM: heterogeneous attribute distance metric

To construct the unified distance metric $D^r(x_i^r, x_j^r)$ concerning different heterogeneous attributes A_r, we employ a transformation cost as a metric. This cost quantifies the effort needed to convert one Conditional Probability Distribution (CPD) into another. To provide a more intuitive explanation of the principles, we will begin by specifying the CPD and formulating the distance between possible values of a categorical attribute. Afterward, we demonstrate how this distance unifies both categorical and numerical cases. Lastly, we will derive the object-level distance. The CPD of an attribute A_t, which has v^t possible values $V_t = \{v_1^t, v_2^t, ..., v_{\mathrm{v_t}}^t\}$, given a possible value v_g^r of another attribute A_r, can be expressed as

$$\mathbf{P}_g^{r \leftarrow t} = [p(v_1^t | v_g^r), p(v_2^t | v_g^r), ..., p(v_{\mathrm{v_t}}^t | v_g^r)], \tag{6}$$

where $p(v_h^t | v_g^r)$ is the conditional probability of v_h^t as given v_g^r. We represent the CPD as $\mathbf{P}_g^{r \leftarrow t}$. The superscript $r \leftarrow t$ indicates that this CPD describes the gth possible value of A_r according to the values of A_t. For simplicity, we denote $r \leftarrow t$ as rt from here on. The difference between two CPDs, such as $\mathbf{P}_g^{rt}$ and $\mathbf{P}_q^{rt}$, effectively reflects the dissimilarity between two possible values, i.e., v_g^r and v_q^r. Hence, we employ Earth Mover's Distance (EMD) [47, 48], originally developed for calculating the transformation cost between two histogram descriptors, to quantitatively measure the difference between the CPDs describing two categorical possible values. Consequently, the distance between v_g^r and v_q^r as represented by A_t can be denoted using EMD as

$$D^{rt}(v_i^r, v_j^r) = \sigma((\mathbf{P}_i^{rt} - \mathbf{P}_j^{rt}, \mathbf{0})) \cdot \mathbf{1}, \tag{7}$$

where $\sigma(\cdot,\cdot)$ matches each pair of corresponding bits in two vectors and selects the maximum value. The $\mathbf{0}$ and $\mathbf{1}$ are v^t-dimensional vectors with all elements equal to 0 and 1, respectively. Due to varying degrees of inter-attribute dependence, different attributes A_ts may have different contributions based on their respective weights w^{rt} when forming the overall distance $D^{rt}(v_g^r, v_h^r)$ between v_g^r and v_h^r by

$$D^r(v_i^r, v_j^r) = \sum_{A_t \in A} D^{rt}(v_i^r, v_j^r) \cdot w^{rt}. \tag{8}$$

The contribution w^{rt} indicated in Eq. (8) by using this dependence, which measures the inter-attribute dependence between A_r and A_t, can be written as

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

Considering the possible values of a categorical attribute as concepts proposed in [49], Eq. (9) essentially measures the average inter-concept distances of A_r, as indicated by A_t. To clarify Eq. (9), we consider an extreme example. When A_r and A_t are identical, they are perfectly interdependent, and the corresponding $D^{rt}(v_g^r, v_h^r)$ always reaches the maximum value of "1" for any combinations of g and h with $g \neq h$, as stated in Eq. (7). Therefore, the corresponding w^{rt} also reaches the maximum value of "1", signifying a complete 100% inter-attribute dependence. Based on Eqs. (7)–(9), the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ w.r.t. attribute A can be calculated using Eq. (5). Now we demonstrate the applicability of the defined distance metric to both categorical and numerical attributes while satisfying the properties of a metric.

Theorem 1 *Equation (8) treats both categorical and numerical attributes in a homogeneous way, under the assumption that the distance space for each numerical attribute is an independent one-dimensional continuous Euclidean distance space with a domain of [0, 1].*

Proof Each numerical attribute space can be considered a linear space with an infinite number of evenly spaced possible values. Because they are independent, w^{rt} is always equal to 0 for numerical attributes where $r \neq s$. Therefore, we should only consider the special case, i.e., r equals s, where $w^{rt} = \lim_{h\to\infty}(\sum_{g=1}^{h} 1/h)/1 = 1$ according to Eq. (9). Here, $h \to \infty$ represents the number of intervals between adjacent possible values. The denominator "1" reflects the presence of only one inter-concept distance because there are only two concepts, "0" and "1". Therefore, when $r = s$, the EMD with the use of Euclidean distance simplifies to $D^{rt}(v_g^r, v_h^r) = |v_g^r - v_h^r|$, which is essentially equivalent to the Euclidean distance in $D(\mathbf{x}_i, \mathbf{x}_j)$ as described in Eq. (5).□

Theorem 2 *$D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a distance metric.*

Proof As Eq. (7) is a metric, it follows that Eq. (8) based on Eq. (7) is also a metric. Additionally, Eq. (5), which is computed using finite arithmetic operations based on Eq. (8), ensures that $D(\mathbf{x}_i, \mathbf{x}_j)$ satisfies all the properties of a metric:

(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) $D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_k) + D(\mathbf{x}_k, \mathbf{x}_j)$.

□

NRSMP: neighborhood rough set-based micro partition

The existing conventional neighborhood rough set $N^k(\mathbf{x}_i)$ and $N^\delta(\mathbf{x}_i)$ (i.e., Eqs. (1) and (2)) create n neighborhood sets that may partially overlap with each other. This overlapping can be problematic for partitions because it introduces high computational costs when considering each current attribute. Furthermore, in the case of unevenly distributed data objects, $N^k(\mathbf{x}_i)$ might inappropriately group dissimilar objects into the same neighborhood set, while $N^\delta(\mathbf{x}_i)$ could create neighborhood sets that overlap with boundaries. These neighborhood sets located in the central regions, may not effectively support our partitioning but would require a significant computational cost.

To effectively distinguish class boundaries and partition objects more efficiently while reducing computational costs, we introduce a novel approach called the neighborhood rough set-based micro partition. This approach aims to create non-overlapping neighborhood sets. To achieve this, we begin by selecting representative data objects, which gather neighbors based on the concept of a density gap.

Definition 1 Density gap: Density gap ξ_i, for an object $\mathbf{x}_i$ with a density of ρ_i, represents the minimum distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ with a higher density ρ_j. It can be expressed as

$$\xi_i = minD(\mathbf{x}_i, \mathbf{x}_j), \quad s.t.\ \rho_i < \rho_j \ and\ \mathbf{x}_i \in X \backslash \mathbf{x}_j, \tag{10}$$

where ρ_i and ρ_j are the density of $\mathbf{x}_i$ and $\mathbf{x}_i$, respectively. And the $X\backslash\mathbf{x}_j$ is the dataset excluding $\mathbf{x}_j$. Moreover, for the object with the highest density, its density gap is defined as the maximum $D(\mathbf{x}_i, \mathbf{x}_j)$ with $\mathbf{x}_j \in X$.

The density ρ_i in Definition 1 can be computed by

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

where g_i represents the rank of a neighbor object $\mathbf{x}_{\langle i,g_i\rangle} \in D$ w.r.t. $\mathbf{x}_i$. This rank signifies that $\mathbf{x}_{\langle i,g_i\rangle}$ is the g_i-th closest object to $\mathbf{x}_i$ among all n data objects. More specifically, we initially construct the *ascending* micro partition set MP_i concerning $\mathbf{x}_i$ as $MP_i = \{\mathbf{x}_{\langle i,1\rangle}, \mathbf{x}_{\langle i,2\rangle}, ..., \mathbf{x}_{\langle i,n\rangle}\}$, where $\mathbf{x}_{\langle i,1\rangle} \equiv \mathbf{x}_i$ and $MP_i(j) = \mathbf{x}_{\langle i,j\rangle}$. Next, while iterating through MP_i from left to right, we select the object $\mathbf{x}_{\langle i,h\rangle}$ that initially fulfills the condition $D(\mathbf{x}_i, \mathbf{x}_{\langle i,h\rangle})/ D(\mathbf{x}_{\langle i,h-1\rangle})/(h\text{-}1)$ determines the value of g_i as $g_i = h - 2$. The density computation intuitively selects neighbor objects, ensuring objects beyond a significant class boundary are not incorporated into the $\mathbf{x}_i$ neighborhood set.

Clearly, a representative object in a neighborhood set should be encircled by neighbors of lower density and also be situated at a considerable distance from other representative objects. As per Definition 1, objects with higher density gaps are better suited to serve as representative objects. Consequently, we initiate the micro partitioning process by ranking data objects in *descending* order based on density gaps and then creating micro partition accordingly

$$N^\xi(\mathbf{x}_i) = \left\{ \bigcup_{j=1}^{g_i} MP_i(j) \right\} \setminus \left\{ \bigcup_{\xi_q \leq \xi_i} N^\xi(\mathbf{x}_q) \right\}. \tag{12}$$

This process continues until the existing neighborhood sets incorporate all objects. The representative objects in all neighborhood sets are denoted as $NR = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_s\}$, where $\mathbf{z}_g$ is a representative object in g-th neighborhood set. The proposed neighborhood rough set-based micro partition is summarized in Algorithm 1. Another critical issue that will be addressed below is how to merge the obtained neighborhood set $N^\xi(\mathbf{x}_i)$ by the merging mechanism.

Hierarchical merging mechanism

Based on the neighborhood rough set-based micro partition proposed in Section "namerefsc:NRSMP", we introduce the hierarchical mechanism to merge micro partitions appropriately. We iteratively update neighborhood sets $N^\xi(\mathbf{x}_i)$ and dataset X at each layer, following these two steps: 1) fix X,

Algorithm 1 NRSMP: Neighborhood Rough Set-based Micro Partition

Input: Dataset X , distance metric D.
Output: Neighborhood sets $N^{\xi}(\mathbf{x}_i)$, neighborhood representative object set NR.

1: **for** $i = 1$ to n **do**
2: Compute density ρ_i for $\mathbf{x}_i$ according to Eq. (11);
3: **end for**
4: **for** $i = 1$ to n **do**
5: Compute density gap ξ_i for $\mathbf{x}_i$ according to Eq. (10);
6: **end for**
7: **for** $i = 1$ to n **do**
8: **if** $\rho_i > 0$ **then**
9: Add $\mathbf{x}_i$ as representative object to NR;
10: **end if**
11: Compute $N^{\xi}(\mathbf{x}_i)$ according to Eq. (12);
12: **end for**

Algorithm 2 AMPHM: Adaptive Micro Partition and Hierarchical Merging Clustering Algorithm

Input: Dataset X, number of clusters k.
Output: Partition $\mathbf{U}$.

1: Set neighborhood sets for each object;
2: **while** $|NR| > k$ **do**
3: Compute D according to Eq. (5);
4: Compute $N^{\xi}(\mathbf{x}_i)$ and NR by Algorithm 1;
5: Update $X = NR$;
6: **end while**
7: Update representative object set $R = NR$;
8: Compute $\mathbf{U}$ according to Eq. (3).

compute $N^{\xi}(\mathbf{x}_i)$ and representative object set NR by Algorithm 1, and 2) fix NR, update X by the representative object set NR, i.e., $X = NR$. These two steps repeat until the number of neighborhood representative objects m is equal to the number of clusters k. A more specific explanation will be shown in Section "Clustering algorithm" below.

Clustering algorithm

With the neighborhood rough set-based micro partition in Section "namerefsc:NRSMP" proposed based on the heterogeneous attribute distance metric in Section "HADM: heterogeneous attribute distance metric" and the merging mechanism introduced in Section "Hierarchical merging mechanism", we propose the overall clustering algorithm. The whole clustering algorithm named Adaptive Micro Partition and Hierarchical Merging (AMPHM) is summarized in Algorithm 2.

AMPHM first partitions data into micro clusters, obtaining the neighborhood representative object set NR, then uses NR as the next layer dataset X to further partition and merge the NR of the previous layer while maintaining the overall cluster result in each layer. This algorithm continues until the number of neighborhood representative objects m is equal to the number of clusters k. Since our proposed approach

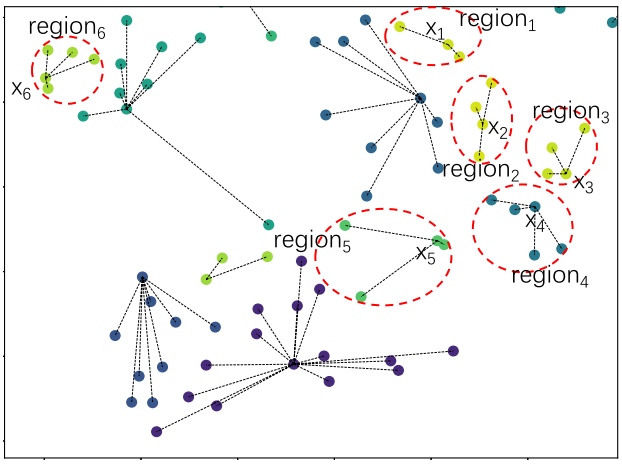

Fig. 2 Hierarchical merging with the micro sets in the form of the region. The regions are marked by red dashed circles, and different regions are partitioned by black solid lines. The points where black lines meet in each region are representative objects, with merging limited by density gap

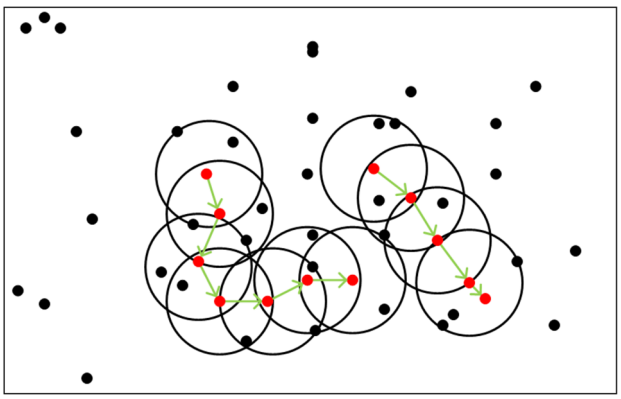

Fig. 3 In hierarchical clustering, data points or clusters are merged in each layer. The most common approach is one-by-one pairwise merging. In contrast to AMPHM, which enables the combination of multiple clusters simultaneously, the process of merging in this case involves the combination of individual clusters one at a time

shares similarities with hierarchical clustering, Figs. 2 and 3 illustrate their partial clustering processes to highlight distinctions between the two methods.

Theorem 3 *The time complexity of the AMPHM clustering algorithm is* $O(d_c^2 n + n \log n)$ *at any round.*

Proof For the worst-case analysis, we assume that all the attributes are categorical, i.e., $d_c = |A|$, and $v = max(v^1, v^2, ..., v^{d_c})$. The distances D correspond to attributes in A and are updated previously before each time we partition the neighborhood set N^{ξ}. Additionally, the time complexity at each round is the same. Therefore, we analyze the complexity of updating D and N^{ξ} once, respectively.

To update D requires obtaining $d_c \times d_c$ pairs of CPDs by scanning the n objects in X once, resulting in a time complexity of $O(nd_c^2)$. To compute the distances between a pair of

intra-attribute possible values using EMD in Eq. (7), it would take $O(v)$ time for each attribute. To calculate the weights for each pair of attributes, based on Eq. (8), it would require $v(v-1)/2$ calculations, resulting in $O(v^2)$ complexity for each attribute. For all the d_c attributes, we need to prepare weights between d_c pairs of attributes, and each attribute has $v(v-1)/2$ distances to compute. Therefore, the complexity for acquiring D is $O(nd^2 + v^3 d_c^2 + v^2 d_c^2)$, which can be simplified to $O(nd^2 + v^3 d_c^2)$.

To compute N^{ξ}, we need to parallelly compute an $n \times n$ distance matrix and sort all its rows in ascending order, which takes $O(n + n \log n)$ time. Then, for all the n objects, each one considers the remaining $n-1$ ones to identify its k_i value for obtaining the density ρ_i in Eq. (11), with a $O(n)$ complexity. To determine n density gaps for all objects, we need to examine the remaining $n-1$ objects for each one to identify the object with the highest density and the shortest distance. This process still has a complexity of $O(n)$. Next, the n density gaps need to be sorted in descending order, which requires $O(n \log n)$ time. Following that, we can create up to n neighborhood sets, where each set considers at most k_i objects based on the already sorted distances, resulting in a complexity of $O(nk_i)$. Therefore, the total complexity for calculating N^{ξ} is $O(n + n \log n + nlogn + nk_i)$, which simplifies down to $O(n \log n)$.

Thus, the overall time complexity of AMPHM at each round is $O(nd^2 + v^3 d_c^2 + n \log n)$. As v is a very small constant, typically with values in the interval $[1, 6]$, satisfying $v^3 \ll n$ for almost all real datasets, the final time complexity of AMPHM is $O(d_c^2 n + n \log n)$. □

Note that the computation cost of Algorithm 2 during the AMPHM process will drop significantly after updating the dataset X to NR in the first round.

Experiments

To evaluate the performance of AMPHM, three experiments have been conducted. (1) Clustering Performance Evaluation: AMPHM was compared with existing methods on 9 real datasets using two validity indices to evaluate the clustering performance; (2) Ablation Study: Various ablated versions of AMPHM are compared to prove its effectiveness from different aspects; and (3) Visualization: The cluster discrimination ability of the proposed method is compared with various existing counterparts.

Experimental settings

In the experiments, we compare the proposed AMPHM with ten other methods. These include eight clustering approaches from existing literature and two variants of AMPHM. Five counterparts proposed in recent years, including Jia's Distance Metric (JDM) [15], Coupled Metric Similarity (CMS) [26], Entropy-based Distance Metric (EDM) [16], Unified Distance Metric (UDM) [18] combined with k-modes (KMD) [50] and k-prototypes (KPT) [5], according to the attribute composition of datasets, while Object-Cluster Iterative Learning (OCIL) [20] designed for datasets that contain both numerical attributes and categorical attributes been selected, where JDM, CMS, EDM, and UDM are the state-of-the-arts. We also select three conventional clustering algorithms for comparison, i.e., w-k-means clustering algorithm (WKM) [12], which are representative algorithms in the attribute-weighting clustering stream, and original versions of KMD and KPT. We configured the parameters of the aforementioned methods to align with the values recommended in their respective papers. Additionally, we created two variations of AMPHM, denoted AMPHM$^{\mathrm{I}}$ and AMPHM$^{\mathrm{II}}$, to conduct ablation studies. More information about these two variants is available in Section "Ablation study".

Two validity indices are utilized to evaluate the clustering performance. CA [51] is a standard metric that calculates the match rate by determining the optimal permutation mapping between the identified clusters and their corresponding true labels. Adjusted Rand Index (ARI) [52–54] is an advanced validity index that is more discriminative in quantifying the performance of different counterparts. More specifically, ARI is an enhanced version of the Rand Index (RI) that mitigates issues related to random labeling. The RI measures the similarity between the clusters obtained by the algorithm and the actual data labels. It is calculated based on the proportion of object pairs that are assigned to the same or different clusters in both the predicted and actual clusters. CA falls within the range of [0, 1] and ARI has a range of values from $[-1, 1]$. All the selected validity indices show better clustering performance with higher values.

We use real-world datasets from various domains, including medicine, biology, and sociology, to conduct comprehensive experiments. All the datasets are publicly available and commonly used. The statistics for these datasets are summarized in Table 4. Datasets 1 to 7 are publicly available datasets obtained from the University of California, Irvine (UCI) machine learning repository.[4] Datasets 8, 9 collected from the Weka website.[5] All the datasets are preprocessed by removing objects with missing values. All the experiments are performed in the same computer with an Apple M1 CPU and 8GB RAM and are programmed using Python 3.9.6 and MATLAB 2021B.

Clustering performance evaluation

The clustering performance has been evaluated using the CA and ARI indexes, with the results presented in Tables 5 and 6,

[4] https://archive.ics.uci.edu/datasets.

[5] https://waikato.github.io/weka-wiki/datasets/.

Table 4 Statistics of the 9 Utilized Datasets. d_c, d_u, and n are the numbers of categorical, numerical attributes, and data objects, respectively. k^* is the true number of clusters and we set $k = k^*$ in the experiments

No.	Dataset	Abbrev	d_c	d_u	n	k^*
1	Dermatology	DT	33	1	366	6
2	Autism-adolescent	AA	2	7	104	2
3	Common Toad	CT	12	2	189	2
4	Hayes-roth	HR	4	0	132	3
5	Breast cancer	BC	9	0	286	2
6	Lymphography	LG	18	0	148	4
7	Congressional voting	VT	16	0	435	2
8	Employee selection	ES	4	0	488	9
9	Social workers	SW	10	0	1000	4

respectively. The best and second-best outcomes for each dataset are highlighted in **boldface** and underline, respectively. To maintain compactness, the results of KMD for categorical data and KPT for heterogeneous attribute data are combined in the same column. The "Ave. Rank" row in Tables 5 and 6 computes the average rank of the performance of the counterparts on all datasets.

The notable observations are: (1) AMPHM performs the best on most datasets in general, indicating its competitive clustering performance. (2) Although AMPHM does not perform the best on CT, VT, and ES datasets in the CA index, the best one differs on these datasets while AMPHM always outperforms them on other datasets, which also demonstrates that AMPHM is highly competitive. (3) On CT, LG, and VT datasets, AMPHM does not perform the best in terms of the ARI index. The reason could be that many attribute values of these datasets concentrate on two sides of the possible values. For such attribute values, AMPHM only has almost one intra-attribute to be weighted during clustering, which makes AMPHM downgrade into a conventional attribute weighting approach. As the ES dataset has few data objects in each class, AMPHM may not have sufficient information to form enough micro clusters and distinguish the compact clusters thus data objects have less ability to reject to be merged into dissimilar objects. The reason why AMPHM has competitive clustering performance compared to other counterparts is that our proposed heterogeneous attribute distance metric enables AMPHM to appropriately exploit distance structure information collaboratively represented by the heterogeneous attributes, and thus partition objects into prominent compact clusters, which are separable to achieve better clustering accuracy. Thus, it can be concluded that our AMPHM is very competitive in terms of clustering accuracy.

Table 5 Clustering performance evaluated by CA index. The average performance rank is reported in the Ave. Rank row

Dataset	KMD/KPT	WKM	OCIL	JDM	CMS	EDM	UDM	AMPHM
DT	0.5538±0.10	0.6234±0.09	0.6750±0.10	0.6654±0.10	0.6017±0.14	0.5872±0.10	0.6845±0.11	**0.7677±0.00**
AA	0.5301±0.03	0.5254±0.02	0.5186±0.03	0.5788±0.05	0.5244±0.03	0.5577±0.03	0.5579±0.02	**0.5962±0.00**
CT	0.5298±0.02	0.5234±0.03	0.5056±0.00	0.5224±0.02	0.5259±0.02	0.5365±0.03	**0.5781±0.02**	0.5450±0.00
HR	0.3638±0.01	0.4078±0.05	0.3761±0.04	0.3753±0.02	0.4048±0.04	0.4070±0.03	0.4040±0.03	**0.4167±0.00**
BC	0.5187±0.02	0.5835±0.09	0.5414±0.06	0.5823±0.10	0.5281±0.04	0.5304±0.02	0.5685±0.19	**0.7657±0.00**
LG	0.4531±0.04	0.4385±0.05	0.5003±0.04	0.4730±0.04	0.4189±0.05	0.4523±0.04	0.4704±0.04	**0.5608±0.00**
VT	0.8643±0.00	0.8570±0.07	**0.8810±0.00**	0.8681±0.00	0.8649±0.01	0.8317±0.10	0.8721±0.00	0.8736±0.00
ES	0.3674±0.03	0.3750±0.04	0.3839±0.04	0.3505±0.03	**0.4016±0.04**	0.3664±0.02	0.3681±0.03	0.3934±0.00
SW	0.3924±0.03	0.3750±0.03	0.3732±0.03	0.3337±0.03	0.3310±0.03	0.3321±0.01	0.3744±0.03	**0.4350±0.00**
Ave. Rank	5.7778	4.6667	4.5556	4.8889	5.6667	5.6667	3.4444	1.3333

Table 6 Clustering performance evaluated by ARI index. The average performance rank is reported in the Ave. Rank row

Dataset	KMD/KPT	WKM	OCIL	JDM	CMS	EDM	UDM	AMPHM
DT	0.4222±0.12	0.5091±0.09	0.6063±0.10	0.6140±0.13	0.5176±0.17	0.4393±0.12	0.6266±0.15	**0.6776±0.00**
AA	−0.0031±0.01	−0.0055±0.01	−0.0073±0.01	0.0182±0.03	−0.0086±0.01	0.0063±0.01	−0.0153±0.01	**0.0185±0.00**
CT	−0.0081±0.01	−0.0079±0.01	−0.0208±0.00	−0.0142±0.01	−0.0075±0.01	0.0017±0.01	**0.0133±0.02**	−0.0010±0.00
HR	−0.0121±0.00	0.0069±0.02	−0.0035±0.02	−0.0063±0.01	0.0075±0.02	0.0080±0.02	0.0070±0.02	**0.0094±0.00**
BC	−0.0041±0.00	0.0396±0.07	0.0112±0.04	0.0405±0.07	0.0029±0.02	0.0066±0.01	0.0620±0.02	**0.1092±0.00**
LG	0.1128±0.04	0.0845±0.04	**0.1817±0.05**	0.1231±0.04	0.0884±0.04	0.0893±0.03	0.1317±0.05	0.1293±0.00
VT	0.5297±0.00	0.5273±0.11	**0.5795±0.00**	0.5411±0.01	0.5319±0.03	0.4776±0.17	0.5527±0.01	0.5568±0.00
ES	0.1621±0.02	0.1717±0.03	0.1929±0.02	0.1666±0.02	**0.2115±0.03**	0.1627±0.04	0.2106±0.02	0.1725±0.00
SW	0.0571±0.02	0.0456±0.02	0.0516±0.02	0.0519±0.01	0.0505±0.02	0.0591±0.01	0.0760±0.01	**0.0847±0.00**
Ave. Rank	6.3333	5.8889	4.4444	4.5556	5.1111	4.8889	2.8889	1.8889

Ablation study

Ablation studies were carried out, and the clustering performance was assessed using the ARI index. To evaluate the effectiveness of the heterogeneous attribute dissimilarity metric, we modified AMPHM to employ only the conventional Euclidean distance and Hamming distance, creating a variant AMPHM$^{\mathrm{I}}$. To evaluate the effectiveness of the hierarchical merging mechanism, we conducted a comparison between AMPHM and a variation called AMPHM$^{\mathrm{II}}$, which incorporates the prototype selection strategy from the typically k-prototypes approach to cluster the representative objects after the initial formation of neighborhood sets. The clustering performance and the Ave. Rank of AMPHM and its two variations are shown in Fig. 4.

It can be observed from Fig. 4 that AMPHM outperforms its two variations, indicating the effectiveness of the heterogeneous attribute distance metric and hierarchical merging mechanism. More specifically, AMPHM outperforms the AMPHM$^{\mathrm{I}}$ on all the nine datasets. This illustrates that the heterogeneous attribute distance metric $D(\cdot,\cdot)$ proposed in Sect. 3.2 can sufficiently harness the original data information for different data distributions, which testifies to the effectiveness of our distance metric. Moreover, AMPHM outperforms AMPHM$^{\mathrm{II}}$ on eight datasets and AMPHM$^{\mathrm{I}}$ performs better than AMPHM$^{\mathrm{II}}$ on six datasets. This shows that the hierarchical merging mechanism effectively merges micro clusters into compact clusters, indicating the competitiveness of our proposed merging mechanism compared to the conventional k-means-type clustering process. The reason why AMPHM$^{\mathrm{I}}$ performs worse than AMPHM$^{\mathrm{II}}$ in some datasets (i.e. CT, LG, and VT) would be that the former adopts the simplest Euclidean and Hamming distances.

Visualization

In Fig. 5, t-SNE [55] is utilized to illustrate the cluster discrimination ability of AMPHM. We first cluster the dataset DT using KPT, WKM, JDM, CMS, UDM, and AMPHM. Then, data is processed by t-SNE into two-dimensional according to the distance matrix of data objects generated by the distance metrics adopted by the compared methods. It is visualized with data points marked in green for those that cluster correctly and red for those that cluster incorrectly, while different clusters are marked in different markers. Intuitively, the fewer red data points indicate a better clustering accuracy. We can observe from Fig. 5 that AMPHM has significantly fewer red points than other counterparts, which indicates the effectiveness of AMPHM. More specifically, by using NRSMP proposed in Sect. "NRSMP: neighborhood rough set-based micro partition" to partition micro clusters, AMPHM can collaboratively distinguish dissimilar objects. As a result, the merging mechanism appropriately merges

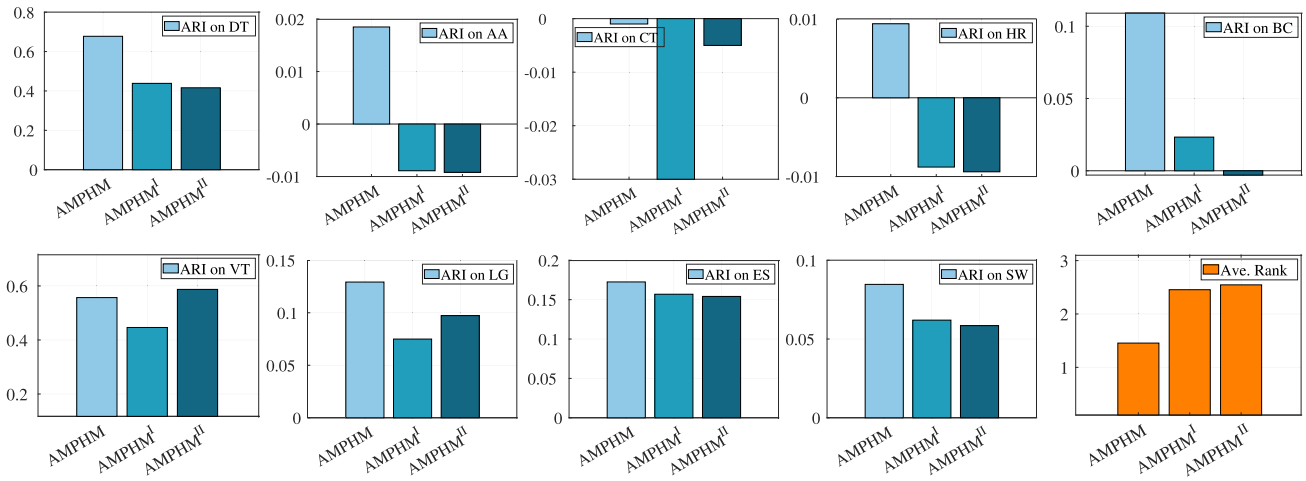

Fig. 4 Clustering performance and Ave. rank of AMPHM, AMPHM$^{\mathrm{I}}$ and AMPHM$^{\mathrm{II}}$ on all the 9 datasets, where a better measure yields a higher value

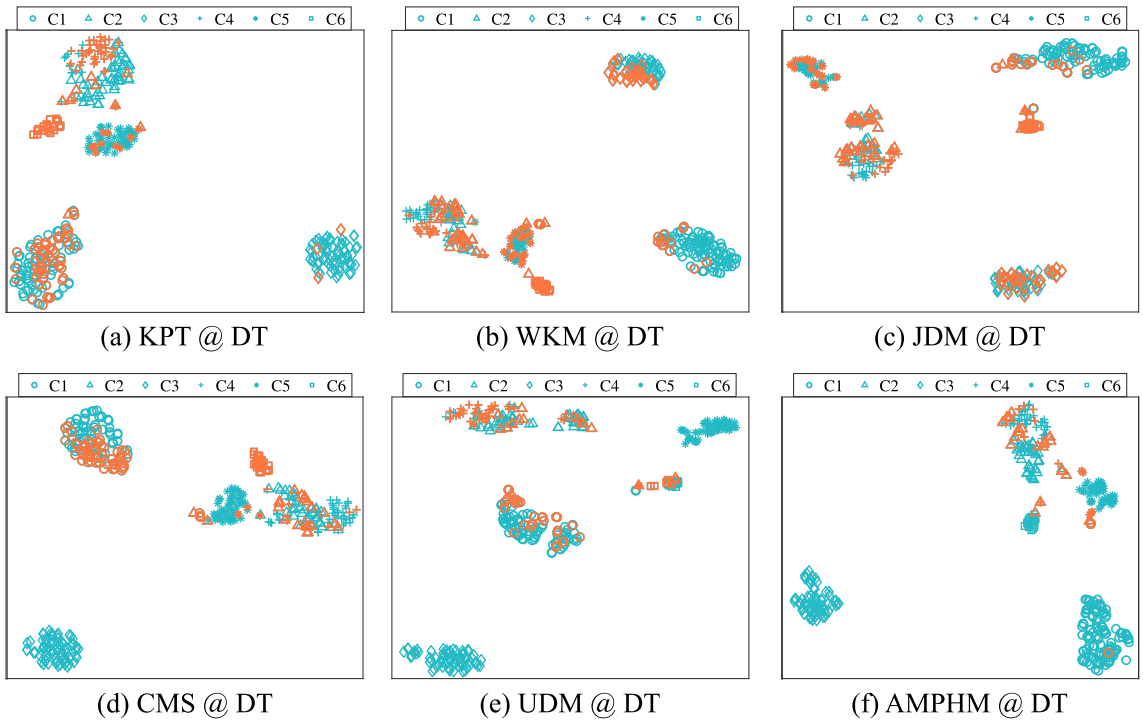

Fig. 5 t-SNE visualization of the DT datasets after KPT (**a**), WKM (**b**), JDM (**c**), CMS (**d**), UDM (**e**), and AMPHM (**f**) clustering, where objects are marked in different markers indicate their ground truth labels, while data points marked in red for those that cluster incorrectly

similar micro clusters and finally forms the final "true" clusters, indicating stronger cluster discrimination compared to other methods.

Conclusion

In this paper, we have proposed a novel clustering algorithm called Adaptive Micro Partition and Hierarchical Merging (AMPHM) for heterogeneous attribute data clustering. AMPHM mainly solved the two challenges in clustering real datasets, i.e., most existing clustering approaches relying on certain assumptions or models, and the clustering limitation on heterogeneous attributes, by adopting (1) a unified distance metric to quantify distances represented by heterogeneous attributes, (2) a micro partition mechanism based on neighborhood set to appropriately partition data objects into many compact micro clusters, and (3) a merging mechanism to hierarchically merge micro clusters. We simultaneously address the bias brought by prior clustering knowledge and attribute heterogeneity, which is far more challenging than coping with only one of them. The effectiveness of AMPHM

is illustrated by the comprehensive experiments including comparison with state-of-the-art counterparts and ablation study on various public datasets. Moreover, AMPHM is accurate and interpretable. Our future research will focus on the analysis of heterogeneous attribute data with more challenging issues, e.g., clustering mixed data under dynamic or federated environments.

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 2023A1515012855, the General Research Fund of RGC under Grants: 12201321, 12202622, and 12201323, and the RGC Senior Research Fellow Scheme under grant SRFS2324-2S02.

Data availability The datasets used in all the experiments of this paper are publicly available. The links of the source repositories are provided in Sect. "Experimental settings".

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Hongshi X, Ma C, Lian J, Kui X, Chaima E (2018) Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou. Chin J Hydrol 563:975–986. https://doi.org/10.1016/j.jhydrol.2018.06.060
2. Alan A (2002) Categorical data analysis. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/0471249688
3. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J (2023) K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf Sci 622:178–210. https://doi.org/10.1016/j.ins.2022.11.139
4. Arabie P, Baier Newark D, Critchley Cottbus F, Keynes M (2006) Studies in classification, data analysis, and knowledge organization. Springer, New York
5. Huang Z (1997) Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining, pp 21–34
6. Mohamed ABHK, Chiheb-Eddine BN, Nadia E (2015) MapReduce-based k-prototypes clustering method for big data. In: Proceedings of the IEEE international conference on data science and advanced analytics, pages 1–7. IEEE. https://doi.org/10.1109/DSAA.2015.7344894
7. Qian Y, Li F, Liang J, Liu B, Dang C (2016) Space structure and clustering of categorical data. IEEE Trans Neural Netw Learn Syst 27(10):2047–2059. https://doi.org/10.1109/TNNLS.2015.2451151
8. Jian S, Pang G, Cao L, Kai L, Gao H (2019) CURE: flexible categorical data representation by hierarchical coupling learning. IEEE Trans Knowl Data Eng 31(5):853–866. https://doi.org/10.1109/TKDE.2018.2848902
9. Zhu C, Cao L, Yin J (2022) Unsupervised heterogeneous coupling learning for categorical representation. IEEE Trans Pattern Anal Mach Intell 44(1):533–549. https://doi.org/10.1109/TPAMI.2020.3010953
10. Lin D (1998) An Information-Theoretic Definition of Similarity. In: Proceedings of the 15th International Conference on Machine Learning, pages 296–304
11. Le SQ, Ho TB (2005) An association-based dissimilarity measure for categorical data. Pattern Recogn Lett 26(16):2549–2557. https://doi.org/10.1016/j.patrec.2005.06.002
12. Huang JZ, Ng MK, Rong H, Li Z (2005) Automated variable weighting in k-means type clustering. IEEE Trans Pattern Anal Mach Intell 27(5):657–668. https://doi.org/10.1109/TPAMI.2005.95
13. Ienco D, Pensa Ruggero G, Meo R (2009) Context-Based Distance Learning for Categorical Data Clustering. In: Advances in Intelligent Data Analysis VIII 83–94. https://doi.org/10.1007/978-3-642-03915-7_8
14. Ienco D, Pensa RG, Meo R (2012) From context to distance. ACM Trans Knowl Discov Data 6(1):1–25. https://doi.org/10.1145/2133360.2133361
15. Jia H, Cheung Y, Liu J (2016) A new distance metric for unsupervised learning of categorical data. IEEE Trans Neural Netw Learn Syst 27(5):1065–1079. https://doi.org/10.1109/TNNLS.2015.2436432
16. Zhang Y, Cheung Y-M, Tan KC (2020) A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering. IEEE Trans Neural Netw Learn Syst 31(1):39–52. https://doi.org/10.1109/TNNLS.2019.2899381
17. Zhang Y, Cheung Y (2020) An ordinal data clustering algorithm with automated distance learning. In: Proceedings of the AAAI Conference on Artificial Intelligence 34:6869–6876. https://doi.org/10.1609/AAAI.V34I04.6168
18. Zhang Y, Cheung Y-M (2022) A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. IEEE Trans Cybern 52(2):758–771. https://doi.org/10.1109/TCYB.2020.2983073
19. Zhang Y, Cheung Y (2022) Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. IEEE Trans Pattern Anal Mach Intell 44(7):3560–3576. https://doi.org/10.1109/TPAMI.2021.3056510
20. Cheung Y, Jia H (2013) Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recogn 46(8):2228–2238. https://doi.org/10.1016/j.patcog.2013.01.027
21. Zhu C, Zhang Q, Cao L, Abrahamyan A (2020) Mix2Vec: unsupervised mixed data representation. In: Proceedings of the IEEE 7th International Conference on Data Science and Advanced Analytics, pp 118–127. IEEE. https://doi.org/10.1109/DSAA49011.2020.00024
22. Zhang Y, Cheung Y, Zeng A (2022) Het2Hom: representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pages 3758–3765, California. https://doi.org/10.24963/ijcai.2022/522
23. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323. https://doi.org/10.1145/331499.331504

24. Alamuri M, Surampudi BR, Negi A (2014) A survey of distance/similarity measures for categorical data. In: Proceedings of the International Joint Conference on Neural Networks, pages 1907–1914. IEEE. https://doi.org/10.1109/IJCNN.2014.6889941
25. Ahmad A, Dey L (2007) A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recogn Lett 28(1):110–118. https://doi.org/10.1016/j.patrec.2006.06.006
26. Jian S, Cao L, Kai L, Gao H (2018) Unsupervised Coupled Metric Similarity for Non-IID Categorical Data. IEEE Trans Knowl Data Eng 30(9):1810–1823. https://doi.org/10.1109/TKDE.2018.2808532
27. Chen L, Guo Q, Liu Z, Zhang S, Zhang H (2021) Enhanced synchronization-inspired clustering for high-dimensional data. Complex Intell Syst 7(1):203–223. https://doi.org/10.1007/s40747-020-00191-y
28. Zhang Y, Cheung Y (2018) Exploiting order information embedded in ordered categories for ordinal data clustering. In: Proceedings of the 24th International Symposium on Methodologies for Intelligent Systems, pp. 247–257. https://doi.org/10.1007/978-3-030-01851-1_24
29. Wang P, Zhang Y, Zhang Y, Lu Y, Li M, Cheung Y (2024) Clustering by learning the ordinal relationships of qualitative attribute values. In: Proceedings of the IEEE 34th International Joint Conference on Neural Networks, pp 1–8
30. Zhang Y, Luo X, Chen Q, Zou R, Zhang Y, Cheung Y (2024) Towards unbiased minimal cluster analysis of categorical-and-numerical attribute data. In: Proceedings of the 28th International Conference on Pattern Recognition, pages 1–16
31. Sangma JW, Sarkar M, Pal V, Agrawal A, Yogita (2022) Hierarchical clustering for multiple nominal data streams with evolving behaviour. Complex Intell Syst 8(2):1737–1761. https://doi.org/10.1007/s40747-021-00634-0
32. Zhao M, Feng S, Zhang Y, Mengkei L, Lu Y, Cheung Y (2024) Learning order forest for qualitative-attribute data clustering. In: Proceedings of the 27th European Conference on Artificial Intelligence, pages 1–8
33. Peng Z, Song X, Song S, Stojanovic V (2023) Hysteresis quantified control for switched reaction–diffusion systems and its application. Complex Intell Syst 9(6):7451–7460. https://doi.org/10.1007/s40747-023-01135-y
34. Sun P, Song X, Song S, Stojanovic V (2023) Composite adaptive finite-time fuzzy control for switched nonlinear systems with preassigned performance. Int J Adapt Control Signal Process 37(3):771–789. https://doi.org/10.1002/acs.3546
35. Zhang Z, Song X, Sun X, Stojanovic V (2023) Hybrid-driven-based fuzzy secure filtering for nonlinear parabolic partial differential equation systems with cyber attacks. Int J Adapt Control Signal Process 37(2):380–398. https://doi.org/10.1002/acs.3529
36. Abdel-Basset M, Abouhawwash M, Askar S, Hawash H, Nayyar A (2023) A system of the internet of nano-dental things
37. Shenghong C, Yiqun Z, Xiaopeng L, Yiu-ming C, Hong J, Peng L (2024) Robust categorical data clustering guided by multi-granular competitive learning. In: Proceedings of the IEEE 44th International Conference on Distributed Computing Systems, pp. 288–299
38. Zou R, Zhang Y, Zhang Y, Lu Y, Li M, Cheung Y (2024) Federated clustering with unknown number of clusters. In: Proceedings of the IEEE 6th International Conference on Data-driven Optimization of Complex Systems, pp. 671–677
39. Chen J, Ji Y, Zou R, Zhang Y, Cheung Y (2024) QGRL: quaternion graph representation learning for heterogeneous feature data clustering. In: Proceedings of the 30th SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1–10
40. Mohamed M (2023) Agricultural Sustainability in the Age of Deep Learning: Current Trends, Challenges, and Future Trajectories. Sustain Mach Intell J 4. https://doi.org/10.61185/SMIJ.2023.44102
41. Mohamed M (2023) Empowering deep learning based organizational decision making: a survey. Sustain Mach Intell J 3. https://doi.org/10.61185/SMIJ.2023.33105
42. AlNuaimi N, Masud MM, Serhani MA, Zaki N (2022) Streaming feature selection algorithms for big data: a survey. Appl Comput Inf 18(1/2):113–135. https://doi.org/10.1016/j.aci.2019.01.001
43. Eskandari S, Javidi MM (2016) Online streaming feature selection using rough sets. Int J Approx Reason 69:35–57. https://doi.org/10.1016/j.ijar.2015.11.006
44. Mohammad Masoud Javidi and Sadegh Eskandari (2019) Online streaming feature selection: a minimum redundancy, maximum significance approach. Pattern Anal Appl 22(3):949–963. https://doi.org/10.1007/s10044-018-0690-7
45. Saúl Solorio-Fernández J, Carrasco-Ochoa A, Martínez-Trinidad JF (2022) A survey on feature selection methods for mixed data. Artif Intell Rev 55(4):2821–2846. https://doi.org/10.1007/s10462-021-10072-6
46. Zhang P, Li T, Yuan Z, Luo C, Liu K, Yang X (2022) Heterogeneous Feature Selection Based on Neighborhood Combination Entropy. IEEE Trans Neural Netw Learn Syst, pp. 1–14. https://doi.org/10.1109/TNNLS.2022.3193929
47. Rubner Y, Tomasi C, Guibas LJ (2000) The Earth Mover's Distance as a Metric for Image Retrieval. Int J Comput Vision 40(2):99. https://doi.org/10.1023/A:1026543900054
48. Jia X, Bin Lei YG, Winslett M, Ge Yu, Zhang Z (2015) Efficient Similarity Join Based on Earth Mover's Distance Using MapReduce. IEEE Trans Knowl Data Eng 27(8):2148–2162. https://doi.org/10.1109/TKDE.2015.2411281
49. Zhang Y, Cheung Y-M (2023) Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data. IEEE Trans Neural Netw Learn Syst 34(9):6530–6544. https://doi.org/10.1109/TNNLS.2022.3202700
50. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Disc 2(3):283–304. https://doi.org/10.1023/A:1009769707641
51. He Xiaofei, Cai Deng, Niyogi Partha (2005) Laplacian Score for Feature Selection. In *Advances in Neural Information Processing Systems*, pages 507–514
52. Santos Jorge M, Mark Embrechts (2009) On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Proceedings of the 19th International conference on artificial neural networks, pages 175–184. https://doi.org/10.1007/978-3-642-04277-5_18
53. Rand WM (1971) Objective Criteria for the Evaluation of Clustering Methods. J Am Stat Assoc 66(336):846–850
54. Gates AJ, Ahn YY (2017) The impact of random models on clustering similarity. J Mach Learn Res 18:3049–3076. https://doi.org/10.5555/3122009.3176831
55. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.