Expert Systems With Applications 273 (2025) 126738

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Learning unified distance metric for heterogeneous attribute data clustering

Yiqun Zhang [a], Mingjie Zhao [a], Yizhou Chen [b], Yang Lu [b,*], Yiu-ming Cheung [c]

[a] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, 510006, Guangdong, China
[b] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, 361102, Fujian, China
[c] Department of Computer Science, Hong Kong Baptist University, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Mixed data clustering
Heterogeneous attribute
Distance structure reconstruction
Learnable weighting

ABSTRACT

Datasets composed of numerical and categorical attributes (also called mixed data hereinafter) are common in real clustering tasks. Differing from numerical attributes that indicate tendencies between two concepts (e.g., high and low temperature) with their values in well-defined Euclidean distance space, categorical attribute values are different concepts (e.g., different occupations) embedded in an implicit space. Simultaneously exploiting these two very different types of information is an unavoidable but challenging problem, and most advanced attempts either encode the heterogeneous numerical and categorical attributes into one type, or define a unified metric for them for mixed data clustering, leaving their inherent connection unrevealed. This paper, therefore, studies the connection among any-type of attributes and proposes a novel Heterogeneous Attribute Reconstruction and Representation (HARR) learning paradigm accordingly for cluster analysis. The paradigm transforms heterogeneous attributes into a homogeneous status for distance metric learning, and integrates the learning with clustering to automatically adapt the metric to different clustering tasks. Differing from most existing works that directly adopt defined distance metrics or learn attribute weights to search clusters in a subspace. We propose to project the values of each attribute into unified learnable multiple spaces to more finely represent and learn the distance metric for categorical data. HARR is parameter-free, convergence-guaranteed, and can more effectively self-adapt to different sought number of clusters k. Extensive experiments illustrate its superiority in terms of accuracy and efficiency.

1. Introduction

Mixed datasets comprising numerical and categorical attributes are widespread in real cluster analysis tasks (Jain, Murty, & Flynn, 1999). Differing from the numerical attributes with precisely quantified values, categorical attributes are usually with a certain number of qualitative values representing some fuzzy concepts (Agresti, 2003), which are embedded in an implicit distance space. Categorical attributes can be further distinguished into two sub-types, i.e., nominal and ordinal attributes, where ordinal attribute values are with additional intrinsic order information (Qian, Xu, Liang, Liu and Wang, 2015). The conventional taxonomy of attributes is shown in Fig. 1. As the distance spaces of categorical attributes are implicit, how to combine them with the explicit distance spaces of numerical attributes is a key issue for mixed data clustering (dos Santos & Zárate, 2015). Existing efforts can be roughly categorized into two streams: (i) Encode the qualitative categorical attribute values into numerical ones for clustering; (ii) Define a new hybrid distance measure or metric for clustering (Boriah, Chandola, & Kumar, 2008).

For the encoding-based clustering, the popular and simple one-hot encoding converts a possible value of a categorical attribute into a new numerical attribute, with "1" representing this possible value and "0" representing the others. In recent years, more informative encoding methods (Jian, Pang, Cao, Lu & Gao, 2018; Qian, Li, Liang, Liu and Dang, 2015) have been proposed for extracting and embedding more statistical information including occurrence frequencies of possible values, co-occurrence frequencies of possible values from dependent attributes, etc., to achieve better representations. Recently, the powerful encoding method (Zhu, Cao, & Yin, 2022) further introduces multiple kernels to comprehensively encode the value-level and attribute-level coupling relationships of categorical attributes. The above-mentioned encoding results are usually fed to the existing clustering algorithms, e.g., the k-means algorithm (Ball & Hall, 1967) and its variants (Fränti & Sieranoja, 2019; Huang, Ng, Rong, & Li, 2005; Jing, Ng, & Huang, 2007), for clustering. However, since they are designed for purely categorical data only, the relationship information between numerical and categorical attributes in mixed data will be

* Corresponding author.
E-mail addresses: yqzhang@gdut.edu.cn (Y. Zhang), zmingjie775@gmail.com (M. Zhao), imboaba@gmail.com (Y. Chen), luyang@xmu.edu.cn (Y. Lu), ymc@comp.hkbu.edu.hk (Y.-m. Cheung).

https://doi.org/10.1016/j.eswa.2025.126738
Received 3 September 2024; Received in revised form 24 January 2025; Accepted 30 January 2025
Available online 8 February 2025
0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

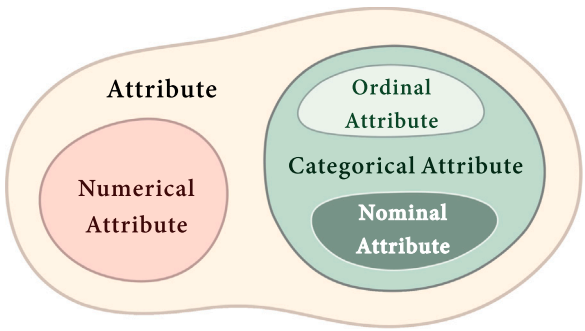


Fig. 1. Conventional taxonomy of attributes, which shows the traditional classification of attributes into Numerical and Categorical types. Categorical Attributes are further divided into Ordinal Attributes, with inherent order (e.g., small, medium, large), and Nominal Attributes, without natural order (e.g., colors or labels). This taxonomy emphasizes the distinct characteristics of each attribute type.

somewhat omitted by using them.

For the approaches based on dissimilarity defining, Hamming distance metric (Arabie, Baier, Critchley, & Keynes, 2006) that only simply distinguishes the distance between identical and different categorical values is the most common. To more finely distinguish dissimilarity degrees of different values, some measures (Goodall, 1966; Lin, 1998; Zhang, Cheung, & Tan, 2020) have been proposed based on the occurrence probabilities of possible values. There are also measures (Ienco, Pensa, & Meo, 2012; Jia, Cheung, & Liu, 2016; Jian, Cao, Lu & Gao, 2018; Zhang & Cheung, 2022b) that take the interdependence of categorical attributes into account further. Most recently, more advanced dissimilarity learning methods (Zhang & Cheung, 2020, 2022a) have been proposed, which not only reasonably define the dissimilarities, but also enable the adaptive learning of dissimilarities during clustering. As all the above-mentioned solutions are designed for categorical attributes only, the defined dissimilarities are usually combined with Euclidean distances of numerical attributes by the k-prototypes algorithm (Huang, 1997) for mixed data clustering. In the literature, some measures (Cheung & Jia, 2013; Gower, 1971; Jia & Cheung, 2018) also attempt to directly quantify the object-cluster dissimilarities for mixed data.

Although existing works provide a variety of solutions, they may still suffer from at least one of the following two intractable issues: (i) Adaptability: As the encoding or dissimilarity defining phase is independent of the clustering phase, the encoded data or the dissimilarities may not suit certain clustering tasks well. How to adapt the representation or the defined dissimilarities of mixed data to the clustering task is a key issue to enhance the scalability and accuracy of the clustering approach; (ii) Homogeneity: Numerical attribute values finely grade the tendencies between two concepts (e.g., {high, low} of attribute "temperature"), while categorical attribute values usually represent multiple concepts (e.g., {driver, lawyer, nurse, . . . } of attribute "occupation") in a coarse-grained manner. To tackle these issues, a mixed data clustering method (Zhang, Cheung, & Zeng, 2022) has been proposed, which first reconstructs categorical attributes into a form similar to that of numerical attributes, and then uniformly learns the weights of all the attributes. However, it overlooks the potential ordinal relationship of categorical attribute values and only learns the importance of each whole attribute without finely exploring the distance structure among attribute values.

This paper, therefore, studies the differences and connections among numerical, nominal, and ordinal attributes. For all heterogeneous attributes to provide comparable distances for clustering, we reconstruct each categorical attribute that is originally arranged in a nonlinear space into a set of possible one-dimensional spaces that are homogeneous to the distance space of the numerical attribute. In comparison with the mainstream approaches that rely on external knowledge or hypothesized distributions for categorical attribute distance space representation like Gower's distance (Gower, 1971), our reconstruction is only based on the basic data statistics to avoid priori bias. It turns out that a categorical attribute can be adequately represented from different views of its value pairs, and represented attributes are with distance spaces like that of numerical attributes. Such design provides a homogeneous basis for jointly considering heterogeneous attributes, and we further arrange learnable weights to the distances among reconstructed categorical values to leverage adaptive learning w.r.t. clustering. The main contributions of this paper are four-fold:

- Connection among numerical, nominal, and ordinal attributes is revealed from a new perspective of the intrinsic semantic concepts represented by the attribute values, which offers the potential to jointly understand the heterogeneous attributes in mixed datasets.
- A projection-based attribute reconstruction method is proposed to informatively convert the heterogeneous distance spaces of different types of attributes to be homogeneous, which provides an effective data basis for the processing and analysis of mixed data.
- The adaptation of reconstructed representation to the data objects clustering is designed into a learnable task. As all types of attributes can be homogeneously and interactively represented by learning, such clustering paradigm is suitable for any-type-attributed data.[1]
- Two learning algorithms are instantiated from the learning paradigm, which circumvents the hyper-parameters for learning, and searches clusters in attribute subspaces to improve the degree of learning freedom, respectively.

The rest of this paper is organized as follows. Section 2 provides an overview of the related works. In Section 3, the proposed representation method and the learning algorithms are presented. Section 4 presents theoretical analyses. Section 5 conducts experimental evaluation. Section 6 draws a conclusion of this paper.

2. Related work

This section provides an overview of the closely related research, i.e., data encoding approaches and dissimilarity measures, for the clustering of categorical and mixed data.

Encoding approaches have two common procedures: (1) encode the target dataset according to a certain strategy, (2) conduct cluster analysis by treating the encoded dataset as numerical-valued data. In this stream, one-hot encoding is the simplest but commonly used approach, which converts the sth possible value o_s^r of attribute a^r into a v^r-bit vector with the sth bit set at "1" and the remainders set at "0". Here, v^r indicates the total number of possible values of attribute a^r. Since such encoding process is equivalent to uniformly assigning distance "1" to any pair of different values, one-hot encoding is infeasible to produce different dissimilarity degrees (Boriah et al., 2008; dos Santos & Zárate, 2015), and may thus cause information loss. Space structure-based encoding proposed in Qian, Li et al. (2015) encodes a target data object $\mathbf{x}_i$ by concatenating the distances from $\mathbf{x}_i$ to all the n data objects into an n-dimensional vector. Later, coupling-based representation (Jian, Cao, Pang, Lu, & Gao, 2017; Jian, Pang et al., 2018) has been proposed considering the relationship among categorical attributes for encoding. Most recently, a more advanced method (Zhu et al., 2022) adopting different kernels for more comprehensive relationship representation has been proposed. As all the above-mentioned approaches are designed for categorical attributes

[1] Any-type-attributed data is a dataset composed of any combination of numerical, nominal, and ordinal attributes, which is also called any-type data interchangeably in this paper.

only, some approaches (Jian, Hu, Cao & Lu, 2018; Zhu, Zhang, Cao, & Abrahamyan, 2020) have also been designed to simultaneously encode the numerical and categorical attributes with considering their dependency. However, sensitivity to the hyper-parameters may still influence their performance, and the potential order relationship information of categorical attributes remains to be exploited.

For the approaches that directly define the dissimilarities, Hamming distance is the most popular metric. It always assigns distance "1" to two different values and set distance at "0" for identical values. Similar to the one-hot encoding, the boolean distance values prevent it from distinguishing different dissimilarity degrees. Accordingly, similarity measures based on probability (Cheung & Jia, 2013; Goodall, 1966; Li et al., 2023) have been proposed, which can more finely define similarities according to the occurrence frequencies of the possible values. Entropy-based similarity measures (Khan, Mishra, & Mohanty, 2023; Lin, 1998; Zhang & Cheung, 2018; Zhang et al., 2020) proposed to consider distance measurement from the aspect of information theory (Kullback, 1997). They also compute similarities based on the occurrence frequencies of possible values, which take a common idea that two more dissimilar values may result in a higher information amount reflected by a larger entropy value. Since they all treat attributes independently and ignore the valuable information provided by the inter-attribute dependence (dos Santos & Zárate, 2015), some measures Ahmad and Dey (2007), Ienco, Pensa, and Meo (2009), Ienco et al. (2012) and Le and Ho (2005) have been proposed to define similarities between possible values according to their conditional probability distributions obtained from the relevant attributes. Among them, Ahmad's measure (Ahmad & Dey, 2007) and the association-based measure (Le & Ho, 2005) adopt a similar idea that the similarity between two possible values from an attribute can be reflected by their corresponding conditional probability distributions on the other attributes. Context-based measure (Ienco et al., 2009, 2012) further filters out the irrelevant attributes. However, since they rely on the sole interdependence of attributes, they may fail in distance measurement when all the attributes are independent of each other (Jia et al., 2016; Yu, Fu, Cheng, & Ye, 2022). To address such issue, a series of distance measures proposed in Jia and Cheung (2018), Jia et al. (2016), Jian, Cao et al. (2018) and Zhang and Cheung (2022b, 2023) consider both the intra- and inter-attribute statistical information. Some advanced clustering algorithms with attributes weighting mechanism (Bai, Liang, Dang, & Cao, 2011; Chen, Wang, Wang, & Zhu, 2016) have also been proposed to learn the importance of categorical attributes during clustering.

Since encoding and dissimilarity defining operations are usually performed separately from clustering, the encoding results or the defined dissimilarities cannot adapt well to different clustering tasks. Accordingly, some recently proposed methods (Zhang & Cheung, 2020, 2022a; Zhao et al., 2024) have been proposed to first define the similarities, and then iteratively learn the similarities and data partitions. As a result, this type of method achieves competitive clustering accuracy compared to the existing clustering approaches. However, the method proposed in Zhang and Cheung (2020) only considers pure ordinal data while the methods presented in Zhang and Cheung (2022a) and Zhao et al. (2024) are only designed for mixed data composed of ordinal and nominal attributes. Most recently, a new clustering approach (Hu, Jiang, Liu, & He, 2022) has been proposed to explore significant clusters of categorical data without requiring a known true number of clusters.

3. Proposed method

We first provide preliminaries and formulate the problem, then we introduce the attribute representation method and the learning algorithms in detail. Frequently used symbols in this paper and the corresponding explanations are sorted out in Table 1.

3.1. Preliminaries

Given a mixed dataset X containing n data objects $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. Each data object $\mathbf{x}_i = [x_i^1, x_i^2, \dots, x_i^d]^\top$ can be viewed as a d-dimensional vector with values from d attributes $A = \{a^1, a^2, \dots, a^d\}$, including d_u numerical attributes and d_c categorical attributes. For simplicity, assuming that the former d_u attributes in A are numerical ones denoted as an attribute set A_u, and the latter d_c are categorical attributes denoted as A_c. Thus we have $A_u = \{a^1, a^2, \dots, a^{d_u}\}$ and $A_c = \{a^{d_u+1}, a^{d_u+2}, \dots, a^d\}$. Note that the superscript denotes the index of attribute throughout the paper. Accordingly, we have $A = A_u \cup A_c$, and $d = d_u + d_c$. The v^r possible values of each categorical attribute $a^r \in A_c$ are represented as a set $O^r = \{o_1^r, o_2^r, \dots, o_{v^r}^r\}$.

Since categorical attributes may also include ordinal attributes, we assume the former d_n attributes in A_c are nominal, and the latter d_o are ordinal. Accordingly, we have $A_c = A_n \cup A_o$ and $d_c = d_n + d_o$. The sequential numbers $\{1, 2, \dots, v^r\}$ further reflect the rank of possible values $O^r = \{o_1^r, o_2^r, \dots, o_{v^r}^r\}$ of an ordinal attribute a^r. For dataset without ordinal attribute, it is possible that $d_o = 0$ and $d_c = d_n$.

A general goal of clustering is to partition the n data objects into k compact clusters $C = \{c_1, c_2, \dots, c_k\}$, where the intra-cluster objects are similar and inter-cluster objects are distinct. An $n \times k$ matrix $\mathbf{Q}$ is usually maintained during clustering indicating the affiliations between objects and clusters. More specifically, the value of the (i, l)th entry of $\mathbf{Q}$ can be obtained by

$$q_{il} = \begin{cases} 1 & , \quad \text{if } l = \arg\min_y \Phi(\mathbf{x}_i, \mathbf{m}_y) \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{1}$$

As we focus on the crisp partitional clustering, q_{il} should satisfy $\sum_{l=1}^k q_{il} = 1$ and $q_{il} \in \{0, 1\}$. $\Phi(\mathbf{x}_i, \mathbf{m}_l)$ denotes the dissimilarity between object $\mathbf{x}_i$ and cluster c_l. We describe a cluster c_l using a vector $\mathbf{m}_l = [m_l^1, m_l^2, \dots, m_l^d]^\top$ called prototype, which is computed following the conventional k-prototypes algorithm (Huang, 1997):

$$m_l^r = \begin{cases} \frac{1}{\sigma(c_l)} \sum_{i=1}^n q_{il} x_i^r & , \quad \text{if } r \in \{1, 2, \dots, d_u\} \\ o_h^r & , \quad \text{if } r \in \{d_u + 1, d_u + 2, \dots, d\}, \end{cases} \tag{2}$$

where $\sigma(\cdot)$ counts the number of elements in a set, and thus $\sigma(c_l)$ counts the number of objects assigned to c_l. o_h^r is obtained by

$$h = \arg\max_u \sum_{i=1}^n q_{il} \delta(x_i^r = o_u^r), \tag{3}$$

where $\delta(x_i^r = o_u^r)$ is a logical function, which returns value "1" if $x_i^r = o_u^r$ is true, otherwise, returns "0". The object-cluster dissimilarity $\Phi(\mathbf{x}_i, \mathbf{m}_l)$ in Eq. (1) can be written as

$$\Phi(\mathbf{x}_i, \mathbf{m}_l) = \sum_{r=1}^d \phi(x_i^r, m_l^r), \tag{4}$$

where the inner-sum $\phi(x_i^r, m_l^r)$ is the dissimilarity between the rth values of $\mathbf{x}_i$ and $\mathbf{m}_l$.

Since numerical, nominal, and ordinal attributes are with different distance space structures, they may provide heterogeneous information in terms of clustering. Thus, how to homogeneously represent them to eliminate their awkward information gap is the key problem that should be solved for mixed data clustering.

3.2. Homogeneous attribute representation

To eliminate the awkward information gap between heterogeneous attributes, we need to represent them homogeneously. From the perspective of concepts described by the values of an attribute, the values of a numerical attribute describe tendencies between two (usually opposite) concepts (e.g., {high, low} of attribute "temperature") while the values of a categorical attribute usually describe multiple concepts (e.g., {driver, lawyer, nurse, police} of an attribute "occupation"). The ordinal attributes are also common among categorical attributes, where

Table 1
Frequently used symbols. Note that we uniformly use lowercase, uppercase, bold lowercase, and bold uppercase to indicate value, set, vector, and matrix, respectively.

Symbol	Explanation	Symbol	Explanation
X, A, O, M, and C	Dataset, attribute set, possible value set, cluster prototypes set, and cluster set	$\mathbf{x}_i$ and x_i^r	ith data object and rth value of $\mathbf{x}_i$
a^r, v^r	rth attribute, possible value number of a^r	A_u and A_c	Numerical and categorical attribute set
A_n and A_o	Nominal and ordinal attribute set of categorical attributes, $A_c = A_n \cup A_o$	d, k and n	Number of attributes($d = d_u + d_c$), clusters, and data objects
d_c and d_u	Number of categorical attributes ($d_c = d_n + d_o$), and numerical attributes	o_h^r	hth possible value of a^r
$\mathbf{Q}$ and c_l	Object-cluster affiliation matrix and lth cluster	q_{il}	A value indicating the affiliation between $\mathbf{x}_i$ and c_l
$\mathbf{m}_l$	A vector describing data objects of c_l	γ^r	Number of endogenous spaces corresponding to a^r
$\mathcal{R}^r$ and $\mathcal{R}_b^r$	Endogenous space set of a^r and its bth endogenous space	$\Phi(\cdot,\cdot)$ and $\phi(\cdot,\cdot)$	Data object-level dissimilarity and value-level distance
$\mathbf{W}$ and w^r	Attribute weight set and weight indicating the importance of a^r	$\kappa(\cdot,\cdot)$	Base distance

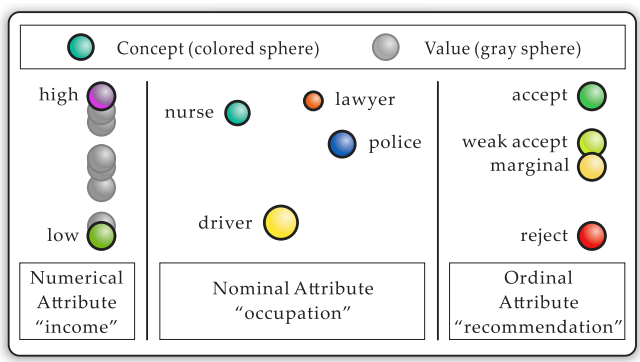


Fig. 2. Comparison of the "concepts" of numerical attribute, nominal attribute, and ordinal attribute, which illustrates how "concepts" and "values" are represented across different attribute types. Numerical attributes, like "income", maintain a continuous order (e.g., high to low). Nominal attributes, such as "occupation", consist of distinct categories without inherent order (e.g., nurse, lawyer, driver). Ordinal attributes, like "recommendation", contain ordered categories (e.g., accept, weak accept, marginal, reject), highlighting their hierarchical structure.

multiple ordered concepts (e.g., {accept, weak accept, marginal, reject}) describe tendencies between two concepts (e.g., {accept, reject} of attribute "recommendation"). Fig. 2 shows the concepts of the three types of attributes.

More specifically, the values of a numerical attribute $a^r \in A_u$ are linearly arranged between two conceptual values o_{min}^r and o_{max}^r in a one-dimensional space, where $o_{min}^r = 0$ and $o_{max}^r = 1$ if the values are normalized into the interval $[0, 1]$. By contrast, the values of a nominal attribute $a^r \in A_c$ can be viewed as concentrating on v^r non-linearly arranged conceptual values in O^r. The case between numerical and nominal attributes is the ordinal attribute, where the attribute values concentrate on v^r linearly arranged conceptual values in O^r. Numerical attribute is with well-defined Euclidean distance space, which features easy-to-represent and computational convenience. Therefore, how to represent a categorical attribute in one-dimensional space like numerical attribute but still preserve the original information of the attribute is the goal of our representation.

The information carried by categorical attributes refers to their multidimensional distance structure, which is rich in information. However, when converted into a one-dimensional Euclidean distance structure, this can easily lead to a loss of the original information. To preserve this information, we use a multiple spaces projection approach, which retains the rich structure while enabling the conversion into a one-dimensional distance structure similar to numerical attributes. This allows categorical attributes to participate seamlessly in arithmetic operations during clustering. However, when One-Hot Encoding (OHE) is used, it imposes an equidistant constraint on the distance structure, preventing the original information of categorical attributes from being fully preserved.

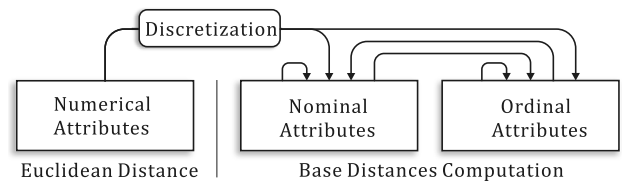


Fig. 3. Relationships among different types of attributes in base distances computation. "A→B" means that base distances of Type-B attributes are computed with the contribution of Type-A attributes.

For a categorical attribute $a^r \in A_c$, distance between any pair of values o_g^r and o_h^r can be informatively measured based on dataset statistics by

$$\kappa(o_g^r, o_h^r) = \sum_{s=1}^{d} \sum_{j=1}^{v^s} |p(o_j^s|o_g^r) - p(o_j^s|o_h^r)|, \tag{5}$$

where the term $p(o_j^s|o_g^r)$ is actually the occurrence probability of the value o_j^s as given o_g^r, which can be written as

$$p(o_j^s|o_g^r) = \frac{\sum_{u=1}^{n} \delta(x_u^s = o_j^s \wedge x_u^r = o_g^r)}{\sum_{i=1}^{n} \delta(x_i^r = o_g^r)}. \tag{6}$$

Eq. (5) measures the overall difference between two Conditional Probability Distributions (CPDs) obtained from attribute a^s as given two possible values o_g^r and o_h^r from a^r. Distance defined in this form has been commonly adopted by the recent works (Jian, Cao et al., 2018; Zhang & Cheung, 2022a) proposed for categorical data, and is generalized to HOmogeneous Distance (HOD) (Zhang & Cheung, 2022a) under the case that the dataset is composed of both nominal and ordinal attributes. To further exploit the information offered by numerical attributes, we discrete numerical attributes and treat them as ordinal attributes to reflect distances of categorical attributes. This distance is called base distance as the following projection-based representation will be implemented based on these distances. Fig. 3 shows the reflection relationship between different types of attributes for computing $\kappa(o_g^r, o_h^r)$.

However, distance spaces of categorical attributes yielded by HOD are not guaranteed to be one-dimensional spaces like that of numerical attributes, see Fig. 2. To represent the heterogeneous space structures of numerical and categorical attributes in a homogeneous way, we dismantle the space structure of a categorical attribute a^r by projecting all its n attribute values onto each of the one-dimensional spaces $\mathcal{R}^r = \{\mathcal{R}^{r,1}, \mathcal{R}^{r,2}, \ldots, \mathcal{R}^{r,\gamma^r}\}$ spanned by γ^r pairs of conceptual values (i.e. possible values) where $\gamma^r = v^r(v^r - 1)/2$. The projection is implemented by geometrically projecting the n values onto each space based on the base distance $\kappa(o_g^r, o_h^r)$. Specifically, projection point of a value o_t^r on a space $\mathcal{R}^{r,b}$ ($b \in \{1, 2, \ldots, \gamma^r\}$) spanned by o_g^r and o_h^r can be described by the distance between o_t^r and o_g^r, and also interchangeably by the distance

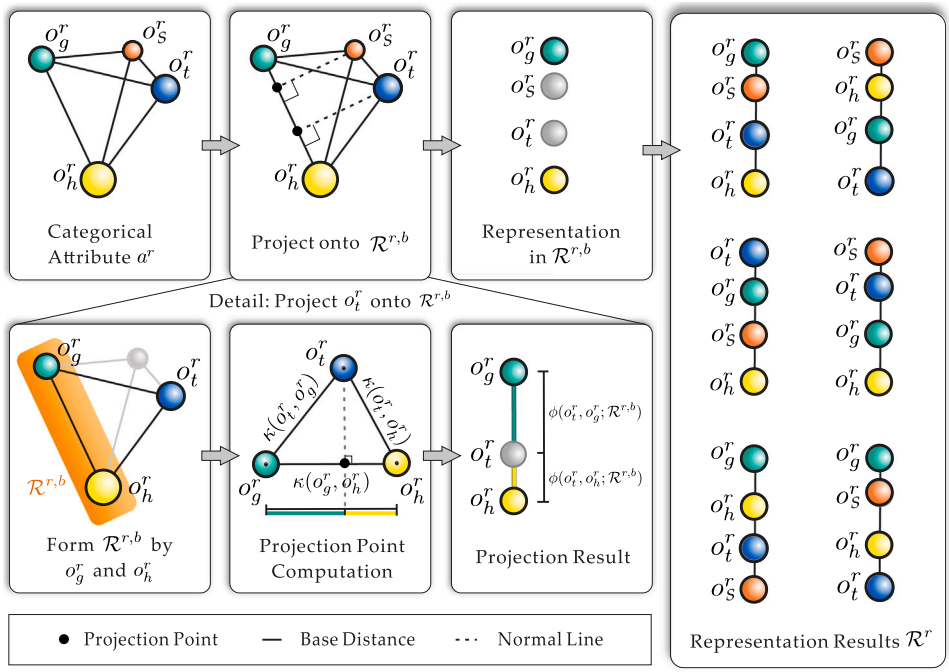


Fig. 4. Projection processes of a categorical attribute a^r. An attribute a^r with 4 possible values is expanded into 6 distinct attributes after the projection process. Each projected attribute represents unique information, highlighting the rich information content of the original attribute.

between o_t^r and o_h^r, which can be computed by:

$$\phi(o_t^r, o_g^r; \mathcal{R}^{r,b}) = \frac{|\kappa(o_t^r, o_g^r)^2 - \kappa(o_t^r, o_h^r)^2 + \kappa(o_g^r, o_h^r)^2|}{2\kappa(o_g^r, o_h^r)} \tag{7}$$

and

$$\phi(o_t^r, o_h^r; \mathcal{R}^{r,b}) = \frac{|\kappa(o_t^r, o_h^r)^2 - \kappa(o_t^r, o_g^r)^2 + \kappa(o_g^r, o_h^r)^2|}{2\kappa(o_g^r, o_h^r)}, \tag{8}$$

respectively. Eqs. (7) and (8) are derived by directly applying the Pythagorean theorem. A more intuitive demonstration of such a process of a categorical attribute a^r is shown in Fig. 4. An attribute a^r with 4 possible values is expanded into 6 distinct attributes after the projection process. Each projected attribute represents unique information, highlighting the rich information content of the original attribute, which naturally justifies the expansion factor. This scaling is determined by the number of possible values for the categorical attribute. Therefore, dividing its contribution equally among the 6 projected attributes would dilute its inherent richness and is not appropriate.

The projection for ordinal attributes is a special case for the projection of categorical attributes. Since we assume that the ordinal values are arranged linearly as shown in Fig. 2, the γ^r one-dimensional spaces spanned by the conceptual value pairs are exactly overlapped, and so do the corresponding projection results. Thus, only one one-dimensional space $\mathcal{R}^r = \{\mathcal{R}^{r,1}\}$ where $\gamma^r = 1$ will be enough for representing an ordinal attribute a^r, and the projection point of a value o_t^r can be directly reflected by the based distances:

$$\phi(o_t^r, o_g^r; \mathcal{R}^{r,b}) = \kappa(o_t^r, o_g^r) \tag{9}$$

and

$$\phi(o_t^r, o_h^r; \mathcal{R}^{r,b}) = \kappa(o_t^r, o_h^r) \tag{10}$$

without loss of the generality for both the two types of categorical attributes.

Given a certain space $\mathcal{R}^{r,b}$ spanned by o_g^r and o_h^r, all the n values of a^r are linearly arranged after the projection, and the distance between any two values o_u^r and o_f^r represented in $\mathcal{R}^{r,b}$ can be written based on Eqs. (7)–(10) as

$$\begin{aligned}\phi(o_u^r, o_f^r; \mathcal{R}^{r,b}) &= |\phi(o_u^r, o_g^r; \mathcal{R}^{r,b}) - \phi(o_f^r, o_g^r; \mathcal{R}^{r,b})| \\ &= |\phi(o_u^r, o_h^r; \mathcal{R}^{r,b}) - \phi(o_f^r, o_h^r; \mathcal{R}^{r,b})|.\end{aligned} \tag{11}$$

The projection of a^r actually generates γ^r sub-attributes $A^r = \{a_1^r, a_2^r, \ldots, a_{\gamma^r}^r\}$ with common possible values O^r represented in γ^r spaces $\mathcal{R}^r$. Note that A^r with superscript r indicates sub-attributes obtained after the representation of original attribute a^r, while A_c with subscript c indicates that the attribute type is categorical. After representing all the d_c categorical attributes, attribute set A is updated to

$$\hat{A} = A_u \cup \hat{A}_c, \tag{12}$$

where

$$\hat{A}_c = A^{d_u+1} \cup A^{d_u+2} \cup \cdots \cup A^d. \tag{13}$$

The symbol "hat" denotes that an original set or quantity is updated after the representation, see $\hat{A}$ and $\hat{A}_c$. Accordingly, we have $\hat{d} = d_u + \hat{d}_c$, where $\hat{d}_c = \sum_{r=d_u+1}^{d} \gamma^r$.

Based on the updated attribute set $\hat{A}$, we divide all distances of each $a^r \in \hat{A}_c$ by the maximum distance of a^r to make the distances comparable to the distances of normalized numerical attributes in A_u. The obtained distances $\phi(o_u^r, o_f^r; \mathcal{R}^{r,b})$ can be interchangeably denoted as $\phi(o_u^r, o_f^r)$ for simplicity without causing ambiguity. $\phi(o_u^r, o_f^r)$ calculates the absolute value of the dissimilarity between the encoded coordinates of o_u^r and o_f^r (see the "Projection Result" part of Fig. 4), which is equivalent to calculating the absolute value of the distance between two values on a numerical attribute. Because the differences among variables (attributes) in mixed data is often large, we use the Manhattan distance for Eq. (4).

An intuitive comparison of attributes represented by the proposed method and the conventional methods is shown in Fig. 5. The representation of Hamming distance is significantly different from the structure of the numerical attribute, and all the different values are with identical distances. One-hot encoding has the additional problem of making different values indistinguishable. It is obvious that our representation is more informative, and provides a homogeneous basis for distance computation on heterogeneous attributes. To obtain more clustering-friendly heterogeneous attribute representations, the key lies in how to integrate the representation process with clustering through joint learning.

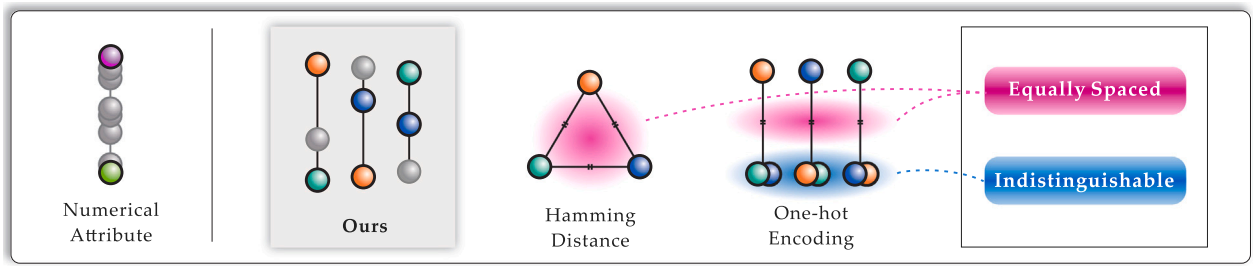


Fig. 5. Comparison of Hamming distance, one-hot encoding, and categorical attribute representation obtained through our method. It is obvious that our representation is more informative, and provides a homogeneous basis for distance computation on heterogeneous attributes.

3.3. Learning algorithms

To facilitate the learning of representations in clustering, the distance $\Phi(\mathbf{x}_i, \mathbf{m}_l)$ in Eq. (4) is re-written as the linear combination of the dissimilarities between x_i^r and m_l^r reflected by different attributes in $\hat{A}$:

$$\Phi_w(\mathbf{x}_i, \mathbf{m}_l) = \sum_{r=1}^{\hat{d}} \phi(x_i^r, m_l^r) \cdot w^r, \tag{14}$$

where w^r controls the contribution of a^r, $\mathbf{w} = [w^1, w^2, \dots, w^{\hat{d}}]^\top$, and $\sum_{r=1}^{\hat{d}} w^r = 1$. To learn the representations while adapting to the clustering task, data objects partition $\mathbf{Q}$, cluster prototypes $M = \{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_k\}$, and the weights of attributes $\mathbf{w}$ are iteratively updated in the following processes: (1) Let $\mathbf{w}$ and M be fixed, compute $\mathbf{Q}$; (2) Let $\mathbf{w}$ and $\mathbf{Q}$ be fixed, compute M; (3) Iteratively implement (1) and (2) until convergence, then fix $\mathbf{Q}$ and M, and update $\mathbf{w}$. The above process follows the common learning steps for clustering with attributes weighting (Huang et al., 2005), and the updating strategy of $\mathbf{Q}$ and M have been presented by Eqs. (1)–(3) in Section 3.1.

The main difficulty lies in the updating of $\mathbf{w}$. As mentioned in Section 3.2, we represent an original attribute into a set of same-source sub-attributes A^t with common possible values O^t described by different distance structures computed by Eqs. (7)–(11), the attributes in A^t can thus be considered to be more dependent. If we directly compute $\mathbf{w}$ using the Lagrangian multiplier method or the gradient-decent method in the step (3), the weights of the same-source sub-attributes with similar represented distance structures will over-enhance each other in the next step (1). More specifically, suppose $\hat{a}^r$ and $\hat{a}^s$ are the same-source attributes in A^t, if the total intra-cluster distance contributed by a^r and a^s, i.e., $D^r = \sum_{i=1}^n \sum_{l=1}^k q_{il}\phi(x_i^r, m_l^r)$ and $D^s = \sum_{i=1}^n \sum_{l=1}^k q_{il}\phi(x_i^s, m_l^s)$, are both relatively low and are close to each other, then the effects of the weights w^r and w^s computed according to D^r and D^s, respectively, will over-enhance each other in the next step (1) because they may have similar effects in forming clusters. Such a synergy effect can easily cause a corrupt solution.

Accordingly, we adopt a novel weight updating strategy that more comprehensively considers the intra-cluster compactness and inter-cluster separation contributed by attributes to obtain discriminative quantification of the attribute weights. Specifically, a weight w^r is computed by

$$w^r = \frac{I^r}{\sum_{s=1}^{\hat{d}} I^s}, \tag{15}$$

where I^r is defined as

$$I^r = \frac{S^r}{D^r}. \tag{16}$$

We adopt average intra-cluster distance

$$D^r = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il}\phi(x_i^r, m_l^r) \tag{17}$$

to indicate the importance of a^r in terms of the intra-cluster compactness, and adopt average inter-cluster distance

$$S^r = \frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{l=1}^{k} (1 - q_{il})\phi(x_i^r, m_l^r) \tag{18}$$

to quantify the importance of a^r in terms of the inter-cluster separation. We compute average intra- and inter-cluster distance to make D^r and S^r comparable, and a more important attribute a^r should have a smaller D^r and a larger S^r. Moreover, $\frac{1}{n}$ in Eq. (17) and $\frac{1}{n(k-1)}$ in Eq. (18) can be simultaneously omitted, as their division $\frac{1}{n(k-1)} / \frac{1}{n}$ is a constant in Eq. (16). The proposed learning algorithm is summarized in Algorithm 1.

Algorithm 1 **H**eterogeneous **A**ttribute **R**econstruction and **R**epresentation learning (HARR)

Input: Dataset X, number of clusters k.

Output: Data partition $\mathbf{Q}$ and attribute weights $\mathbf{w}$.

Step 1: Represent all the categorical attributes according to Eqs. (7) - (11), obtain $\hat{A}$;

Step 2: Initialize $\mathbf{Q}'$ and $\mathbf{Q}''$ by setting all their values to 0 (for judging the convergence); Initialize M with randomly selected k objects; Initialize $\mathbf{w}$ by $w^r = 1/\hat{d}$;

Step 3: Fix $\mathbf{w}$ and M, compute $\mathbf{Q}$ by Eq. (1). If $\mathbf{Q} \neq \mathbf{Q}'$, set $\mathbf{Q}' = \mathbf{Q}$, and turn to Step 4; Otherwise, turn to Step 5;

Step 4: Fix $\mathbf{w}$ and $\mathbf{Q}$, compute M by Eq. (2), go to Step 3;

Step 5: If $\mathbf{Q} \neq \mathbf{Q}''$, set $\mathbf{Q}'' = \mathbf{Q}$, fix $\mathbf{Q}$ and M, compute $\mathbf{w}$ according to Eqs. (15)-(18), and turn to Step 3; Otherwise, stop and output $\mathbf{Q}$ and $\mathbf{w}$.

We focus on the learning of linear combinations of represented attributes. It would waste useful learning evidence if we directly use the contribution I^r of each attribute a^r on all the clusters measured by Eq. (16), but ignore the specific coupling relation of each attribute-cluster pair. Therefore, we further propose a more advanced learning strategy to adequately exploit the learning evidence provided by the distances caused by the attributes on different clusters. We expand the original weight vector $\mathbf{w}$ into a $k \times \hat{d}$ matrix $\mathbf{W}$ with its (l, r)th entry stores the weight w_l^r representing the importance of a^r in forming cluster c_l. Following the updating strategy of $\mathbf{w}$ described by Eqs. (15)–(18), we quantify the importance of w_l^r by

$$w_l^r = \frac{I_l^r}{\sum_{s=1}^{\hat{d}} I_l^s}, \tag{19}$$

where I_l^r quantifies the contribution of a^r in making the objects of cluster c_l more compact, and more distinct from the objects of the other clusters, which is defined as

$$I_l^r = \frac{S_l^r}{D_l^r}. \tag{20}$$

We adopt average intra-cluster distance on cluster c_l

$$D_l^r = \frac{1}{\sigma(c_l)} \sum_{i=1}^{n} q_{il} \phi(x_i^r, m_l^r) \tag{21}$$

to indicate the importance of a^r in compacting the $\sigma(c_l)$ objects of c_l. We adopt average distance formed between c_l and objects outside c_l

$$S_l^r = \frac{1}{n - \sigma(c_l)} \sum_{i=1}^{n} (1 - q_{il}) \phi(x_i^r, m_l^r) \tag{22}$$

to quantify the importance of a^r in separating the $\sigma(c_l)$ objects in c_l from the rest $n - \sigma(c_l)$ objects. The corresponding algorithm is summarized as Algorithm 2. As Algorithm 1 updates a weight **V**ector **w**, while Algorithm 2 updates a weight **M**atrix **W**, these two algorithms are denoted as HARR-V and HARR-M, respectively, to distinguish.

Algorithm 2 HARR to update weight **M**atrix (HARR-M)

Input: Dataset X, number of clusters k.

Output: Data partition indicator matrix **Q** and attribute weights **W**.

Step 1: Represent all the categorical attributes according to Eqs. (7) - (11), obtain $\hat{A}$;

Step 2: Initialize $\mathbf{Q}'$ and $\mathbf{Q}''$ by setting all their values to 0 (for judging the convergence); Initialize M with randomly selected k objects; Initialize **W** by $w_l^r = 1/\hat{d}$;

Step 3: Fix **W** and M, compute **Q** by Eq. (1). If $\mathbf{Q} \neq \mathbf{Q}'$, set $\mathbf{Q}' = \mathbf{Q}$, and turn to Step 4; Otherwise, turn to Step 5;

Step 4: Fix **W** and **Q**, compute M by Eq. (2), go to Step 3;

Step 5: If $\mathbf{Q} \neq \mathbf{Q}''$, set $\mathbf{Q}'' = \mathbf{Q}$, fix **Q** and M, compute **W** according to Eqs. (19)-(22), and turn to Step 3; Otherwise, stop the algorithm, output **Q** and **W**.

4. Theoretical analysis

Theorem 1. *Distance measure yielded by the proposed projection-based representation is a distance metric.*

Proof. Since all the values of a categorical attribute are projected onto one-dimensional Euclidean spaces, the value-level distances computed by Eq. (11) satisfy:

1. $\phi(o_g^r, o_f^r) \geq 0$
2. $o_g^r = o_f^r \Leftrightarrow \phi(o_g^r, o_f^r) = 0$
3. $\phi(o_g^r, o_f^r) = \phi(o_f^r, o_g^r)$
4. $\phi(o_g^r, o_f^r) \leq \phi(o_g^r, o_t^r) + \phi(o_t^r, o_f^r)$

for any $g, f, t \in \{1, 2, \ldots, v^r\}$ and $r \in \{d_u + 1, d_u + 2, \ldots, \hat{d}\}$. The object-level distance can be written based on the value-level distance as $\Phi(\mathbf{x}_a, \mathbf{x}_b) = \sum_{r=1}^{\hat{d}} \phi(x_a^r, x_b^r)$ according to Eq. (4), which can be generalized to $\Phi_w(\mathbf{x}_a, \mathbf{x}_b) = \sum_{r=1}^{\hat{d}} \phi(x_a^r, x_b^r) w^r$ where w^r ($w^r \geq 0$) is the weight corresponding to the rth represented attribute. According to the metric properties of value-level distances, it is clear that the object-level distances satisfy:

1. $\Phi_w(\mathbf{x}_a, \mathbf{x}_b) \geq 0$
2. $\mathbf{x}_a = \mathbf{x}_b \Leftrightarrow \Phi_w(\mathbf{x}_a, \mathbf{x}_b) = 0$
3. $\Phi_w(\mathbf{x}_a, \mathbf{x}_b) = \Phi_w(\mathbf{x}_b, \mathbf{x}_a)$
4. $\Phi_w(\mathbf{x}_a, \mathbf{x}_b) \leq \Phi_w(\mathbf{x}_a, \mathbf{x}_c) + \Phi_w(\mathbf{x}_c, \mathbf{x}_b)$

for any $a, b, c \in \{1, 2, \ldots, n\}$ and $r \in \{1, 2, \ldots, \hat{d}\}$. Accordingly, the distance measure yielded by the proposed projection-based representation is a distance metric. □

4.1. Degree of learning freedom

Optimal representation of categorical attributes can be obtained by encode each possible value of the attributes with an optimal numerical value. To search for such optimal representation, each possible value is treated as an encoding variable that will be learned during clustering. For the convenience of analysis without loss of generality, we assume each categorical attribute is with the identical number of possible values $\upsilon = \sum_{r=1}^{d_c} \upsilon^r / d_c$.

Lemma 1. *Given X composed of d_c categorical attributes A_c, the Maximum number of Possible encoding Variables (MoPV) w.r.t. the clustering task that searches for k clusters is $kd_c(\upsilon - 1)$, as the Euclidean space is a one-dimensional space for each attribute, and determining the $\upsilon - 1$ distances between each pair of adjacent possible values is equivalent to the determining of the encoded values of all the υ possible values.*

A higher Degree of Learning Freedom (DoLF) of the independent representation variables can provide a more potential/promising basis for obtaining the optimal attribute representations w.r.t. a clustering task. We analyze the DoLF of the proposed HARR-V/M/T as follows.

Lemma 2. *Given X composed of d_c categorical attributes A_c, DoLF of HARR-V is $d_c\upsilon(\upsilon - 1)/2$ as HARR-V represents each of the d_c attribute a^r into each of the $\upsilon(\upsilon - 1)/2$ one-dimensional spaces, where each such space is spanned by a possible pair of the υ possible values of a^r.*

Theorem 2. *HARR-V yields a hyper-DoLF clustering if $\upsilon > 2k$.*

Proof. Assuming DoLF of HARR-V is larger than MoPV of clustering, then we have $d_c\upsilon(\upsilon - 1)/2 > kd_c(\upsilon - 1)$. By simplification, we have $\upsilon/2 > k$, and thus $\upsilon > 2k$. □

Lemma 3. *Given X composed of d_c categorical attributes A_c, DoLF of HARR-M is $kd_c\upsilon(\upsilon - 1)/2$ as HARR-M represents each attribute a^r into each of the $\upsilon(\upsilon - 1)/2$ one-dimensional spaces w.r.t. each of the k clusters.*

Theorem 3. *HARR-M yields a hyper-DoLF clustering if $\upsilon > 2$.*

Proof. Assuming DoLF of HARR-M is larger than MoPV of clustering, then we have $kd_c\upsilon(\upsilon - 1)/2 > kd_c(\upsilon - 1)$. By simplification, we have $\upsilon/2 > 1$, and thus $\upsilon > 2$. □

4.2. Time complexity

It can be seen from Algorithm 1 that under the same n, d, and k, HARR-V has the same time complexity as the conventional clustering algorithm with attributes weighting mechanism (Huang et al., 2005). However, due to the representation of attributes, $\hat{d}$ of HARR-V will not be smaller than the original d, i.e., $\hat{d} - d = \sum_{r=d_u+1}^{d} \upsilon^r(\upsilon^r - 1)/2 - d_c \geq 0$ as discussed in Section 3.2. But since υ^r is a very small integer ranging from 2 to 8 for most categorical attributes as shown in Table 2, $\hat{d}$ will not significantly increase the computation cost of HARR-V. The time complexity of HARR-V is analyzed below.

Theorem 4. *The time complexity of HARR-V is $O(d^2 n + EInkd)$.*

Table 2
Statistics of the 14 real public datasets, including the number of numerical attributes d_u, categorical attributes d_c, nominal and ordinal categorical attributes d_n and d_o, i.e., $d^c = d^n + d^o$, data objects n, true number of clusters k^*, and the number of possible values of categorical attributes.

No.	Dataset	Abbrev.	d_u	d_c	d_n	d_o	n	k^*	No. of possible values of categorical attributes
1	Inflammations diagnosis	DS	1	5	5	0	120	2	{2, 2, 2, 2, 2}
2	Heart Failure	HF	7	5	5	0	299	2	{2, 2, 2, 2, 2}
3	Autism-Adolescent	AA	2	7	7	0	104	2	{2, 9, 2, 2, 33, 2, 6}
4	Amphibians	AP	2	12	6	6	189	2	{8, 5, 8, 7, 8, 3, 5, 6, 6, 6, 3, 2}
5	Dermatology	DT	1	33	1	32	358	6	{2, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4}
6	Australia Credit	AC	6	8	1	7	690	2	{2, 2, 2, 2, 3, 14, 8, 3}
7	Soybean	SB	0	35	35	0	266	15	{7, 2, 3, 3, 2, 4, 4, 3, 3, 3, 2, 2, 3, 3, 3, 2, 2, 3, 2, 2, 4, 4, 2, 2, 2, 3, 2, 3, 4, 2, 2, 2, 2, 2, 3}
8	Solar Flare	SF	0	9	9	0	323	6	{6, 4, 2, 3, 2, 2, 2, 2, 2}
9	Tic-Tac-Toe	T3	0	9	9	0	958	2	{3, 3, 3, 3, 3, 3, 3, 3, 3}
10	Hayes-Roth	HR	0	4	2	2	132	3	{3, 4, 4, 4}
11	Lymphography	LG	0	18	15	3	148	4	{3, 2, 2, 2, 2, 2, 2, 2, 2, 3, 4, 4, 8, 3, 2, 2, 2, 4, 8}
12	Mushroom	MR	0	20	20	0	8124	2	{6, 4, 10, 2, 9, 2, 2, 2, 12, 2, 4, 4, 9, 9, 4, 3, 5, 9, 6, 7}
13	Lecturer Evaluation	LE	0	4	0	4	1000	5	{5, 5, 5, 5}
14	Social Works	SW	0	10	0	10	1000	4	{3, 3, 4, 3, 4, 2, 3, 3, 3, 3}

Proof. To obtain the worst-case time complexity, we assume that X is a pure nominal data, i.e., $d = d_u$, and let $v = max(v^1, v^2, \ldots, v^d)$. Then we separately prove the time complexity of three parts of HARR in the following.

(1) Projection-based representation (i.e., Step 1 of Algorithm 1): To compute base distances using Eq. (5), CPDs from each of the d attributes given each of the dv possible values should be obtained first by counting the n attribute values with time complexity $O(d^2vn)$. Then, Eq. (5) (with time complexity $O(dv)$) will be triggered $dv(v-1)/2$ times to compute the base distance between $v(v-1)/2$ pairs of intra-attribute possible values for each of the d attributes. So the computation of base distances is with time complexity $O(d^2v^3)$. By using the prepared base distances, all the v possible values of each of the d attributes will be represented onto $v(v-1)/2$ one-dimensional spaces according to Eqs. (7) and (8), with time complexity $O(dv^3)$. Therefore, the time complexity of the projection-based representation is $O(d^2vn + d^2v^3 + dv^3)$.

(2) Learning of $\mathbf{Q}$ and M with fixed $\mathbf{w}$ (i.e., Step 3 and 4 of Algorithm 1): Let I represent the number of iterations for making Step 3 and 4 converge. In each iteration, the distance between each of the n objects and each of the k clusters reflected by each of the $dv(v-1)/2$ represented sub-attributes is computed. Therefore, time complexity for this part is $O(Inkdv^2)$.

(3) Computation of $\mathbf{w}$ (i.e., Step 5 of Algorithm 1): For each a^r from the $dv(v-1)/2$ represented attributes, D^r and S^r should be computed by summing up n and $n(k-1)$ distances according to Eqs. (17) and (18), respectively. As all the value-level distances have been prepared in the projection-based representation phase, the time complexity for computing all the $dv(v-1)/2$ weights by Eqs. (15) and (16) is $O(nkdv^2)$.

Let E be the number of iterations for making the loop formed by Steps 3–5 converge (i.e., times for repeating the computation of part (2) and part (3) in this proof). The time complexity of the HARR algorithm is $O(d^2vn + d^2v^3 + dv^3 + E(Inkdv^2 + nkdv^2))$. Since v is a very small constant ($v \leq 8$ in most cases, see Table 2), the final time complexity of the HARR algorithm is $O(d^2n + EInkd)$. □

Similar to the proof of Theory 4, HARR-M described in Algorithm 1 is with same time complexity as HARR-V.

5. Experiments

We first introduce the experimental settings and then present the evaluation results with analysis.

5.1. Experimental settings

Experimental design, datasets, counterparts to be compared, and validity indices are introduced in the following. All the experiments are coded by MATLAB R2021a. Reported clustering results are the average performance of 20 runs of the compared approaches.

5.1.1. Experimental design

Three types of experiments have been designed to evaluate the proposed HARR-V, HARR-M, and their key technical components. We first compare the clustering performance with the conventional and state-of-the-art counterparts to illustrate the superiority of the proposed approaches. Then, two types of ablation studies are conducted to evaluate the effectiveness of key technical contributions of this work. The total object-cluster distances of HARR-V and HARR-M are also plotted to show their convergence and effectiveness.

5.1.2. Datasets

14 datasets are utilized for the experimental evaluation. All the datasets are real public ones obtained from the UCI machine learning repository[2] (Dua & Karra Taniskidou, 2017) to ensure the reliability and reproducibility of the experiments. Statistics of the datasets are shown in Table 2. The former six mixed datasets are utilized to validate the superiority of the proposed approaches in mixed data clustering. The latter eight categorical datasets are used to specifically verify the effectiveness of the proposed attribute representation and learning mechanisms, since the handling of categorical attributes is the key to mixed data clustering.

Categorical attributes of the 14 datasets can be divided into nominal type and ordinal type, and are with different numbers of possible values. All this statistical information is provided in Table 2. Datasets AP, DT, AC, HR, LG, LE, and SW contain potential ordinal attributes, which are indicated by the official description files of the datasets, or the values of a categorical attribute have an obvious order relationship, e.g., an attribute with possible values {low, mid, high}.

The utilized datasets come from many fields including finance, medicine, biology, etc.; The proportion of different types of attributes varies greatly in the datasets; The scale of the datasets is not large, thus often cannot provide sufficient statistical information. Therefore, analyzing these datasets remains very challenging for state-of-the-art techniques.

5.1.3. Counterparts

Twelve clustering approaches are compared including ten counterparts and the proposed HARR-V and HARR-M. The conventional k-means (Ball & Hall, 1967) clustering algorithm is combined with One-Hot Encoding (OHE, for nominal attributes) and Order Coding (OC, for ordinal attributes) to be a counterpart (OHE+OC), where OC encodes the values of an ordinal attribute into their order values, and then normalizes them. OHE+OC is one of the most straightforward and common solutions for practical mixed data clustering. Two other variants

[2] https://archive.ics.uci.edu.

Table 3
ARI performance of 12 different clustering approaches on six mixed datasets. The interval of ARI is $[-1, 1]$, and a larger value of ARI indicates a better clustering performance. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively. The experimental results of KMD and KPT are combined and reported as "KMD/KPT".

Approach	DS	HF	AA	AP	DT	AC
KMD/KPT	0.2881 ± 0.2718	0.0693 ± 0.0662	−0.0027 ± 0.0125	−0.0068 ± 0.0129	0.4839 ± 0.1044	0.3253 ± 0.1025
OHE+OC	0.2745 ± 0.3484	0.0014 ± 0.0082	−0.0042 ± 0.0087	−0.0162 ± 0.0080	0.6402 ± 0.1597	0.1412 ± 0.2065
SBC	0.1728 ± 0.1719	−0.0022 ± 0.0067	−0.0106 ± 0.0038	−0.0073 ± 0.0005	0.5837 ± 0.1508	0.3622 ± 0.0856
JDM	0.2243 ± 0.1923	0.0039 ± 0.0099	−0.0028 ± 0.0137	−0.0043 ± 0.0162	0.5926 ± 0.1446	0.1451 ± 0.1376
CMS	0.3815 ± 0.1064	0.0782 ± 0.0066	0.0018 ± 0.0142	−0.0034 ± 0.0125	0.5099 ± 0.0970	0.1530 ± 0.0165
UDM	0.3943 ± 0.2863	0.0788 ± 0.0061	0.0069 ± 0.0208	−0.0005 ± 0.0293	0.6207 ± 0.1285	0.3769 ± 0.0436
HOD	0.3943 ± 0.2863	0.0788 ± 0.0061	0.0018 ± 0.0142	−0.0016 ± 0.0258	0.6207 ± 0.1285	0.3769 ± 0.0436
GWD	0.3178 ± 0.2130	0.0951 ± 0.1146	0.0422 ± 0.2001	−0.0089 ± 0.0303	0.0232 ± 0.0185	0.2673 ± 0.0592
GBD	0.5121 ± 0.0962	0.0770 ± 0.0051	0.0025 ± 0.0216	−0.0044 ± 0.0229	0.6191 ± 0.1216	0.3683 ± 0.0401
FBD	0.1799 ± 0.2544	0.0717 ± 0.0745	−0.0110 ± 0.0051	0.0030 ± 0.0000	0.6439 ± 0.1311	0.0597 ± 0.0412
HARR-V	0.3849 ± 0.1897	**0.0882 ± 0.0000**	0.0278 ± 0.0283	−0.0016 ± 0.0456	0.6403 ± 0.1694	0.3453 ± 0.0000
HARR-M	**0.4599 ± 0.2409**	**0.0882 ± 0.0000**	**0.0289 ± 0.0285**	**0.0412 ± 0.0534**	**0.6826 ± 0.1429**	**0.4200 ± 0.0000**

Table 4
ARI performance of 12 different clustering approaches on eight categorical datasets. The interval of ARI is $[-1, 1]$, and a larger value of ARI indicates a better clustering performance. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively. The experimental results of KMD and KPT are combined and reported as "KMD/KPT".

Approach	SB	SF	T3	HR	LG	MR	LE	SW
KMD/KPT	0.3185 ± 0.0362	0.2083 ± 0.0649	0.0247 ± 0.0318	−0.0054 ± 0.0085	0.0683 ± 0.0986	0.3420 ± 0.2311	0.0353 ± 0.0180	0.0485 ± 0.0168
OHE+OC	0.3913 ± 0.0446	0.1961 ± 0.0509	0.0109 ± 0.0098	0.0099 ± 0.0252	0.0892 ± 0.0563	0.4629 ± 0.2502	0.0343 ± 0.0207	0.0004 ± 0.0017
SBC	0.3871 ± 0.0254	0.1483 ± 0.0333	0.0297 ± 0.0182	−0.0097 ± 0.0098	0.0065 ± 0.0029	0.4591 ± 0.2494	0.0295 ± 0.0127	0.0577 ± 0.0115
JDM	0.3826 ± 0.0366	0.1225 ± 0.0355	0.0244 ± 0.0267	−0.0011 ± 0.0086	0.0395 ± 0.0602	0.3191 ± 0.1839	0.0359 ± 0.0115	0.0506 ± 0.0191
CMS	0.3356 ± 0.0449	0.2972 ± 0.0660	0.0171 ± 0.0217	−0.0011 ± 0.0063	0.0824 ± 0.0759	0.4210 ± 0.2690	0.0384 ± 0.0193	0.0608 ± 0.0159
UDM	0.3673 ± 0.0411	0.2800 ± 0.0703	0.0178 ± 0.0250	0.0340 ± 0.0354	0.0860 ± 0.0788	0.3675 ± 0.2519	0.0425 ± 0.0185	0.0620 ± 0.0132
HOD	0.3686 ± 0.0397	0.2799 ± 0.0848	0.0178 ± 0.0250	0.0325 ± 0.0386	0.0762 ± 0.0854	0.3626 ± 0.2348	0.0425 ± 0.0185	0.0661 ± 0.0130
GWD	0.2530 ± 0.0520	0.1829 ± 0.0551	0.0177 ± 0.0210	0.0103 ± 0.0222	0.0619 ± 0.0590	0.1371 ± 0.1649	0.0361 ± 0.0207	0.0379 ± 0.0137
GBD	0.3574 ± 0.0416	0.2395 ± 0.0688	0.0147 ± 0.0223	0.0328 ± 0.0413	0.1012 ± 0.0471	0.3197 ± 0.2474	0.0435 ± 0.0190	0.0627 ± 0.0128
FBD	0.3377 ± 0.0457	0.1996 ± 0.0552	0.0198 ± 0.0013	0.0390 ± 0.0217	0.1530 ± 0.0274	0.3795 ± 0.2169	0.0323 ± 0.0176	0.0598 ± 0.0048
HARR-V	0.4196 ± 0.0441	0.3216 ± 0.0628	0.0211 ± 0.0273	**0.0580 ± 0.0211**	0.1687 ± 0.0226	0.5667 ± 0.1742	**0.0615 ± 0.0415**	**0.0880 ± 0.0136**
HARR-M	**0.4367 ± 0.0495**	**0.3254 ± 0.0748**	**0.0338 ± 0.0252**	0.0524 ± 0.0228	**0.1849 ± 0.0000**	**0.6122 ± 0.0678**	**0.0615 ± 0.0345**	0.0778 ± 0.0215

of k-means, i.e., k-modes (KMD) (Huang, 1998) and k-prototypes (KPT) (Huang, 1997) are also chosen. As KMD is feasible for pure categorical data and KPT is designed for mixed data, the experimental results of KMD and KPT are combined and reported as "KMD/KPT". GoWer's Distance(GWD) (Gower, 1971), Structure-Based Categorical (SBC) data encoding (Qian, Li et al., 2015), Jia's Distance Metric (JDM) (Jia et al., 2016), Coupled Metric Similarity (CMS) (Jian, Cao et al., 2018), Unified Distance Metric (UDM) (Zhang & Cheung, 2022b), HOmogeneous Distance (HOD) metric (Zhang & Cheung, 2022a), Graph-Based Distance (GBD) metric (Zhang & Cheung, 2023), and Forest-Based Distance(FBD) metric (Zhao et al., 2024), combined with KMD and KPT for categorical and mixed data, respectively, have also been chosen as counterparts.

5.1.4. Validity indices

To comprehensively evaluate the clustering performance of compared approaches, we adopt two popular validity indices, i.e., Adjusted Rand Index (ARI) (Santos & Embrechts, 2009) and Clustering Accuracy (CA) (He, Cai, & Niyogi, 2005) for the evaluation. CA is a conventional index in the interval $[0, 1]$, which first searches the optimal mapping between label and clustering results, and then measures the accuracy. ARI in the interval $[-1, 1]$ is a more discriminative index based on the conventional Rand Index (Gates & Ahn, 2017; Rand, 1971). For both ARI and CA, a larger value indicates a better clustering performance.

5.2. Comparison of clustering performance

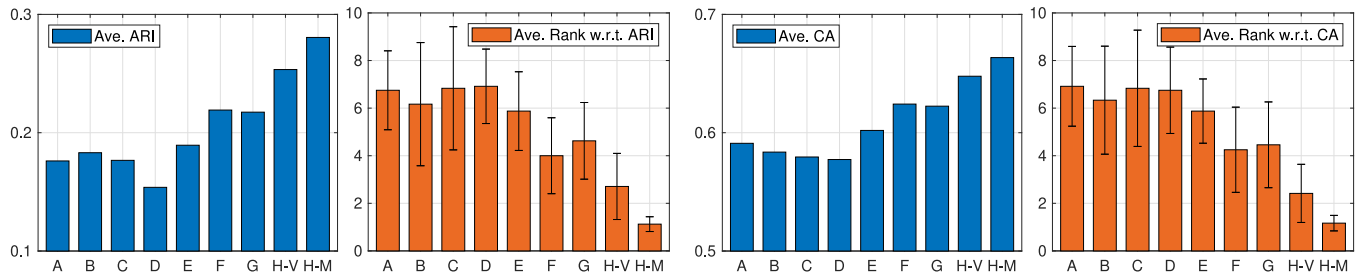
Clustering performance of the compared approaches is shown in Tables 3–6. In general, HARR-M outperforms all the counterparts, and HARR-V is very competitive in most comparisons. More detailed observations are provided in the following: (1) It can be seen that HARR-M outperforms HARR-V in general. This is because that HARR-M more finely considers the contribution of representation to the formation of different clusters; (2) Superiority of the performance of HARR-M and HARR-V is more obvious on mixed data than on categorical data because the advantage of the proposed representation method that homogeneously represents heterogeneous numerical and categorical attributes will surely be weakened on datasets composed of only categorical attributes; (3) Results in Tables 4 and 6 show that both HARR-V and HARR-M perform well on categorical datasets.

Based on the results shown in Tables 3–6, we also compute: (1) the average ARI and CA performance, and (2) the ARI and CA performance rankings, on the datasets for all the counterparts to facilitate an intuitive comparison in Fig. 6. To better organize the histogram results, KMD/KPT, OHE+OC, SBC, JDM, CMS, UDM, HOD, HARR-V, and HARR-M, are interchangeably denoted as A, B, ..., G, H-V, H-M, respectively, hereinafter. The standard deviation of average ARI and CA are not reported because the clustering performance on different datasets may be very different. It can be observed that HARR-V and HARR-M perform obviously better than the other counterparts. Moreover, HARR-M always ranks the first with a very small standard deviation, which indicates its superiority and stability.

To confirm that improvements over baseline methods are significant and consistent, we first implement the Friedman (Friedman, 1937) test on the ARI and CA results in Tables 3–6. The corresponding p-values are 0.00000011 and 0.0000013 respectively, both passing the test under the confidence of 99% (p-value = 0.01). On this basis, results of the two-tailed BD test (Demšar, 2006) with confidence intervals 0.95 ($\alpha = 0.05$) and 0.9 ($\alpha = 0.1$) are shown in Fig. 7, with the corresponding Critical Difference (CD) intervals CD_95% = 3.8158 and CD_90% = 3.4069, respectively, for comparing 12 methods across 14 datasets. As can be seen from Fig. 7, all 10 compared methods fall outside the right boundary of the CD intervals, except for the UDM method w.r.t. ARI performance under $\alpha = 0.05$. But it is worth mentioning that UDM is very close to the boundary of $\alpha = 0.05$ and stays outside the boundary of $\alpha = 0.1$. In general, the test results indicate that our method, HARR-M, significantly outperforms all the other counterparts except for the other version of our method, HARR-V, which achieves comparable performance.

Table 5
CA performance of 12 different clustering approaches on six mixed datasets. The interval of CA is [0, 1], and a larger value of CA indicates a better clustering performance. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively. The experimental results of KMD and KPT are combined and reported as "KMD/KPT".

Approach	DS	HF	AA	AP	DT	AC
KMD/KPT	0.7375 ± 0.1352	0.6224 ± 0.0668	0.5361 ± 0.0330	0.5299 ± 0.0198	0.5964 ± 0.0893	0.7784 ± 0.0658
OHE+OC	0.7196 ± 0.1527	0.5415 ± 0.0330	0.5322 ± 0.0283	0.5079 ± 0.0097	0.6753 ± 0.1240	0.6267 ± 0.1442
SBC	0.6846 ± 0.1085	0.5251 ± 0.0235	0.5173 ± 0.0067	0.5410 ± 0.0024	0.6264 ± 0.1276	0.7946 ± 0.0646
JDM	0.7154 ± 0.1105	0.5580 ± 0.0259	0.5385 ± 0.0391	0.5479 ± 0.0326	0.6536 ± 0.1137	0.6633 ± 0.1031
CMS	0.7686 ± 0.0479	0.6436 ± 0.0054	0.5673 ± 0.0099	0.5397 ± 0.0284	0.6096 ± 0.0768	0.6971 ± 0.0102
UDM	0.7900 ± 0.1302	0.6441 ± 0.0050	0.5745 ± 0.0202	0.5614 ± 0.0402	0.6679 ± 0.0995	0.8069 ± 0.0175
HOD	0.7900 ± 0.1302	0.6441 ± 0.0050	0.5673 ± 0.0099	0.5569 ± 0.0349	0.6679 ± 0.0995	0.8069 ± 0.0175
GWD	0.7433 ± 0.1137	0.6492 ± 0.0950	0.5654 ± 0.1039	0.5550 ± 0.0441	0.2824 ± 0.0261	0.7570 ± 0.0336
GBD	0.7767 ± 0.0447	0.6428 ± 0.0043	0.5611 ± 0.0285	0.5640 ± 0.0321	0.6662 ± 0.0899	0.8034 ± 0.0161
FBD	0.6908 ± 0.1371	0.6284 ± 0.0602	0.5120 ± 0.0162	0.5450 ± 0.0000	0.6828 ± 0.1001	0.6154 ± 0.0494
HARR-V	0.7942 ± 0.1084	**0.6522 ± 0.0000**	**0.6062 ± 0.0305**	0.5550 ± 0.0561	0.6793 ± 0.1411	0.7942 ± 0.0000
HARR-M	**0.8187 ± 0.1247**	**0.6522 ± 0.0000**	**0.6062 ± 0.0302**	**0.6111 ± 0.0535**	**0.7161 ± 0.1130**	**0.8246 ± 0.0000**



(a) Average ARI performance and average ranks. (b) Average CA performance and average ranks.

Fig. 6. Average performance and ranks of different approaches based on the results in Tables 3–6. The methods KMD/KPT, OHE+OC, SBC, JDM, CMS, UDM, HOD, HARR-V, and HARR-M are denoted as A, B, C, D, E, F, G, H-V, and H-M, respectively.

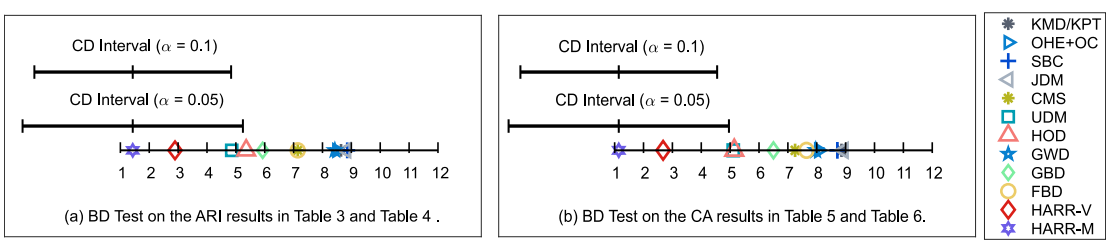


Fig. 7. Results of the two-tailed BD tests on ARI (a) and CA (b) performances across 12 methods on 14 datasets, based on the results from Tables 3 and 6. Confidence intervals are set at 0.95 ($\alpha = 0.05$) and 0.9 ($\alpha = 0.1$), the corresponding Critical Difference (CD) intervals CD_95% = 3.8158 and CD_90% = 3.4069, respectively..

Table 6
CA performance of 12 different clustering approaches on eight categorical datasets. The interval of CA is [0, 1], and a larger value of CA indicates a better clustering performance. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively. The experimental results of KMD and KPT are combined and reported as "KMD/KPT".

Approach	SB	SF	T3	HR	LG	MR	LE	SW
KMD/KPT	0.5021 ± 0.0436	0.4707 ± 0.0523	0.5692 ± 0.0499	0.3826 ± 0.0240	0.6028 ± 0.0933	0.7645 ± 0.1282	0.3276 ± 0.0344	0.3777 ± 0.0348
OHE+OC	0.5348 ± 0.0433	0.4506 ± 0.0393	0.5548 ± 0.0266	0.3989 ± 0.0561	0.6444 ± 0.0568	0.8167 ± 0.1277	0.3455 ± 0.1500	0.0210 ± 0.0941
SBC	0.5368 ± 0.0300	0.4130 ± 0.0284	0.5825 ± 0.0378	0.3591 ± 0.0349	0.5588 ± 0.0062	0.8145 ± 0.1295	0.3172 ± 0.0263	0.3812 ± 0.0195
JDM	0.5417 ± 0.0412	0.4046 ± 0.0306	0.5664 ± 0.0455	0.3932 ± 0.0269	0.5842 ± 0.0707	0.7613 ± 0.1101	0.3142 ± 0.0240	0.3738 ± 0.0274
CMS	0.5177 ± 0.0476	0.5042 ± 0.0484	0.5551 ± 0.0399	0.3962 ± 0.0209	0.6282 ± 0.0783	0.7958 ± 0.1368	0.3274 ± 0.0379	0.3866 ± 0.0292
UDM	0.5376 ± 0.0391	0.5020 ± 0.0532	0.5544 ± 0.0433	0.4470 ± 0.0521	0.6289 ± 0.0828	0.7757 ± 0.1296	0.3324 ± 0.0391	0.3819 ± 0.0231
HOD	0.5393 ± 0.0405	0.5050 ± 0.0641	0.5543 ± 0.0434	0.4405 ± 0.0568	0.6190 ± 0.0833	0.7780 ± 0.1190	0.3324 ± 0.0391	0.3880 ± 0.0257
GWD	0.4462 ± 0.0505	0.4515 ± 0.0550	0.5643 ± 0.0389	0.4163 ± 0.0384	0.6176 ± 0.0611	0.6430 ± 0.1208	0.3352 ± 0.0319	0.3721 ± 0.0254
GBD	0.5295 ± 0.0401	0.4873 ± 0.0577	0.5478 ± 0.0419	0.4383 ± 0.0599	0.6588 ± 0.0447	0.7523 ± 0.1310	0.3341 ± 0.0320	0.3810 ± 0.0199
FBD	0.5199 ± 0.0459	0.4565 ± 0.0444	0.5784 ± 0.0007	0.3905 ± 0.0376	0.6982 ± 0.0211	0.7827 ± 0.0604	0.2912 ± 0.0233	0.3887 ± 0.0172
HARR-V	0.5492 ± 0.0408	0.5127 ± 0.0540	0.5688 ± 0.0386	**0.4883 ± 0.0409**	0.7085 ± 0.0138	0.8642 ± 0.0976	0.3658 ± 0.0585	0.4089 ± 0.0246
HARR-M	**0.5564 ± 0.0533**	**0.5139 ± 0.0605**	**0.5863 ± 0.0405**	0.4674 ± 0.0416	**0.7183 ± 0.0000**	**0.8905 ± 0.0246**	**0.3687 ± 0.0463**	**0.4095 ± 0.0260**

5.3. Ablation studies

To evaluate the effectiveness of the key technical components proposed in this paper, the performance of different ablated versions of HARR-V and HARR-M is compared in Tables 7 and 8. We combine the Homogeneous Attribute Representation (HAR) proposed in Section 3.2 with KMD/KPT (without representation learning), and compare its performance with HARR-V and HARR-M to illustrate the effectiveness of the weights updating mechanisms proposed in Section 3.3. Performance of the Base Distance (BD) described by Eq. (5) and Fig. 3 combined with KMD/KPT is compared with HAR to verify the effectiveness of the projection mechanism of the representation. Performance of KMD/KPT

Table 7
ARI performance of five ablated HARR versions on 14 datasets. The interval of ARI is $[-1, 1]$, and a larger value of ARI indicates a better clustering performance. $\overline{AR}$ row reports the average ranks.

Data	KMD/KPT	BD	HAR	HARR-V	HARR-M
DS	0.2881 ± 0.2718	0.3374 ± 0.2798	0.3374 ± 0.2798	0.3849 ± 0.1897	0.4599 ± 0.2409
HF	0.0693 ± 0.0662	0.0793 ± 0.0181	0.0793 ± 0.0181	0.0882 ± 0.0000	0.0882 ± 0.0000
AA	−0.0027 ± 0.0125	−0.0030 ± 0.0106	0.0430 ± 0.0313	0.0278 ± 0.0283	0.0289 ± 0.0285
AP	−0.0068 ± 0.0129	−0.0020 ± 0.0288	−0.0002 ± 0.0300	−0.0016 ± 0.0456	0.0412 ± 0.0534
DT	0.4839 ± 0.1044	0.6322 ± 0.1343	0.6322 ± 0.1343	0.6403 ± 0.1694	0.6826 ± 0.1429
AC	0.3253 ± 0.1025	0.3773 ± 0.0455	0.3773 ± 0.0455	0.3453 ± 0.0000	0.4200 ± 0.0000
SB	0.3185 ± 0.0362	0.3686 ± 0.0397	0.3912 ± 0.0355	0.4196 ± 0.0441	0.4367 ± 0.0495
SF	0.2083 ± 0.0649	0.2799 ± 0.0848	0.3090 ± 0.0879	0.3216 ± 0.0628	0.3254 ± 0.0748
T3	0.0247 ± 0.0318	0.0178 ± 0.0250	0.0196 ± 0.0226	0.0211 ± 0.0273	0.0338 ± 0.0252
HR	−0.0054 ± 0.0085	0.0325 ± 0.0386	0.0244 ± 0.0177	0.0580 ± 0.0211	0.0524 ± 0.0228
LG	0.0683 ± 0.0986	0.0762 ± 0.0854	0.1187 ± 0.0676	0.1687 ± 0.0226	0.1849 ± 0.0000
MR	0.3420 ± 0.2311	0.3626 ± 0.2348	0.4165 ± 0.2588	0.5667 ± 0.1742	0.6122 ± 0.0678
LE	0.0353 ± 0.0180	0.0425 ± 0.0185	0.0439 ± 0.0167	0.0615 ± 0.0415	0.0615 ± 0.0345
SW	0.0485 ± 0.0168	0.0661 ± 0.0130	0.0809 ± 0.0135	0.0880 ± 0.0136	0.0778 ± 0.0215
$\overline{AR}$	4.7143 ± 0.8254	3.8571 ± 0.6630	2.9286 ± 0.8287	2.1071 ± 0.8810	1.3929 ± 0.6257

Table 8
CA performance of five ablated HARR versions on 14 datasets. The interval of CA is $[0, 1]$, and a larger value of CA indicates a better clustering performance. $\overline{AR}$ row reports the average ranks. $\overline{IP}$ row reports the improved percentage achieved by the current one w.r.t. the previous one. In Table 7, since ARI is in the interval $[-1, 1]$, $\overline{IP}$ results are not reported.

Data	KMD/KPT	BD	HAR	HARR-V	HARR-M
DS	0.7375 ± 0.1352	0.7650 ± 0.1289	0.7650 ± 0.1289	0.7942 ± 0.1084	0.8187 ± 0.1247
HF	0.6224 ± 0.0668	0.6436 ± 0.0145	0.6436 ± 0.0145	0.6522 ± 0.0000	0.6522 ± 0.0000
AA	0.5361 ± 0.0330	0.5404 ± 0.0302	0.6197 ± 0.0475	0.6062 ± 0.0305	0.6062 ± 0.0302
AP	0.5299 ± 0.0198	0.5561 ± 0.0381	0.5614 ± 0.0446	0.5550 ± 0.0561	0.6111 ± 0.0535
DT	0.5964 ± 0.0893	0.6811 ± 0.1071	0.6811 ± 0.1071	0.6793 ± 0.1411	0.7161 ± 0.1130
AC	0.7784 ± 0.0658	0.8070 ± 0.0182	0.8070 ± 0.0182	0.7942 ± 0.0000	0.8246 ± 0.0000
SB	0.5021 ± 0.0436	0.5393 ± 0.0405	0.5258 ± 0.0457	0.5492 ± 0.0408	0.5564 ± 0.0533
SF	0.4707 ± 0.0523	0.5050 ± 0.0641	0.5053 ± 0.0770	0.5127 ± 0.0540	0.5139 ± 0.0605
T3	0.5692 ± 0.0499	0.5543 ± 0.0434	0.5602 ± 0.0405	0.5688 ± 0.0386	0.5863 ± 0.0405
HR	0.3826 ± 0.0240	0.4405 ± 0.0568	0.4496 ± 0.0402	0.4883 ± 0.0409	0.4674 ± 0.0416
LG	0.6028 ± 0.0933	0.6190 ± 0.0833	0.6687 ± 0.0551	0.7085 ± 0.0138	0.7183 ± 0.0000
MR	0.7645 ± 0.1282	0.7780 ± 0.1190	0.7940 ± 0.1368	0.8642 ± 0.0976	0.8905 ± 0.0246
LE	0.3276 ± 0.0344	0.3324 ± 0.0391	0.3330 ± 0.0324	0.3658 ± 0.0585	0.3687 ± 0.0463
SW	0.3777 ± 0.0348	0.3880 ± 0.0257	0.4011 ± 0.0233	0.4089 ± 0.0246	0.4095 ± 0.0260
$\overline{AR}$	4.7857 ± 0.8018	3.6429 ± 0.6914	2.9286 ± 0.7810	2.4286 ± 0.9579	1.2143 ± 0.4688
$\overline{IP}$	–	5.5546%	2.4933%	3.5547%	2.3478%

is also reported for completeness.

In Tables 7 and 8, to make the results easy to observe, the row of Average Rank ($\overline{AR}$) reports the average performance ranks of compared approaches, and the row of Improved Percentage ($\overline{IP}$) reports the average improvement each approach achieves over its left approach. Since ARI is in the interval [−1,1], $\overline{IP}$ results are not reported in Table 7. It can be observed that the performance of KMD/KPT, BD, HAR, HARR-V, and HARR-M increases in general. More specifically, the following four observations: (1) BD outperforms KMD/KPT, (2) HAR outperforms BD, (3) HARR-V outperforms HAR, and (4) HARR-M outperforms HARR-V, can clearly verify the effectiveness of: (1) base distance adopted by the projection, (2) projection-based representation, (3) weights learning mechanism of HARR-V, and (4) weights learning mechanism of HARR-M, respectively.

This work distinguishes categorical attributes into nominal and ordinal ones. To evaluate the necessity of doing so, the clustering performance of approaches is compared by treating potential ordinal attributes in the following two ways: (1) distinguish nominal and ordinal attributes, and (2) treat all the categorical attributes as nominal ones. Note that only the approaches feasible for distinguishing nominal and ordinal attributes are compared here. According to the results visualized in Fig. 8, treating nominal and ordinal attributes differently can obviously boost clustering performance in most cases, which confirms the necessity of our distinction between ordinal and nominal attributes.

5.4. Learning process evaluation

The values of total object-cluster distance $z = \sum_{i=1}^{n}\sum_{l=1}^{k^*} q_{il}\Phi_w(\mathbf{x}_i, \mathbf{m}_l)$ produced by HARR-V and HARR-M on the six mixed datasets are plotted in Fig. 9. The x-axis is the number of learning iterations, and the y-axis is the value of z. The circles and boxes indicate that the weights were updated by HARR-V and HARR-M in the corresponding iterations, respectively.

It can be observed from the figures that the z values of HARR-V and HARR-M monotonically decrease on all the datasets, and both the two algorithms always converge quickly within 15 iterations. This indicates that although HARR-V and HARR-M need to learn the representations of $\hat{d}$ attributes ($\hat{d} \geq d$), they are still very efficient, which is a very important characteristic in practical applications. It can also be observed that the z values of HARR-M are never higher than the z values of HARR-V at convergence, which again indicates the effectiveness of the weights updating strategy of HARR-M.

5.5. Efficiency evaluation

To specifically evaluate the efficiency of HARR-V and HARR-M, we compare their execution time with their state-of-the-art counterparts. To better observe the scalability and verify the time complexity, we compare the approaches on a large randomly generated synthetic dataset with $n = 10,0000$, $d = d^c = 5$, $v^1 = v^2 = \cdots = 5$ and $k = 5$. All the compared approaches are implemented by first computing the similarity/distance matrices of each attribute and then performing clustering by reading off the similarities/distances. The time (y-axis) - n (x-axis) curves in Fig. 10 are formed by sampling the synthetic dataset using the sampling rates $\varphi = \{0.001, 0.2, 0.4, 0.6, 0.8, 1\}$. The efficiency-related experiments are coded by MATLAB R2021b and implemented by a PC (Intel i7-9700 CPU @ 3.00 GHz, 16 GB RAM).

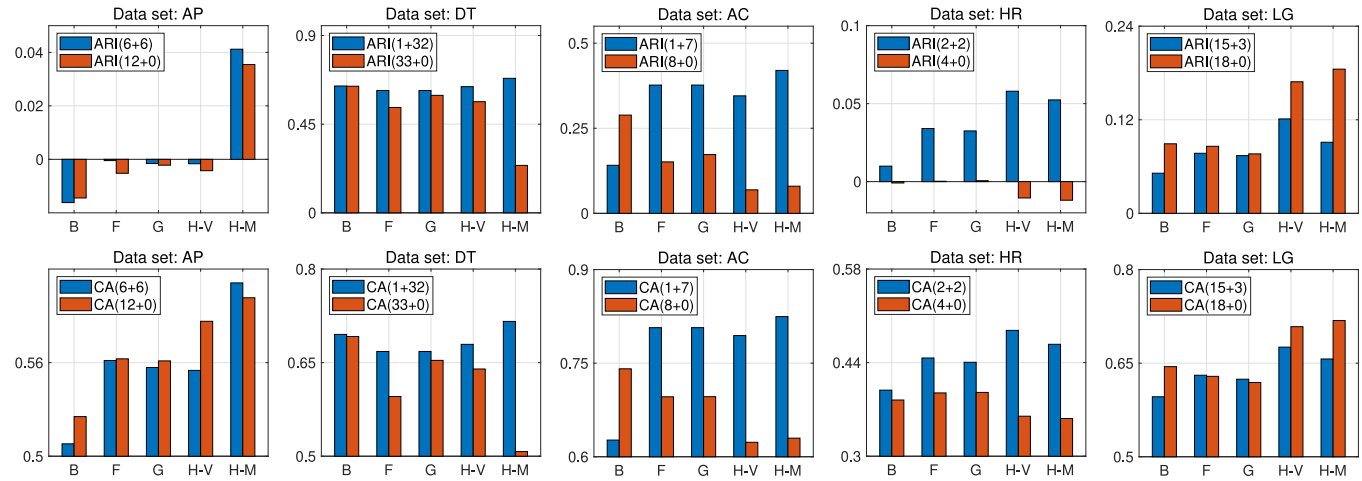


Fig. 8. ARI and CA performance of different approaches with and without distinguishing ordinal attributes on AP, DT, AC, HR, and LG datasets. The two numbers in parentheses in the legend report the value of $d_n + d_o$ used to treat the dataset.

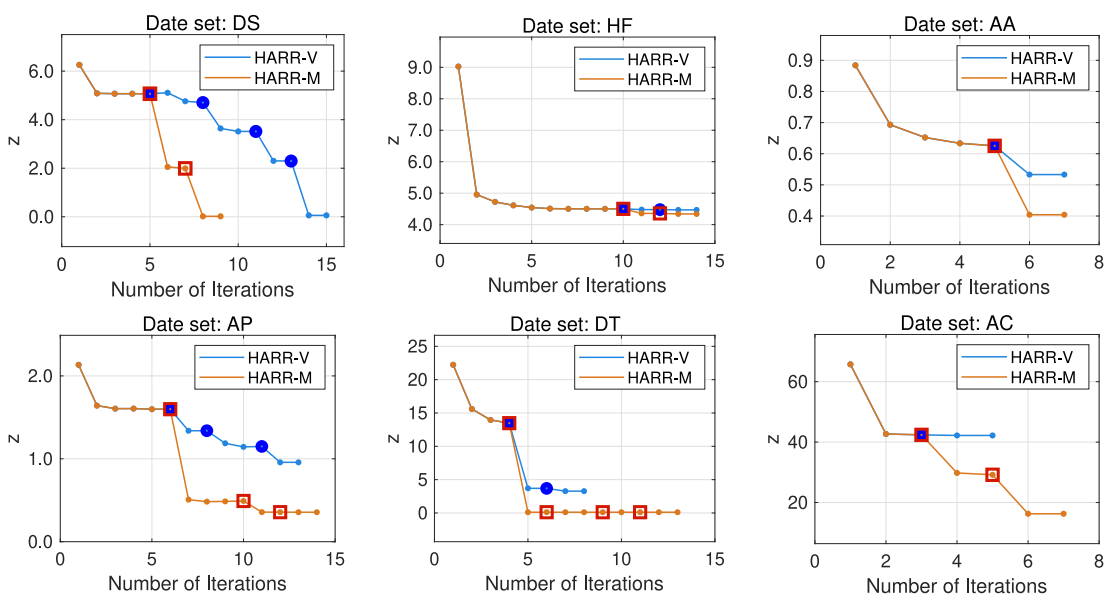


Fig. 9. Total object-cluster distance z of HARR-V and HARR-M on different mixed data sets. The blue circles and red boxes indicate that the weights were updated by HARR-V and HARR-M, respectively.

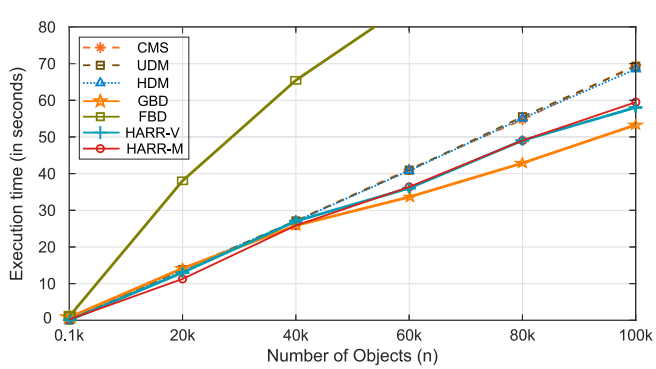


Fig. 10. Comparison of the execution time of CMS, UDM, HDM, GBD, FBD, HARR-V, and HARR-M on a large synthetic dataset ($n = 100,000$) sampled at different sampling rates $\varphi = \{0.001, 0.2, 0.4, 0.6, 0.8, 1\}$.

It is intuitive that the execution time of HARR-V and HARR-M increases linearly, which verifies the time complexity analyzed in Theorem 4. Moreover, the time curves of both methods exhibit nearly identical trends, further validating their practical performance and consistency with the theoretical expectations. This consistency demonstrates the scalability of the proposed methods, making them well-suited for handling large-scale datasets. In addition, it can be seen that even though HARR involves the weights learning procedure, its execution time is still shorter than that of CMS, UDM, HDM, and FBD. This is because the similarity computation cost dominates the execution time for all the compared approaches. HARR directly adopts the CPDs' differences as the base distance measure and is thus relatively more efficient than the similarity measurement strategies adopted by the other three metrics.

5.6. Intuitive study

To showcase improvements of HARR-V and HARR-M and the intuitiveness of the distance structure they obtain, with a practical case. We use the distance between attribute values learned by OHE, GBD, FBD,

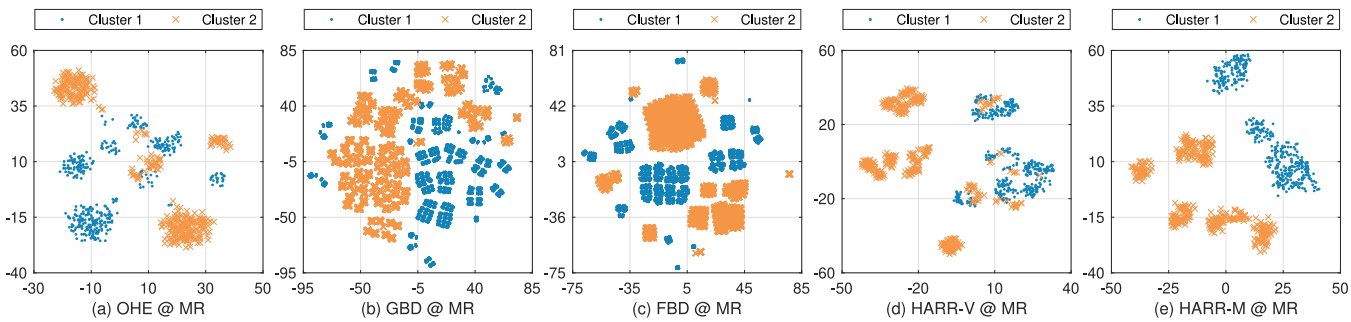


Fig. 11. t-SNE visualization of the MR dataset represented by OHE, GBD, FBD, HARR-V, and HARR-M. Data points are marked in different colors according to the 'true' labels. A higher concentration of data points with the same label indicates that the corresponding method's distance metric better discriminates between clusters.

HARR-V and HARR-M to encode the attributes of the MR dataset. The encoded data are then dimensionally reduced into a 2-D space through t-SNE (Maaten & Hinton, 2008) and visualized in Fig. 11 by marking the data points with 'true' labels provided by the dataset in different colors. If more data points with the same label are gathered in the visualization, then it indicates that the corresponding distance metric is more competent in discriminating different clusters.

It is evident from the figure that the cluster discrimination ability of the HARR-V and HARR-M methods is significantly better than that of OHE, GBD, and FBD. Specifically, compared to the methods shown in Fig. 11(a)–(c), (d) and (e) demonstrate that the distance structures learned by HARR-V and HARR-M enable the encoded MR dataset to exhibit more distinct cluster separation. This advantage stems from the projection mechanism employed by the HARR framework, which effectively preserves the information between categorical attributes. Additionally, by automating the assignment of attribute weights, the HARR methods adequately consider both inter-cluster and intra-cluster relationships, leading to more reasonable and discriminative representations.

These improvements are particularly valuable in real-world scenarios where accurate and efficient clustering of heterogeneous data is critical, such as in market segmentation, healthcare diagnostics, or recommendation systems. In these domains, being able to distinguish clusters with more precision and interpretability can significantly enhance decision-making and result in more actionable insights. For example, in market segmentation, the ability to capture subtle differences between customer groups based on categorical attributes can lead to more targeted marketing strategies.

6. Concluding remarks

This paper proposes an attribute representation learning approach for solving the clustering problem of mixed data composed of heterogeneous attributes. Based on the analysis of the concepts described by the attributes, the values of each attribute are represented in homogeneous linear distance spaces. Such homogeneous representation provides an effective basis for the fusion of heterogeneous information from the different types of attributes in mixed data clustering. To make the representations adapt to clustering tasks, we propose a clustering paradigm to simultaneously search the weights of represented attributes and partitions of data objects. The designed learning mechanisms can effectively circumvent sub-optimal solutions to a certain extent and demonstrate superior clustering performance on both categorical data and more challenging mixed data. Moreover, the proposed clustering algorithms converge quickly and do not involve non-trivial parameter settings.

Despite the effectiveness of our method on static datasets, certain limitations remain. Specifically, our method tends to underperform when applied to datasets containing missing or noisy values. Additionally, in the context of complex and dynamic environments involving streaming data, the multiple spaces projection mechanism employed by our approach may struggle to maintain its effectiveness. Addressing these issues will be a key focus of our future research. In particular, incorporating mechanisms for handling noisy or incomplete data, as well as developing incremental metric fusion techniques for streaming data under dynamic conditions, are promising directions for extending and improving our method.

CRediT authorship contribution statement

Yiqun Zhang: Methodology, Programming, Writing –original draft preparation. **Mingjie Zhao:** Data preparation, Writing –revision and response. **Yizhou Chen:** Visualization, Writing – reviewing & editing. **Yang Lu:** Supervision, Conceptualization, Reviewing and editing. **Yiu-ming Cheung:** Validation, Reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62476063, and 62376233, the NSFC/Research Grants Council (RGC) Joint Research Scheme, China under the grant N_HKBU214/21, the Natural Science Foundation of Guangdong Province, China under grant 2023A1515012855, the Natural Science Foundation of Fujian Province, China under grant 2024J09001, the General Research Fund of RGC, China under grants: 12202622, and 12201323, and the RGC Senior Research Fellow Scheme, China under grant SRFS2324-2S02.

Data availability

All datasets are publicly available.

References

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

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 Recognition Letters*, *28*(1), 110–118.

Arabie, P., Baier, N. D., Critchley, C. F., & Keynes, M. (2006). *Studies in classification, data analysis, and knowledge organization*, Springer.

Bai, L., Liang, J., Dang, C., & Cao, F. (2011). A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognition*, *44*(12), 2843–2861.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science*, *12*(2), 153–155.

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining* (pp. 243–254).

Chen, L., Wang, S., Wang, K., & Zhu, J. (2016). Soft subspace clustering of categorical data with probabilistic distance. *Pattern Recognition*, *51*, 322–332.

Cheung, Y.-m., & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, *46*(8), 2228–2238.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(1), 1–30.

dos Santos, T. R., & Zárate, L. E. (2015). Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, *42*(3), 1247–1260.

Dua, D., & Karra Taniskidou, E. (2017). UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, *93*, 95–112.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gates, A. J., & Ahn, Y.-Y. (2017). The impact of random models on clustering similarity. *Journal of Machine Learning Research*, *18*(1), 3049–3076.

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882–907.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection. In *Proceedings of the 18th international conference on neural information processing systems* (pp. 507–514).

Hu, L., Jiang, M., Liu, Y., & He, Z. (2022). Significance-based categorical data clustering. arXiv preprint arXiv:2211.03956.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the first Pacific-Asia conference on knowledge discovery and data mining* (pp. 21–34).

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657–668.

Ienco, D., Pensa, R. G., & Meo, R. (2009). Context-based distance learning for categorical data clustering. In *Proceedings of the eighth international symposium on intelligent data analysis* (pp. 83–94).

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

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

Jia, H., & Cheung, Y.-m. (2018). Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3308–3325.

Jia, H., Cheung, Y.-m., & Liu, J. (2016). A new distance metric for unsupervised learning of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(5), 1065–1079.

Jian, S., Cao, L., Lu, K., & Gao, H. (2018). Unsupervised coupled metric similarity for non-IID categorical data. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1810–1823.

Jian, S., Cao, L., Pang, G., Lu, K., & Gao, H. (2017). Embedding-based representation of categorical data by hierarchical value coupling learning. In *Proceedings of the 26th international joint conference on artificial intelligence*.

Jian, S., Hu, L., Cao, L., & Lu, K. (2018). Metric-based auto-instructor for learning mixed data representation. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 3318–3325).

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

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*(8).

Khan, A. A., Mishra, A. C., & Mohanty, S. K. (2023). An entropy-based weighted dissimilarity metric for numerical data clustering using the distribution of intra feature differences. *Knowledge-Based Systems*, *280*, Article 110967.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

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

Li, Q., Ji, S., Hu, S., Yu, Y., Chen, S., Xiong, Q., et al. (2023). A multi-view deep metric learning approach for categorical representation on mixed data. *Knowledge-Based Systems*, *260*, Article 110161.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning* (pp. 296–304).

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

Qian, Y., Li, F., Liang, J., Liu, B., & Dang, C. (2015). Space structure and clustering of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(10), 2047–2059.

Qian, Y., Xu, H., Liang, J., Liu, B., & Wang, J. (2015). Fusing monotonic decision trees. *IEEE Transactions on Knowledge and Data Engineering*, *27*(10), 2717–2728.

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

Santos, J. M., & Embrechts, M. (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* (pp. 175–184). Springer.

Yu, Y., Fu, L., Cheng, Y., & Ye, Q. (2022). Multi-view distance metric learning via independent and shared feature subspace with applications to face and forest fire recognition, and remote sensing classification. *Knowledge-Based Systems*, *243*, Article 108350.

Zhang, Y., & Cheung, Y.-m. (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).

Zhang, Y., & Cheung, Y.-m. (2020). An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 6869–6876).

Zhang, Y., & Cheung, Y.-m. (2022a). Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3560–3576.

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

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

Zhang, Y., Cheung, Y.-m., & Tan, K. (2020). A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(1), 39–52.

Zhang, Y., Cheung, Y.-m., & Zeng, A. (2022). Het2Hom: Representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering. In *Proceedings of the 31st international joint conference on artificial intelligence* (pp. 3758–3765).

Zhao, M., Feng, S., Zhang, Y., Li, M., Lu, Y., & Cheung, Y.-M. (2024). Learning order forest for qualitative-attribute data clustering. In *Proceedings of the 27th European conference on artificial intelligence*.

Zhu, C., Cao, L., & Yin, J. (2022). Unsupervised heterogeneous coupling learning for categorical representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 553–549.

Zhu, C., Zhang, Q., Cao, L., & Abrahamyan, A. (2020). Mix2Vec: Unsupervised mixed data representation. In *Proceedings of the seventh international conference on data science and advanced analytics* (pp. 118–127).