Het2Hom: Representation of Heterogeneous Attributes into Homogeneous Concept Spaces for Categorical-and-Numerical-Attribute Data Clustering

Yiqun Zhang[1] , **Yiu-ming Cheung**[2*] , **An Zeng**[1]

[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
yqzhang@gdut.edu.cn, ymc@comp.hkbu.edu.hk, zengan@gdut.edu.cn

Abstract

Data sets composed of a mixture of categorical and numerical attributes (also called mixed data hereinafter) are common in real-world cluster analysis. However, insightful analysis of such data under an unsupervised scenario using clustering is extremely challenging because the information provided by the two different types of attributes is heterogeneous, being at different concept hierarchies. That is, the values of a categorical attribute represent a set of different concepts (e.g., professor, lawyer, and doctor of the attribute "occupation"), while the values of a numerical attribute describe the tendencies toward two different concepts (e.g., low and high of the attribute "income"). To appropriately use such heterogeneous information in clustering, this paper therefore proposes a novel attribute representation learning method called Het2Hom, which first converts the heterogeneous attributes into a homogeneous form, and then learns attribute representations and data partitions on such a homogeneous basis. Het2Hom features low time complexity and intuitive interpretability. Extensive experiments show that Het2Hom outperforms the state-of-the-art counterparts.

1 Introduction

Categorical values, which refer to the qualitative values without explicit numerical meanings, are quite common in machine learning and data analysis tasks [Agresti, 2003]. Given a data set, the attributes that describe the data samples using a set of categorical values are called categorical attributes. It is inevitable to process categorical-and-numerical attribute data in cluster analysis, as clustering is one of the most commonly used machine learning techniques for unsupervised data analysis. Unlike numerical attributes, categorical attribute values are infeasible for arithmetic computation and do not have a well-defined similarity space. It is therefore extremely challenging to appropriately use the information provided by categorical and numerical attributes in cluster analysis. To address this issue, the existing efforts that have been paid can be roughly divided into two types: (i) represent categorical attribute values into numerical values and treat the represented attributes as numerical ones for clustering, and (ii) directly define similarities for categorical attributes and then perform cluster analysis [Xu and Wunsch, 2005][Boriah *et al.*, 2008][dos Santos and Zárate, 2015].

For the representation-based methods, one-hot encoding is the most common one, which encodes categorical values into boolean vectors. In recent years, more powerful representation-based methods [Qian *et al.*, 2015][Jian *et al.*, 2018b] have been proposed to encode categorical attributes by extracting and embedding more valuable information into the representations. Recently, a more advanced representation method [Zhu *et al.*, 2022] further introduces multiple kernel functions to learn the representations. Although the above-mentioned recent progress has achieved considerable improvements in clustering performance, they are all designed for pure categorical data only and have not considered the common mixed data clustering problem.

For defining the similarities between categorical values, the conventional Hamming distance simply assigns distances 0 and 1 to identical and different values. Some other measures [Goodall, 1966][Lin, 1998][Cheung and Jia, 2013][Zhang *et al.*, 2020] define the similarities more finely based on the occurrence probabilities of possible values. To further consider the interdependence of attributes, similarity measures [Ienco *et al.*, 2012][Jia *et al.*, 2016][Jian *et al.*, 2018a][Zhang and Cheung, 2022] have been successively presented in the literature. The above-mentioned measures are usually combined with k-prototype clustering algorithm [Huang, 1997][Kacem *et al.*, 2015], which is designed for mixed data clustering. Most recently, a similarity learning method [Zhang and Cheung, 2021] has been proposed to make the similarities learnable in clustering.

Nevertheless, as far as we know, clustering performance on mixed data is still far from satisfactory because none of the existing methods can perform representation or similarity formulation based on the establishment of a homogeneous connection between the heterogeneous categorical and numerical attributes. From the perspective of the concepts expressed by the attributes, categorical attributes and numerical attributes are in different concept hierarchies. For example, the values of a numerical attribute describe the tendencies toward two different concepts, e.g., {high, low} of attribute "income",

*Corresponding author

while each possible value of a categorical attribute stands for a different concept, e.g., {professor, lawyer, doctor} of attribute "occupation", which can produce three pairs of different concepts. Obviously, with the new insight that categorical and numerical attributes are in different concept hierarchies, existing methods still leave us a considerable space for information mining.

In this paper, we propose a novel method called Het2Hom to learn the representations of categorical attributes for mixed data clustering. Het2Hom first projects all the values of an attribute into the spaces spanned by different concept pairs of this attribute, to obtain an informative homogeneous representation of categorical and numerical attributes. Accordingly, a learning mechanism is elaborately designed so that the learning of attribute representations and data objects partition can adapt to each other more appropriately. Extensive experiments show the efficacy of the proposed Het2Hom, and its main advantages are three-fold:

- Het2Hom represents categorical attributes into the form of numerical attributes while preserving the original relationship information of the possible values, thus providing an appropriate basis for mixed data learning.
- A learning mechanism has been designed to make the attribute representation and data objects partition adapt to each other during clustering, thereby somewhat avoiding sub-optimal solutions in the optimization.
- Het2Hom achieves superior clustering performance on both categorical data and mixed data. Furthermore, its representation learning process is efficient and the results are highly interpretable.

2 Related Work

Representation-based clustering approaches have two common procedures: (1) represent the data set based on a certain strategy, (2) perform clustering by treating the represented categorical attributes as numerical ones. The simplest one-hot encoding is the most commonly used one, which encodes each possible value of an attribute into a vector by setting the bit corresponding to the possible value to 1 and the other bits to 0. Since it assigns an identical distance to any pair of unequal values, it is incapable to distinguish the different dissimilarity degrees. Space structure-based representation [Qian *et al.*, 2015] has been proposed providing a solution for capturing the value and attribute couplings of categorical data. It encodes a target data object by concatenating the distances from it to all the objects. Later, coupling-based representations [Jian *et al.*, 2017][Jian *et al.*, 2018b] have also been proposed to encode the couplings of categorical data. Such methods further perform k-means [Ball and Hall, 1967] clustering and PCA to obtain a more concise representation of the couplings. Recently, a more advanced method [Zhu *et al.*, 2022] adopting different kernel functions has been proposed to more comprehensively represent the couplings. Most recently, the deep learning-based method [Zhu *et al.*, 2020] has also been proposed focusing on the dynamic representation of streaming data with concept-drifts. In summary, none of the above-mentioned approaches can appropriately handle the heterogeneity of mixed data.

For the approaches that directly define the similarities, the widely used Hamming distance uniformly assigns distance 1 to any pair of unequal values and assigns distance 0 to identical values, which has the same drawback as the one-hot encoding. Therefore, probability-based similarity measures [Goodall, 1966][Lin, 1998][Cheung and Jia, 2013][Zhang and Cheung, 2018][Zhang *et al.*, 2020] have been proposed to more finely define similarities based on the occurrence probabilities of possible values. All the above-mentioned measures treat each attribute independently and ignore the valuable information provided by the inter-attribute dependence. Therefore, the measures [Ahmad and Dey, 2007][Le and Ho, 2005][Ienco *et al.*, 2012] have been proposed to define similarities according to conditional probability distributions obtained from different attributes as given target possible values. Since these measures rely on the sole of interdependence of attributes, they will still fail when all the attributes are independent of each other. To solve this problem, the measures [Jia *et al.*, 2016][Jian *et al.*, 2018a][Zhang and Cheung, 2022] that simultaneously consider the intra- and inter-attribute statistical information have been proposed. For all the above-mentioned measures, k-prototypes algorithm [Huang, 1997] are usually used to perform clustering. Nevertheless, since similarity measurement and clustering are performed independently, the measured similarities cannot adapt well to the clustering task. Accordingly, more advanced methods [Zhang and Cheung, 2020][Zhang and Cheung, 2021] have been proposed to interactively learn the similarities and data partitions. Although they achieve superior clustering performance, they are designed for categorical data only.

3 Proposed Method

In this section, we first formulate the problem. Then, we present the Het2Hom with a learning algorithm. Table 1 sorts out the frequently used symbols in this paper. Specific definitions of the symbols will also be given where they first appear in the following text.

3.1 Problem Formulation

Given a data set $S = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ with n data objects, each object $\mathbf{x}_i = [x_i^1, x_i^2, ..., x_i^d]^\top$ is a d-dimensional vector taking values from the d attributes $A = \{a^1, a^2, ..., a^d\}$, where a categorical attribute a^r has v^r possible values $\{o_1^r, o_2^r, ..., o_{v^r}^r\}$. For convenience but without loss of generality, we assume that the former d^c attributes in A are categorical and the latter d^u are numerical. Clustering refers to the task that assigns the n data objects in S to k proper clusters $C = \{c_1, c_2, ..., c_k\}$, which can be formalized as minimizing

$$z(\mathbf{Q}, M, W) = \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il} \cdot \Phi(\mathbf{x}_i, \mathbf{m}_l), \tag{1}$$

where $\mathbf{Q}$ is an $n \times k$ matrix indicating the object-cluster affiliations, and the (i, l)th entry q_{il} of $\mathbf{Q}$ is defined as

$$q_{il} = \begin{cases} 1 & , \quad \text{if } l = \arg\min_{y} \Phi(\mathbf{x}_i, \mathbf{m}_y) \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{2}$$

Symbol	Explanation
$\mathbf{x}_i$	ith data object
x_i^r	rth value of $\mathbf{x}_i$
a^r	rth attribute
o_h^r	hth possible value of a^r
v^r	Number of possible values of a^r
w^r	Weight indicating the importance of a^r
d^c	Number of categorical attributes
d^u	Number of numerical attributes
d	Number of attributes, $d = d^c + d^u$
c_l	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
$\mathcal{R}_b^r$	bth endogenous space corresponding to a^r
$\mathcal{R}^r$	Endogenous space set, $\mathcal{R}^r = \{\mathcal{R}_1^r, \mathcal{R}_2^r, ..., \mathcal{R}_{\gamma^r}^r\}$
w_b^r	Weight indicating the importance of $\mathcal{R}_b^r$
$e_y^r(l)$	Total error contributed by $\mathcal{R}_y^r$ on cluster c_l
$\Phi(\cdot,\cdot)$	Data object-level dissimilarity
$\phi(\cdot,\cdot)$	Value-level distance
$\kappa(\cdot,\cdot)$	Base distance

Table 1: Explanation of symbols.

As we focus on the crisp clustering problem, q_{il} satisfies $\sum_{l=1}^{k} q_{il} = 1$ and $q_{il} \in \{0, 1\}$. $\Phi(\mathbf{x}_i, \mathbf{m}_y)$ is the dissimilarity between data object $\mathbf{x}_i$ and cluster c_l described by a vector $\mathbf{m}_l = [m_l^1, m_l^2, ..., m_l^d]^\top$ from $M = \{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_k\}$. The value of $\mathbf{m}_l$ can be computed following the way of the conventional k-prototypes clustering algorithm. That is, for the numerical case (i.e. $r > d^c$), the value of m_l^r is the mean of the values from a^r in c_l, while for the categorical case (i.e. $r \leq d^c$), the value of m_l^r is equal to the most frequent possible value from a^r in c_l. The dissimilarity can be written in a general form as

$$\Phi(\mathbf{x}_i, \mathbf{m}_l) = \sum_{r=1}^{d} \phi(x_i^r, m_l^r) \cdot w^r, \tag{3}$$

where $\phi(x_i^r, m_l^r)$ is the distance between $\mathbf{x}_i$ and $\mathbf{m}_l$ in terms of their values on attribute a^r, and each weight w^r from $W = \{w^1, w^2, ..., w^d\}$ indicates the importance of a^r in clustering.

3.2 Projection-based Representation

Het2Hom is proposed to represent the heterogeneous attributes into homogeneous forms, thus providing a homogeneous basis for defining $\Phi(\mathbf{x}_i, \mathbf{m}_l)$ and $\phi(x_i^r, m_l^r)$. As discussed in Section I, our goal is to project the values of a categorical attribute into the concept spaces in the hierarchy of numerical attributes. We thus project all the values of a categorical attribute a^r into the one-dimensional space spanned by a pair of possible values o_g^r and o_h^r. Such a space is called endogenous space because it is endogenously generated by the intra-attribute possible values.

As shown in Figure 1, since all the attribute values are projected into an endogenous space, a structural representation of the distance space is thus obtained. Our goal is to obtain as many possible representations of categorical attributes as possible, and make them learnable with the clustering task, thereby achieving more flexible representations. For a categorical

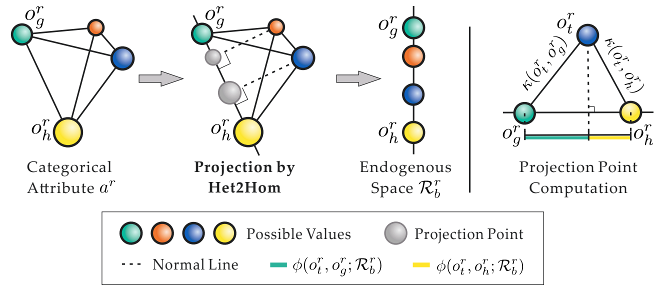

Figure 1: Diagram for projecting attribute values of a^r onto one of the endogenous spaces (i.e. $\mathcal{R}_b^r$) spanned by o_g^r and o_h^r.

attribute with v^r possible values, there are $\gamma^r = v^r(v^r - 1)/2$ endogenous spaces in total. Since each numerical attribute has only one endogenous space, i.e., its original space, we have $\gamma^r = 1$ when $r > d^c$. All the endogenous spaces corresponding to a^r is denoted as $\mathcal{R}^r = \{\mathcal{R}_1^r, \mathcal{R}_2^r, ..., \mathcal{R}_{\gamma^r}^r\}$.

To perform the above-mentioned projection, locations of all the attribute values in the original space should be known in advance. The relative locations of attribute values are indicated by their base distance

$$\kappa(o_g^r, o_h^r) = \sum_{s=1}^{d^c} \sum_{u=1}^{v^s} |p(o_u^s | o_g^r) - p(o_u^s | o_h^r)|, \tag{4}$$

which is the total difference between Conditional Probability Distributions (CPDs) obtained from a^ss as given o_g^r and o_h^r. Such a distance definition has been commonly adopted by most metrics that consider the inter-dependence of attributes, e.g., [Ienco *et al.*, 2012] and [Jian *et al.*, 2018a]. If the processed data set contains both nominal and ordinal attributes, the distance metric proposed in [Zhang and Cheung, 2021], which is a generalized version of the distance in Eq. (4), can be utilized instead. Based on $\kappa(\cdot,\cdot)$, relative location of the projected value o_t^r can be computed as

$$\phi(o_t^r, o_g^r; \mathcal{R}_b^r) = \frac{|\kappa(o_t^r, o_g^r)^2 - \kappa(o_t^r, o_h^r)^2 + \kappa(o_g^r, o_h^r)^2|}{2\kappa(o_g^r, o_h^r)} \tag{5}$$

where $\phi(o_t^r, o_g^r; \mathcal{R}_b^r)$ is the distance between o_g^r and the projection point of o_t^r in the space $\mathcal{R}_b^r$ spanned by o_g^r and o_h^r, and such a formula in Eq. (5) is obtained by simply applying the Pythagorean theorem as shown in the "Projection Point Computation" part of Figure 1. After the projection, since all the values are linearly arranged in $\mathcal{R}_b^r$, the distance between any pair of possible values o_t^r and o_f^r is computed by

$$\phi(o_t^r, o_f^r; \mathcal{R}_b^r) = |\phi(o_t^r, o_g^r; \mathcal{R}_b^r) - \phi(o_f^r, o_g^r; \mathcal{R}_b^r)| \tag{6}$$

if the projection points of o_t^r and o_f^r are on the same side of o_g^r in $\mathcal{R}_b^r$. Otherwise, the distance is computed by

$$\phi(o_t^r, o_f^r; \mathcal{R}_b^r) = \phi(o_t^r, o_g^r; \mathcal{R}_b^r) + \phi(o_f^r, o_g^r; \mathcal{R}_b^r). \tag{7}$$

Such a projection-based attributes representation is compared with the conventional methods in Figure 2. Our representation provides a homogeneous basis for connecting numerical and categorical attributes. After the representation,

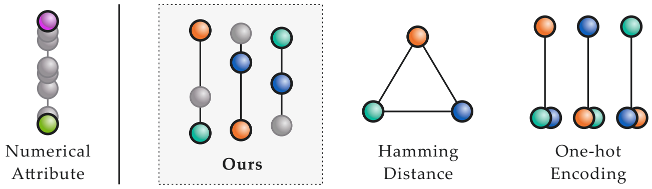

Figure 2: Comparison of the represented distance spaces.

$\phi(x_i^r, m_l^r) \cdot w^r$ in Eq. (3) can be replaced with the linear combination of the distances between x_i^r and m_l^r represented by different endogenous spaces derived from a^r:

$$\phi(x_i^r, m_l^r) \cdot w^r = \sum_{b=1}^{\gamma^r} \phi(x_i^r, m_l^r; \mathcal{R}_b^r) \cdot w_b^r \tag{8}$$

where w_b^r is the importance of the endogenous space $\mathcal{R}_b^r$.

3.3 Learning Algorithm

With Eq. (8), the objective function in Eq. (1) is rewritten as

$$z(\mathbf{Q}, M, \mathcal{W}) = \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il} \sum_{r=1}^{d} \sum_{b=1}^{\gamma^r} \phi(x_i^r, m_l^r; \mathcal{R}_b^r) \cdot w_b^r \tag{9}$$

where the original W is replaced by $\mathcal{W} = \{W^1, W^2, ..., W^d\}$, and $W^r = \{w_1^r, w_2^r, ..., w_{\gamma^r}^r\}$ stores the weights corresponding to the endogenous spaces in $\mathcal{R}^r$. Optimizing under such an objective function facilitates the learning of both data object partition and space linear combination (i.e. representation). We iteratively update the partition $\mathbf{Q}$, cluster descriptor M, and weights $\mathcal{W}$, which can be summarized into three steps: (1) Fix $\mathcal{W}$ and M, compute $\mathbf{Q}$; (2) Fix $\mathcal{W}$ and $\mathbf{Q}$, compute M; (3) Repeat (1) and (2) until convergence, fix $\mathbf{Q}$ and M, update $\mathcal{W}$. These three steps are repeated until the value of $z(\mathbf{Q}, M, \mathcal{W})$ is minimized.

We follow the conventional ways for computing $\mathbf{Q}$ and M as discussed in Section 3.1. The core difficulty is how to appropriately update $\mathcal{W}$, because the updates of the weights in W^r are highly cross-coupled due to the common attribute values shared by the spaces in $\mathcal{R}^r$. More specifically, if $\mathcal{W}$ is directly computed using Lagrangian multiplier method or updated in a normal gradient-decent way in the above-mentioned step (3), the update effect of different weights will somehow offset each other in the next step (1), which can easily lead to a corrupt results, especially when the number of endogenous spaces is large. Accordingly, a strategy is designed to select one weight from each W^r in step (3). Specifically, w_b^r is selected out from W^r by

$$b = \arg\min_y \frac{\sum_{l=1}^{k} \sum_{t=1}^{\gamma^r} |\epsilon_y^r(l) - \epsilon_t^r(l)|}{\sum_{l=1}^{k} e_y^r(l)} \tag{10}$$

where $e_y^r(l)$ is the total error contributed by endogenous space $\mathcal{R}_y^r$ on cluster c_l:

$$e_y^r(l) = \sum_{i=1}^{n} q_{il} \cdot \phi(x_i^r, m_l^r; \mathcal{R}_y^r). \tag{11}$$

$\epsilon_y^r(l)$ and $\epsilon_t^r(l)$ are computed through $\epsilon_y^r(l) = e_y^r(l) / \sum_{j=1}^{\gamma^r} e_j^r(l)$ and $\epsilon_t^r(l) = e_t^r(l) / \sum_{j=1}^{\gamma^r} e_j^r(l)$, respectively, to ensure that the values of $\sum_{t=1}^{\gamma^r} |\epsilon_y^r(l) - \epsilon_t^r(l)|$ on different clusters c_l are comparable, because clusters may have different numbers of data objects.

Remark 1. *Given an attribute a^r, the numerator of Eq. (10) quantifies the overall difference between $\epsilon_y^r(l)$ yielded by $\mathcal{R}_y^r$ and the rest $\epsilon_t^r(l)$s yielded by their corresponding $\mathcal{R}_t^r$s from the perspective of error contribution to $z(\boldsymbol{Q}, M, \mathcal{W})$. Therefore, a smaller numerator reflects that the space $\mathcal{R}_y^r$ is more representative among all the γ^r endogenous spaces of a^r, and updating the corresponding weight w_y^r is expected to yield a more effective reduction on $z(\boldsymbol{Q}, M, \mathcal{W})$.*

Remark 2. *Given an attribute a^r, the denominator of Eq. (10) computes the total error contributed by $\mathcal{R}_y^r$ to $z(\boldsymbol{Q}, M, \mathcal{W})$, which reflects the expected effectiveness of updating w_y^r in reducing $z(\boldsymbol{Q}, M, \mathcal{W})$. Therefore, w_y^r corresponding to a larger $\sum_{l=1}^{k} e_y^r(l)$ is preferred for updating to achieve a more effective reduction on $z(\boldsymbol{Q}, M, \mathcal{W})$.*

All the selected weights are updated by a small step by:

$$\begin{aligned} w_b^{r(new)} &= \max(0, w_b^r - \eta \cdot \frac{\partial z(\mathbf{Q}, M, \mathcal{W})}{\partial w_b^r}) \\ &= \max(0, w_b^r - \eta \cdot \sum_{l=1}^{k} e_b^r(l)). \end{aligned} \tag{12}$$

where η is the learning rate.

Remark 3. *To facilitate a stable learning process, the total weight is always fixed to 1 by $\sum_{r=1}^{d} \sum_{b=1}^{\gamma^r} w_b^r = 1$. Since numerical attributes are with well-defined distance space, the weight of each numerical attribute is fixed at $1/d$. Accordingly, the weights of the endogenous spaces derived from the categorical attributes are uniformly initialized by $w_b^r = d^c / (d \sum_{r=1}^{d^c} \gamma^r)$, where $\sum_{r=1}^{d^c} \gamma^r$ is the total number of the derived endogenous spaces of categorical attributes.*

The whole Het2Hom learning algorithm is summarized in **Algorithm 1**. The measure yielded by Het2Hom learning is a metric, and the learning algorithm is computationally efficient. The corresponding theoretical analysis has been provided in the Supplementary Material[1].

4 Experiments

Experiments have been designed to evaluate the performance of the proposed Het2Hom. Detailed experimental settings and complementary experimental results are provided in the Supplementary Material[1].

4.1 Experimental Settings

Experimental settings are briefly introduced below.

4 + 2 Experiments have been conducted. The four core experiments presented in this paper are (1) clustering performance evaluation, (2) significance study, (3) ablation study,

[1]https://drive.google.com/file/d/1SKNYxutdftgEfDK9CxzZLYkzYet4Jzf8/view?usp=sharing

Algorithm 1 Het2Hom learning for mixed data clustering

Input: Dataset S, number of sought clusters k, learning rate η, stop threshold β
Output: Partition $\mathbf{Q}$, weights $\mathcal{W}$

Project all the values of categorical attributes according to Eq. (4) and (5)
Initialize M by randomly selecting k objects from S, initialize $\mathbf{Q}$ by setting all its values to 0, initialize $\mathcal{W}$ according to Remark 3, set $\theta \to 0$ and $is\_conv = 0$
while $is\_conv = 0$ **do**
  Fix $\mathcal{W}$ and M, compute $\mathbf{Q}^{(new)}$
  if $\mathbf{Q}^{(new)} \neq \mathbf{Q}$ **then**
    Fix $\mathcal{W}$ and $\mathbf{Q}$, compute $M^{(new)}$
  else
    Fix $\mathbf{Q}$ and M, compute $\mathcal{W}^{(new)}$ according to Eq. (10)-(12)
    if $|z(\mathbf{Q}, M, \mathcal{W}) - \theta|/\theta < \beta$ **then**
      Set $is\_conv = 1$
    else
      Set $\theta = z(\mathbf{Q}, M, \mathcal{W})$
    end if
  end if
end while

No.	Data Set	Abbrev.	d^c	d^u	n	k^*
1	Soybean (Large)	SB	35	0	266	15
2	Solar Flare	SF	9	0	323	6
3	Zoo	ZO	15	0	101	7
4	Congressional Voting	VT	16	0	435	2
5	Tic-Tac-Toe	TT	9	0	958	2
6	Mushroom	MR	21	0	8124	2
7	Breast Cancer	BC	5	4	286	2
8	Hayes-Roth	HR	2	2	132	3
9	Lenses	LS	2	2	24	3
10	Lymphography	LG	15	3	148	4
11	Assistant Evaluation	AE	2	2	72	3
12	Fruit Evaluation	FT	2	3	100	5
13	Inflammations Diagnosis	DS	5	1	120	2
14	Heart Failure	HF	5	7	299	2
15	Autism-Adolescent	AA	7	2	104	2
16	Amphibians	AP	12	2	189	2
17	Mammographic	MM	4	1	961	2

Table 2: Statistics of the 17 data sets. d^c, d^u, and n are the numbers of categorical attributes, numerical attributes, and objects, respectively. k^* is the true number of clusters and we set $k = k^*$ here.

and (4) visualization of cluster discrimination ability. The two complementary experiments provided in the Supplementary Material[1] are (i) evaluation of convergence and execution time, and (ii) study of the parameter (i.e., η and β) effects.

9 Counterparts have been compared. One-Hot Encoding (OHE) combined with k-means is chosen because it is a common practical solution for mixed data clustering. Six other counterparts proposed in recent years, including Structure-Based Categorical data encoding (SBC) [Qian *et al.*, 2015], Jia's Distance Metric (JDM) [Jia *et al.*, 2016], Coupled Metric Similarity (CMS) [Jian *et al.*, 2018a], Unified Distance Metric (UDM) [Zhang and Cheung, 2022],

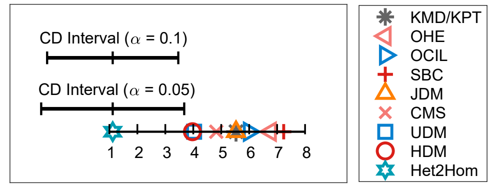

Figure 3: Results of the two-tailed BD tests w.r.t. the CA performance of different clustering approaches shown in Table 3.

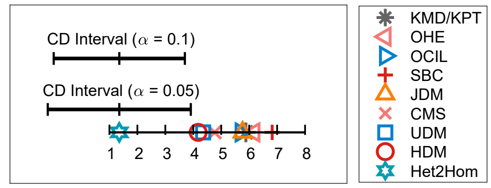

Figure 4: Results of the two-tailed BD tests w.r.t. the ARI performance of different clustering approaches shown in Table 4.

and Homogeneous Distance Metric (HDM) [Zhang and Cheung, 2021] combined with k-modes (KMD) [Huang, 1998] and k-prototypes (KPT) [Huang, 1997] according to the attribute composition of data sets, and Object-Cluster Iterative Learning (OCIL) [Cheung and Jia, 2013], have been selected, where CMS, UDM and HDM are the state-of-the-arts. Two conventional clustering algorithms, i.e., the original versions of KMD and KPT, are also compared.

17 Real Data Sets have been utilized for the experiments, and their statistics are shown in Table 2. All the data sets are obtained from the UCI machine learning repository, except FT from [Zhang and Cheung, 2022] and AE from [Zhang and Cheung, 2021].

4 Validity Indices have been chosen, including Clustering Accuracy (CA) [He *et al.*, 2005], the more discriminative Adjusted Rand Index (ARI) [Gates and Ahn, 2017] (value range [-1,1]), and the Normalized Mutual Information (NMI) [Estévez *et al.*, 2009]. For all these three indices, a larger value indicates better clustering performance. Bonferroni-Dunn (BD) test with computed Critical Difference (CD) interval [Demšar, 2006] is utilized for significance test. NMI results are provided in the Supplementary Material[1].

4.2 Clustering Performance Evaluation

Clustering performance evaluated by CA and ARI has been reported in Table 3 and 4, respectively. Results of KMD for categorical data and KPT for mixed data are combined into the same column for compactness. The observations are: (1) Het2Hom performs the best on almost all the data sets, which indicates its superiority in clustering. (2) Although Het2Hom does not significantly outperform the second-best approaches on TT, AE, and AP data sets, the second-best one differs on these data sets while Het2Hom always performs the best on them. (3) On some data sets, e.g., VT and AA, Het2Hom does not perform the best, but the gaps between Het2Hom and the best-performing counterparts are always tiny (less than 0.01 for both CA and ARI) on these data sets, which still demonstrates the competitiveness of Het2Hom.

Data	KMD/KPT	OHE	OCIL	SBC	JDM	CMS	UDM	HDM	Het2Hom
SB	0.486±0.03	0.529±0.02	0.538±0.04	0.535±0.03	0.529±0.04	0.513±0.03	0.528±0.03	0.531±0.04	**0.546±0.03**
SF	0.502±0.04	0.466±0.04	0.477±0.05	0.431±0.02	0.418±0.02	0.530±0.04	0.526±0.05	0.533±0.05	**0.591±0.03**
ZO	0.679±0.05	0.660±0.05	0.687±0.06	0.656±0.05	0.679±0.05	0.685±0.04	0.674±0.04	0.674±0.05	**0.884±0.03**
VT	0.862±0.01	0.792±0.05	**0.876±0.00**	0.788±0.02	0.869±0.00	0.867±0.00	0.869±0.00	0.869±0.00	0.874±0.00
TT	0.561±0.05	0.552±0.04	0.503±0.17	0.328±0.04	0.562±0.04	0.551±0.04	0.548±0.04	0.548±0.04	**0.569±0.04**
MR	0.811±0.09	0.706±0.14	0.844±0.18	0.511±0.12	0.724±0.22	0.736±0.11	0.731±0.12	0.695±0.12	**0.872±0.09**
BC	0.535±0.01	0.194±0.27	0.511±0.00	0.348±0.34	0.510±0.07	0.502±0.11	0.568±0.19	0.580±0.12	**0.634±0.09**
HR	0.370±0.02	0.424±0.14	0.358±0.03	0.346±0.01	0.383±0.03	0.382±0.02	0.404±0.03	0.405±0.03	**0.466±0.03**
LS	0.524±0.06	0.480±0.10	0.555±0.07	0.547±0.15	0.547±0.07	0.502±0.07	0.575±0.09	0.575±0.09	**0.602±0.12**
LG	0.592±0.14	0.299±0.37	0.582±0.13	0.315±0.28	0.638±0.03	0.619±0.08	0.600±0.10	0.587±0.14	**0.696±0.01**
AE	0.535±0.06	0.507±0.19	0.534±0.08	0.501±0.13	0.524±0.07	0.556±0.07	0.618±0.09	0.618±0.08	**0.620±0.05**
FT	0.468±0.04	0.550±0.04	0.504±0.04	0.536±0.03	0.461±0.05	0.528±0.05	0.550±0.04	0.556±0.04	**0.597±0.05**
DS	0.725±0.12	0.708±0.13	0.579±0.22	0.668±0.11	0.691±0.10	0.772±0.14	0.743±0.11	0.743±0.11	**0.799±0.12**
HF	0.614±0.06	0.543±0.03	0.409±0.23	0.524±0.02	0.548±0.03	0.628±0.06	0.600±0.06	0.600±0.06	**0.644±0.00**
AA	0.535±0.03	0.526±0.02	0.490±0.10	0.517±0.01	0.541±0.04	0.552±0.03	**0.567±0.03**	0.553±0.03	0.560±0.00
AP	0.533±0.02	0.542±0.01	0.531±0.14	0.546±0.00	0.549±0.02	0.533±0.03	0.555±0.01	0.553±0.01	**0.565±0.01**
MM	0.808±0.06	0.759±0.13	0.759±0.23	0.824±0.00	0.787±0.11	0.810±0.06	0.808±0.04	0.817±0.00	**0.831±0.00**
$\overline{AR}$	5.53	6.76	6.00	7.24	5.53	4.82	4.03	3.97	1.12

Table 3: Clustering performance evaluated by CA. "$\overline{AR}$" row reports the average performance ranks.

Data	KMD/KPT	OHE	OCIL	SBC	JDM	CMS	UDM	HDM	Het2Hom
SB	0.315±0.03	0.407±0.03	0.403±0.04	0.388±0.02	0.393±0.03	0.345±0.03	0.379±0.03	0.381±0.03	**0.416±0.03**
SF	0.260±0.05	0.225±0.05	0.234±0.06	0.167±0.03	0.144±0.02	0.331±0.06	0.333±0.07	0.335±0.06	**0.433±0.04**
ZO	0.619±0.05	0.594±0.05	0.644±0.05	0.586±0.05	0.639±0.04	0.637±0.04	0.622±0.04	0.622±0.04	**0.935±0.03**
VT	0.523±0.02	0.520±0.11	**0.565±0.01**	0.508±0.06	0.545±0.00	0.539±0.01	0.545±0.01	0.543±0.01	0.557±0.00
TT	0.023±0.04	0.011±0.02	0.015±0.02	0.018±0.02	0.022±0.03	0.015±0.02	0.015±0.02	0.015±0.03	**0.023±0.02**
MR	0.421±0.19	0.242±0.23	0.564±0.16	0.345±0.24	0.307±0.10	0.491±0.21	0.508±0.22	0.480±0.23	**0.585±0.15**
BC	-0.002±0.00	0.006±0.02	-0.003±0.00	0.060±0.07	-0.001±0.00	0.001±0.01	0.062±0.06	0.047±0.06	**0.085±0.09**
HR	-0.010±0.00	**0.062±0.03**	-0.011±0.01	-0.014±0.00	-0.005±0.01	-0.005±0.01	0.007±0.02	0.008±0.02	0.059±0.02
LS	0.069±0.08	0.053±0.12	0.119±0.11	0.143±0.15	0.117±0.10	0.054±0.09	0.239±0.13	0.239±0.13	**0.277±0.19**
LG	0.070±0.08	0.094±0.12	0.051±0.06	0.005±0.00	0.074±0.03	0.073±0.07	0.051±0.05	0.057±0.07	**0.149±0.01**
AE	0.125±0.06	0.140±0.06	0.123±0.09	0.115±0.12	0.104±0.06	0.173±0.07	0.268±0.10	0.270±0.09	**0.281±0.04**
FT	0.202±0.05	0.334±0.03	0.255±0.05	0.282±0.02	0.188±0.05	0.259±0.04	0.296±0.04	0.297±0.03	**0.366±0.05**
DS	0.255±0.24	0.230±0.28	0.105±0.14	0.153±0.17	0.178±0.17	0.365±0.30	0.277±0.22	0.277±0.22	**0.405±0.25**
HF	0.060±0.06	0.001±0.01	0.001±0.01	-0.003±0.00	0.002±0.01	0.072±0.06	0.044±0.05	0.044±0.05	**0.078±0.00**
AA	-0.003±0.01	-0.006±0.01	-0.009±0.00	-0.011±0.00	-0.003±0.01	0.003±0.01	**0.009±0.02**	0.001±0.01	0.000±0.01
AP	-0.001±0.01	-0.005±0.00	**0.002±0.00**	-0.002±0.00	-0.001±0.01	-0.002±0.01	-0.002±0.01	-0.001±0.01	0.000±0.01
MM	0.394±0.08	0.335±0.18	0.389±0.12	0.419±0.00	0.380±0.15	0.397±0.08	0.387±0.06	0.401±0.00	**0.438±0.00**
$\overline{AR}$	5.88	6.18	5.71	6.82	5.76	4.76	4.35	4.18	1.35

Table 4: Clustering performance evaluated by ARI (with range [-1,1]). "$\overline{AR}$" row reports the average performance ranks.

4.3 Significance Study

Results of the two-tailed BD test [Demšar, 2006] at confidence intervals 0.95 ($\alpha = 0.05$) and 0.9 ($\alpha = 0.1$) are shown in Figure 3 and 4. According to [Demšar, 2006], performance of Het2Hom is considered to be significantly better than that of all the counterparts outside the right bound of CD intervals. It can be observed from Figure 3 and 4 that Het2Hom performs significantly better than all nine counterparts.

4.4 Ablation Study

To explicitly illustrate the effectiveness of the core components of Het2Hom, several variants of Het2Hom are formed for comparison. The version of Het2Hom that only conducts the Projection-Based Representation (PBR) without representation learning is formed. The version further removes the PBR module and only adopts the Difference of CPDs (DCPDs) defined by Eq. (4) for clustering is also compared. Performance of the original KMD/KPT is also reported for completeness. It can be observed from the last sub-figure in Figure 5 that the average performance ranks of KMD/KPT, DCPDs, PBR, and Het2Hom are around 4, 3, 2, and 1, respectively, which intuitively verifies the effectiveness of the core components of Het2Hom. More specifically, the superiority of Het2Hom over PBR indicates the correctness of the learning strategy proposed in Section 3.3. PBR outperforms DCPDs, which proves the effectiveness of the projection mechanism presented in Section 3.2. DCPDs performs better than KMD/KPT, which indicates the reasonableness of adopting the distance defined by Eq. (4) as the base distance for conducting the PBR.

4.5 Visualization

In Figure. 6, t-SNE [Maaten and Hinton, 2008] is utilized to demonstrate the cluster discrimination ability of Het2Hom.

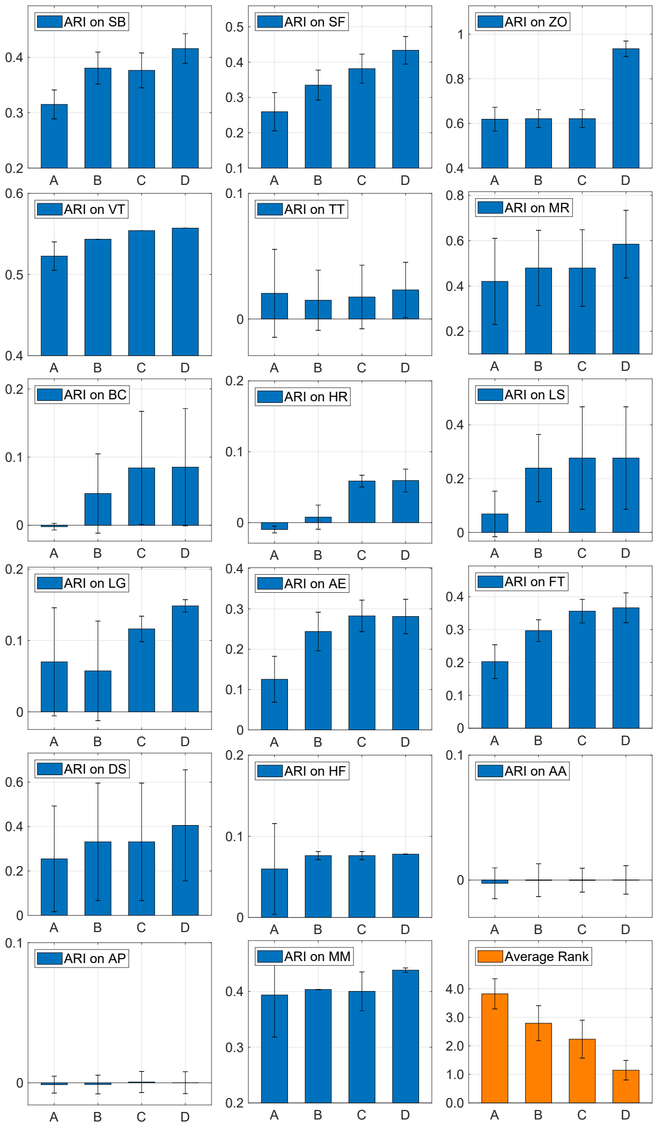

Figure 5: Clustering performance of KMD/KPT, DCPDs, PBR, and Het2Hom (denoted by A, B, C, and D, respectively) on all the 17 data sets. The last sub-figure summarizes the performance ranking of the compared approaches.

We first encode the attributes of the MR data set using OHE, PBR, and Het2Hom. Then the encoded data set is processed by t-SNE into two-dimensional, and visualized by marking the true labels of objects in different colors. It can be observed that the cluster discrimination ability of Het2Hom is obviously stronger than that of PBR and OHE.

5 Conclusion

In this paper, we have proposed Het2Hom, which is composed of a projection-based representation mechanism and a representation learning module, for mixed data clustering. It projects values of categorical attributes onto all the possible endogenous spaces to produce informative representations. Since these elaborately obtained representations are homogeneous with the numerical attributes, attribute representations and the partition of data objects can thus be learned to more appropriately adapt to each other. It turns out that Het2Hom exploits the information more sufficiently based on the established connection between categorical and numerical attributes, and thus achieves a more accurate clustering. Moreover, the intuitive but novel geometry-based projection makes the represented data highly interpretable. Extensive experiments have shown the efficacy of Het2Hom.

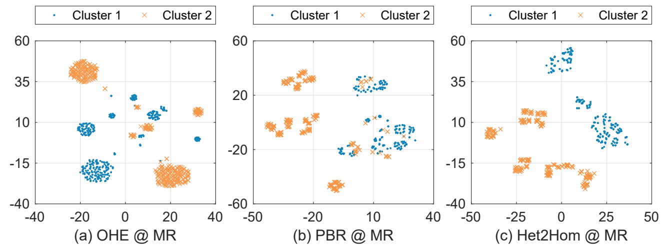

Figure 6: t-SNE visualization of the MR data sets represented by OHE, PBR, and Het2Hom.

Acknowledgements

This work was supported in part by the NSFC under grant 62102097, NSFC/RGC Joint Research Scheme under grant N_HKBU214/21, RGC General Research Fund under grant 12201321, HKBU grants: RC-FNRA-IG/18-19/SCI/03 and RC-IRCMs/18-19/SCI/01, the Key-Area R&D Program of Guangdong Province under grant 2021B0101220006, the Guangdong Basic and Applied Basic Research Foundation under grants: 2022A1515011592 and 2021A1515012300, and the Science and Technology Planning Project of Guangdong Province under grant 2019A050510041.

References

[Agresti, 2003] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2003.

[Ahmad and Dey, 2007] Amir Ahmad and Lipika Dey. 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, 2007.

[Ball and Hall, 1967] Geoffrey H Ball and David J Hall. A clustering technique for summarizing multivariate data. *Behavioral Science*, 12(2):153–155, 1967.

[Boriah *et al.*, 2008] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254, 2008.

[Cheung and Jia, 2013] Yiu-ming Cheung and Hong Jia. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8):2228–2238, 2013.

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

[dos Santos and Zárate, 2015] Tiago RL dos Santos and Luis E Zárate. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3):1247–1260, 2015.

[Estévez *et al.*, 2009] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[Gates and Ahn, 2017] Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *The Journal of Machine Learning Research*, 18(1):3049–3076, 2017.

[Goodall, 1966] David W Goodall. A new similarity index based on probability. *Biometrics*, pages 882–907, 1966.

[He *et al.*, 2005] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 507–514, 2005.

[Huang, 1997] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

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

[Ienco *et al.*, 2012] Dino Ienco, Ruggero G Pensa, and Rosa Meo. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1–25, 2012.

[Jia *et al.*, 2016] Hong Jia, Yiu-ming Cheung, and Jiming Liu. A new distance metric for unsupervised learning of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 27(5):1065–1079, 2016.

[Jian *et al.*, 2017] Songlei Jian, Longbing Cao, Guansong Pang, Kai Lu, and Hang Gao. Embedding-based representation of categorical data by hierarchical value coupling learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1937–1943, 2017.

[Jian *et al.*, 2018a] Songlei Jian, Longbing Cao, Kai Lu, and Hang Gao. Unsupervised coupled metric similarity for non-iid categorical data. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1810–1823, 2018.

[Jian *et al.*, 2018b] Songlei Jian, Guansong Pang, Longbing Cao, Kai Lu, and Hang Gao. Cure: Flexible categorical data representation by hierarchical coupling learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):853–866, 2018.

[Kacem *et al.*, 2015] Mohamed Aymen Ben Haj Kacem, Chiheb-Eddine Ben N'cir, and Nadia Essoussi. Mapreduce-based k-prototypes clustering method for big data. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics*, pages 1–7. IEEE, 2015.

[Le and Ho, 2005] Si Quang Le and Tu Bao Ho. An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 26(16):2549–2557, 2005.

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

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[Qian *et al.*, 2015] Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, and Chuangyin Dang. Space structure and clustering of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 27(10):2047–2059, 2015.

[Xu and Wunsch, 2005] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[Zhang and Cheung, 2018] Yiqun Zhang and Yiu-ming Cheung. Exploiting order information embedded in ordered categories for ordinal data clustering. In *Proceedings of the 24th International Symposium on Methodologies for Intelligent Systems*, pages 247–257, 2018.

[Zhang and Cheung, 2020] Yiqun Zhang and Yiu-ming Cheung. An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6869–6876, 2020.

[Zhang and Cheung, 2021] Yiqun Zhang and Yiu-ming Cheung. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page DOI: 10.1109/TPAMI.2021.3056510, 2021.

[Zhang and Cheung, 2022] Yiqun Zhang and Yiu-ming Cheung. A new distance metric exploiting heterogeneous inter-attribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Cybernetics*, 52(2):758–771, 2022.

[Zhang *et al.*, 2020] Yiqun Zhang, Yiu-ming Cheung, and Kaychen Tan. 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, 2020.

[Zhu *et al.*, 2020] Chengzhang Zhu, Qi Zhang, Longbing Cao, and Arman Abrahamyan. Mix2vec: Unsupervised mixed data representation. In *Proceedings of the 7th International Conference on Data Science and Advanced Analytics*, pages 118–127, 2020.

[Zhu *et al.*, 2022] Chengzhang Zhu, Longbing Cao, and Jianping Yin. Unsupervised heterogeneous coupling learning for categorical representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):553–549, 2022.