Graph-Based Dissimilarity Measurement for Cluster Analysis of Any-Type-Attributed Data

Yiqun Zhang, *Member, IEEE*, and Yiu-Ming Cheung, *Fellow, IEEE*

***Abstract*— Heterogeneous attribute data composed of attributes with different types of values are quite common in a variety of real-world applications. As data annotation is usually expensive, clustering has provided a promising way for processing unlabeled data, where the adopted similarity measure plays a key role in determining the clustering accuracy. However, it is a very challenging task to appropriately define the similarity between data objects with heterogeneous attributes because the values from heterogeneous attributes are generally with very different characteristics. Specifically, numerical attributes are with quantitative values, while categorical attributes are with qualitative values. Furthermore, categorical attributes can be categorized into nominal and ordinal ones according to the order information of their values. To circumvent the awkward gap among the heterogeneous attributes, this article will propose a new dissimilarity metric for cluster analysis of such data. We first study the connections among the heterogeneous attributes and build graph representations for them. Then, a metric is proposed, which computes the dissimilarities between attribute values under the guidance of the graph structures. Finally, we develop a new *k*-means-type clustering algorithm associated with this proposed metric. It turns out that the proposed method is competent to perform cluster analysis of datasets composed of an arbitrary combination of numerical, nominal, and ordinal attributes. Experimental results show its efficacy in comparison with its counterparts.**

***Index Terms*— Cluster analysis, dissimilarity measure, graph space, heterogeneous attributes, representation.**

I. Introduction

CLUSTER analysis is one of the most common methods for data analytics with a variety of applications, including knowledge acquisition from medical databases [1], data analysis of grading and evaluation systems [2], big data preprocessing [3], and so on. Under such circumstances, datasets are usually composed of different features that may have different types of values, e.g., numerical and categorical ones. In general, features with numerical values and categorical values are called numerical attributes and categorical attributes, respectively [4]. Compared with the numerical attributes with well-defined distances in the Euclidean space, similarities among the values of a categorical attribute are, however, not well-defined because the different possible values are more like divergent concepts which are hard to be explicitly located in the space. Moreover, categorical attributes can be further categorized into two subtypes, i.e., the nominal type and the ordinal type, which have extra relative orders between the attribute values [5], [6]. Datasets composed of the above-mentioned heterogeneous attributes (i.e., numerical, nominal, and ordinal attributes) are very common in real-world unsupervised learning tasks. Table I demonstrates two fragments of such datasets. The upper part is a fragment of the medical dataset, which contains the numerical attribute “SaO2,”[1] nominal attribute “Gender,” and ordinal attribute “GCS-E.”[2] The lower part is a fragment of the financial risk evaluation dataset, which also contains the three types of heterogeneous attributes. Such datasets are widespread, and the awkward gap among the heterogeneous attributes brings complexities to cluster analysis.

TABLE I

Fragments of a Medical Dataset (Upper Part) and a Financial Risk Evaluation Dataset (Lower Part)

ID	SaO2[1]	Gender	GCS-E[2]	...
Patient_1	0.23	male	none	...
Patient_2	0.88	female	spontaneous	...
Patient_3	0.72	female	to_pain	...
Patient_4	0.99	male	spontaneous	...
Patient_5	0.75	male	to_speech	...

ID	Income	Occupation	Credit_Rating	...
Client_1	4,000	driver	good	...
Client_2	8,600	freelance	neutral	...
Client_3	12,200	freelance	good	...
Client_4	11,800	doctor	good	...
Client_5	16,800	doctor	very_good	...

Most existing attempts for heterogeneous attribute data clustering focus on exploring more powerful similarity measures or data representation strategies. Both these two orientations have the same goal to achieve more reasonable similarity measurement, which is the basis for the success of cluster analysis. Early attempts usually conduct one-hot encoding to the values of categorical attributes [7] and then process the encoded data as numerical data. Similar attempts also

Manuscript received 18 August 2021; revised 12 January 2022, 27 March 2022, and 6 August 2022; accepted 22 August 2022. Date of publication 12 September 2022; date of current version 1 September 2023. This work was supported in part by the NSFC under Grant 62102097, in part by the NSFC and Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, in part by the General Research Fund of RGC under Grant 12201321, in part by Hong Kong Baptist University under Grant RC-FNRA-IG/18-19/SCI/03, in part by the Natural Science Foundation of Guangdong Province under Grant 2022A1515011592, and in part by the Science and Technology Program of Guangzhou under Grant 202201010548. *(Corresponding author: Yiu-Ming Cheung.)*

Yiqun Zhang is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: yqzhang@gdut.edu.cn).

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2022.3202700.

Digital Object Identifier 10.1109/TNNLS.2022.3202700

[1] SaO2 is the abbreviation of saturation of blood oxygen.

[2] GCS-E is the abbreviation of Glasgow coma scale—Eye-opening reaction. Symptom severity of the four possible values {none, to_pain, to_speech, spontaneous} descending successively.

include the well-known k-prototypes clustering algorithm and its variants [8], [9], which combine the Euclidean distances measured on numerical attributes and Hamming distances measured on categorical attributes using a tradeoff parameter. However, both the one-hot encoding and the Hamming distance uniformly assign distance "1" to any pair of different categorical values, which introduces an over-absolute similarity assumption that all pairs of different values have the identical dissimilarity degree. Furthermore, the similarity measure proposed in [10] quantifies the similarity between a data object and clusters in a unified probability framework for numerical and categorical attributes. Lin [11]'s similarity metric and entropy-based distance metric [12] further take into account the orders between possible values of an ordinal attribute and quantify the dissimilarity degrees between different possible values from the perspective of information theory. Both these two metrics apply to the datasets composed of nominal and ordinal attributes, while their key difference is that the latter unifies the similarity definitions for nominal and ordinal attributes to avoid information loss. Since all the above-mentioned measures treat attributes as independent of each other, which is not always the case in real datasets, the valuable information provided by the interdependence of attributes is, thus, wasted during similarity measurement.

In the literature, some works have been made in an effort to exploit the interdependence of attributes for categorical data clustering. They adopt a common basic idea that two possible values of an attribute are more similar if they have more identical corresponding values on the other attributes. For example, attribute context-based distance metric [16], [17], [18] quantifies the interdependence degree between categorical attributes and then selects attributes accordingly for providing statistical information [i.e., conditional probability distributions (CPDs)] for indicating similarity between possible values of an attribute. In the stream of data representation, approaches have been proposed to encode possible values as the CPDs of their corresponding values from different attributes [19], [20]. However, they have not discriminated against nominal and ordinal attributes, which somewhat leads to information loss. Recently, paper [13] has further studied the interdependence effect between heterogeneous nominal and ordinal attributes and proposed a unified distance metric together with an interdependence measure. Most recently, more advanced solutions, including Mix2Vec [14], homogeneous distance metric [15], and Het2Hom [21], have been proposed. Mix2Vec and Het2Hom represent numerical and categorical attribute values by sufficiently preserving their intrinsic structural information. However, their effectiveness still relies much on the adopted encoding strategies of categorical attributes. In contrast, the homogeneous distance metric [15] uniformly defines distances for nominal and ordinal attributes, but is inapplicable to the common numerical attributes.

In general, all the existing methods for heterogeneous attribute clustering suffer from one or both of the following two limitations: 1) datasets are composed of two specific types of attributes only and 2) the interdependence among different types of attributes has not been taken into account. Table II sorts out the applicability of the existing methods in terms of the types of heterogeneous attributes and the interdependence of heterogeneous attributes. There are a total of four types of heterogeneous attribute datasets, which are composed of: 1) n**U**merical plus **N**ominal attributes (U+N); 2) n**U**merical plus **O**rdinal attributes (U+O); 3) **N**ominal plus **O**rdinal attributes (N+O); and 4) all the three types of attributes (U+N+O). Since datasets composed of heterogeneous and interdependent attributes are common in the real world, it can be seen from Table II that a new method applicable to any-type-attributed data clustering (ADC) is still desired nowadays.

TABLE II

Applicability of the Existing Methods in Terms of Different Types of Heterogeneous Attribute Data (Indicated by "✓") and the Corresponding Interdependence Between Heterogeneous Attributes (Indicated by "‡")

Method	U+N	U+O	N+O	U+N+O
One-Hot Encoding	✓			
Euclidean + Hamming Distance [8]	✓			
Lin's Similarity Metric [11]			✓	
Object-Cluster Similarity [10]	✓			
Entropy-based Distance Metric [12]			✓	
Unified Distance Metric [13]			✓‡	
Mix2Vec Representation [14]	✓‡			
Homogeneous Distance Metric [15]			✓‡	
ADC (proposed)	✓‡	✓‡	✓‡	✓‡

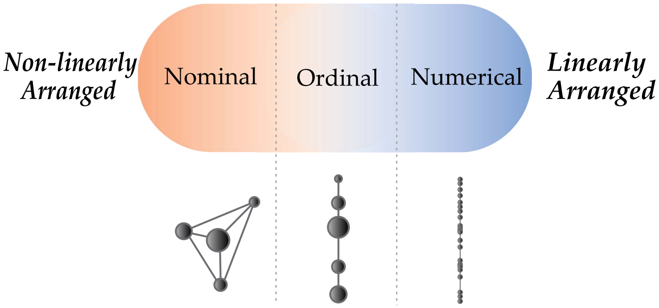

Fig. 1. Attribute types (upper part) and space structures (lower part).

To this end, this article will study the structures of the heterogeneous attributes and their connections, thereby treating them from a homogeneous perspective for clustering. Since most real-world data come in the form of graphs, constructing graph structures has been regarded as a reasonable way of studying the complex relationships among data objects by the recent related works [22], [23], [24]. Our previous work [15] has also successfully applied graphs to the modeling of intraattribute and interattribute relationships (i.e., dissimilarities among attribute values and dependencies among attributes) for ordinal attributes. This article further develops the advantages of graphs for data representation in a more challenging situation, where datasets are composed of an arbitrary combination of heterogeneous attributes. To achieve this, graphs are built for attributes, as shown in Fig. 1, according to the intrinsic attribute types, where nodes with different sizes stand for attribute values with different occurrence frequencies and edges reflect the similarity between attribute values. Numerical and ordinal attributes have line-like graph structures because their sortable values can be viewed as linearly arranged in 1-D spaces, while the values of a nominal attribute are nonlinearly arranged.

To use the constructed graphs, we represent a possible value from an attribute A^r using the probability distribution of its co-occurred values from the other attributes (e.g., A^s). Then, the dissimilarity between two values from A^r is considered as the length of the graph path between them, which is quantified as the difference between the corresponding representations computed according to A^s's graph structure. As a result, the graph-based dissimilarity quantification processes simultaneously preserve the information provided by the intrinsic structures and interdependence of the heterogeneous attributes. Subsequently, a new algorithm is presented for ADC, following the graph structures. Extensive experiments on the benchmark datasets have demonstrated the superiority of the proposed approach. The main contributions of this article are summarized in the following.

- Graph is introduced to homogeneously reveal the value space of numerical, nominal, and ordinal attributes. Based on the graph space representation, the interdependence effect of the heterogeneous attributes is studied to guide forming dissimilarity measures.
- A dissimilarity computation scheme is designed based on the graph structures to convert the dissimilarities of heterogeneous attributes into homogeneous quantities. Accordingly, a unified dissimilarity metric and an interattribute dependence measure are formed.
- Following the graph-based dissimilarity definition, a new k-means-type clustering algorithm is developed, which can iteratively update the heterogeneous-valued cluster prototypes in a unified way. The corresponding theoretical analysis is provided as well.

The remainder of this article is organized as follows. Section II makes an overview of the existing related techniques. In Section III, the graph space, the proposed dissimilarity metric, and the proposed clustering algorithm are presented in detail, together with the analysis and discussions. Then, extensive experiments are conducted in Section IV. Finally, we draw a conclusion in Section V.

II. Related Work

This section makes an overview of the existing related works, focusing on similarity measurement and data representation.

A. Similarity Measures

Euclidean distance and Hamming distance [25] are two conventional distance metrics that have been widely used for the cluster analysis of numerical and categorical data, respectively. For a single numerical attribute, Euclidean distance can appropriately indicate the dissimilarity between attribute values benefited from the well-defined Euclidean space, while for a categorical attribute without a well-defined space, Hamming distance that uniformly assigns distance 1 to any pair of different values demonstrates its limitations in distinguishing the dissimilarities of different value pairs. Therefore, more advanced categorical data similarity measures have been developed in the literature.

Given a target categorical attribute, extracting statistical information from its related attributes for reflecting the similarities has been acknowledged as a feasible solution for more reasonable similarity measurement. Different measures in this stream, including association-based [26], Ahmad and Dey's [27], and context-based [16], [17] measures, have been proposed. They adopt a similar idea to compute the distance between two CPDs obtained from the related attributes giving two values from the target attribute to indicate their dissimilarity. Among these measures, the context-based one can be viewed as the improved version of the former two because it further adopts a context selection module to filter the attributes with lower dependence on the target one. The recently proposed metrics [18], [28] can further reasonably exploit both the intraattribute and interattribute information, which avoids the failure of similarity measurement in the case that all the attributes are independent of each other. However, all the above-mentioned measures have not addressed the heterogeneity of nominal attributes and ordinal attributes, which are two subtypes of the categorical attribute.

Lin's similarity metric is an early attempt to appropriately define similarities for both nominal and ordinal attributes. It computes the entropy of values to indicate their similarity degree. By introducing the order of values for entropy computation, it can preserve the order information of ordinal attributes. The entropy-based distance metric [12], [29] further unifies the distance definition of nominal and ordinal attributes, and provides attributes weighting scheme for distance measurement. Since they have not exploited the information provided by the interattribute relationship, the unified distance metric [13] is proposed to extract useful relationship information for measuring the distances of nominal and ordinal attributes in a unified way. Most recently, homogeneous distance metric [15] is proposed based on the graph structure of attribute values and achieves superior clustering performance on the datasets composed of nominal and ordinal attributes.

However, for the widespread mixed data composed of numerical and categorical attributes, the gap between numerical and categorical attributes is more awkward than that between nominal and ordinal attributes. The conventional way for handling such mixed data is to compute the similarity between data objects contributed by numerical and categorical attributes using Euclidean distance and Hamming distance, respectively, and then combine them using a trade-off parameter [8]. However, the parameter cannot be easily decided and the Hamming distance may introduce unreasonableness for the distance measurement as discussed in the first paragraph of this section. The similarity measure proposed in [10] is a valuable attempt, which transforms the distances of numerical and categorical attributes into the unified probability form for measurement, and achieves better clustering performance than the conventional approaches. However, such a measure has not taken into account the relationship among attributes and is incapable of exploiting the order information of ordinal attributes, which limits its performance and application, respectively.

B. Data Representation Methods

Data representation refers to the methods of encoding the data values to represent the valuable data information and make the data convenient to process. For the problem of heterogeneous attribute data distance measurement, we can encode nominal attributes using one-hot encoding, encode ordinal attributes using the ranking of the possible values, and then treat the encoded data as numerical one for distance measurement. Such a straightforward and simple representation strategy is called numerical coding (NC), and has been widely adopted in real applications [7]. However, NC also suffers several defects, including the lack of theoretical support, loss of the interattribute relationship information, leading to the curse of dimensionality of nominal attributes, and so on. Some empirical studies [30] have already shown that the performance of NC is generally worse than the specially designed counterparts. Although it is a possible choice to ask domain experts to complete the encoding task, for large-scale, high-dimensional, and multivariate datasets, such an encoding solution is extremely laborious and nontrivial.

Therefore, automated representation methods have been proposed to encode the datasets in an unsupervised environment. Since numerical values already have well-defined distances, most existing attempts related to the problem of heterogeneous attribute data clustering focus on how to reasonably represent the categorical attributes. The one proposed in [31] encodes the attribute values according to the interobject dissimilarities of the original dataset, and thus, the interattribute relationship can be taken into account. Later, the representation method proposed in [19] and [32] encodes the original dataset by performing k-means clustering and PCA on intraattribute and interattribute couplings to capture more abundant information of categorical dataset. To ensure a more in-depth exploration of the useful information, a more powerful representation method proposed in [20] represents the dataset by using different types of value couplings learned by multiple kernel spaces and achieves superior clustering performance on categorical data. However, all the above-mentioned categorical data representation methods are designed for nominal data only.

Most recently, Mix2Vec [14] representation approach is proposed for the dataset composed of numerical and categorical attributes. It utilizes a deep model to vectorize the heterogeneous attribute values with sufficiently preserving the structural distribution information of data objects and performs better than the state-of-the-art deep [33] and statistical machine learning [34] representation methods. However, Mix2Vec still encodes categorical and numerical attributes in separate ways and has not addressed the heterogeneous problem of nominal and ordinal attributes. Moreover, although deep representation models are powerful in improving clustering accuracy (CA), their inherent weakness in terms of interpretability may somewhat limit their applications, especially for clustering-based data understanding, knowledge acquisition, and so on.

III. Proposed Method

In this section, we first formulate notations of data composition, dissimilarity measure, interattribute dependence, and objective function of heterogeneous attribute data clustering in Section III-A. Then, the spatial structures of heterogeneous attributes are constructed, based on which the dissimilarity metric and interdependence measure of this article are proposed and discussed in Section III-B. Following the new dissimilarity definition, the clustering algorithm corresponding to the objective function is derived and analyzed in Section III-C.

A. Problem Formulation

Table III sorts out the default notations and symbols used in this article. A heterogeneous attribute dataset S can be represented as a tuple $S =< X, A, O >$, where $X = \{\mathbf{x}_i | i = 1, 2, \ldots, n\}$ is the object set with n objects. $A = \{A^r | r = 1, 2, \ldots, d\}$ is the attribute set composed of d attributes, including $d^{\langle n\rangle}$ nominal, $d^{\langle o\rangle}$ ordinal, and $d^{\langle u\rangle}$ numerical attributes, where $d = d^{\langle n\rangle} + d^{\langle o\rangle} + d^{\langle u\rangle}$. Here, "n," "o," and "u" represent **n**ominal, **o**rdinal, and n**u**merical, respectively. For convenience but without loss of generality, we assume that the attributes are concentrated by type in the order of nominal, ordinal, and numerical attributes for all the datasets, and we have

$$\text{type}(A^r) = \begin{cases} \text{nominal}, & r \le d^{\langle n\rangle} \\ \text{ordinal}, & d^{\langle n\rangle} < r \le d^{\langle n\rangle} + d^{\langle o\rangle} \\ \text{numerical}, & d^{\langle n\rangle} + d^{\langle o\rangle} < r \le d \end{cases} \tag{1}$$

with $r \in \{1, 2, \ldots, d\}$. Formally, we have $A = A^{\langle n\rangle} \cup A^{\langle o\rangle} \cup A^{\langle u\rangle}$, where $A^{\langle n\rangle} = \{A^r | r = 1, 2, \ldots, d^{\langle n\rangle}\}$, $A^{\langle o\rangle} = \{A^r | r = d^{\langle n\rangle} + 1, d^{\langle n\rangle} + 2, \ldots, d^{\langle n\rangle} + d^{\langle o\rangle}\}$, and $A^{\langle u\rangle} = \{A^r | r = d^{\langle n\rangle} + d^{\langle o\rangle} + 1, d^{\langle n\rangle} + d^{\langle o\rangle} + 2, \ldots, d\}$. Corresponding to the attribute set A, $O = \{O^r | r = 1, 2, \ldots, d\}$ is the collection of unique value sets of each attribute, where $O^r = \{o_1^r, o_2^r, \ldots, o_{v^r}^r\}$ is the unique value set of A^r. For nominal and ordinal attributes, v^r is the number of unique values, which is usually a small integer (i.e., $1 \le v^r \ll n$) because the attribute values fall in a limited number of possible values, while v^r is a relatively larger inter (i.e., $1 \le v^r \le n$) for numerical attributes as they have infinite possible values from the real space $\mathcal{R}$. Moreover, compared with nominal attributes, the values in O^r of an ordinal or numerical attribute have an additional order relationship $o_1^r \succ o_2^r \succ \cdots \succ o_{v^r}^r$, where the symbol "$\succ$" indicates that the values on its left rank higher than the values on its right. An intuitive example can be found in Table I that the values of the GCS-E attribute have order relationship: none $\succ$ to_pain $\succ$ to_speech $\succ$ spontaneous. The above-mentioned analysis of O^r of the three types of attributes once again reflects their heterogeneity.

TABLE III

Notations and Symbols Used in This Article

Notations and Symbols	Explanations
Subscript, e.g. "i" of $\mathbf{x}_i$	Element index in a set
Superscript, e.g. "r" of A^r and O^r	Attribute index
Angle-bracketed superscript, e.g. "$\langle o\rangle$" of $d^{\langle o\rangle}$	Annotation
Parentheses, e.g. $dissim(\cdot,\cdot)$	Function
Uppercase, e.g. X, A, and O	Set
Uppercase, bold font, e.g. $\mathbf{Q}$	Matrix
Uppercase, calligraphic font, e.g. $\mathcal{R}$	Space
Lowercase, e.g. d, n, and o_1^r	Value
Lowercase, bold font, e.g. $\mathbf{x}_i$ and $\mathbf{h}_l$	Vector
$\succ$, e.g. $o_1^r \succ o_2^r$	Ranks higher than
$\top$, e.g. $[x_i^1, x_i^2, ..., x_i^d]^\top$,)	Transpose

The ith object of X is a vector $\mathbf{x}_i = [x_i^1, x_i^2, \ldots, x_i^d]^\top$ consists of d values from the d attributes. In partitional clustering task, X should be divided into k clusters $C = \{C_l | l = 1, 2, \ldots, k\}$, which are expressed as a collection of k disjoint subsets of X, where C_l is the set of objects in the lth cluster, and we have $X = \bigcup_{l=1}^k C_l$. To represent each cluster, a $k \times d$ matrix $\mathbf{B}$ is maintained during clustering, and each row $\mathbf{b}_l$ of $\mathbf{B}$ is a d value vector $\mathbf{b}_l = [b_l^1, b_l^2, \ldots, b_l^d]^\top$ representing the lth cluster C_l. An $n \times k$ matrix $\mathbf{Q}$ indicating which of the k cluster the n objects belong to is usually maintained for clustering. The (i, l)th entry of $\mathbf{Q}$ is denoted as q_{il}, and its value is computed by

$$q_{il} = \begin{cases} 1, & \text{if } l = \arg\min_y \text{dissim}(\mathbf{x}_i, \mathbf{b}_y) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

if dissimilar function dissim$(\cdot,\cdot)$ is adopted to evaluate which cluster an object should be assigned to. According to (2), we have $\sum_{l=1}^k q_{il} = 1$, $q_{il} \in \{0, 1\}$. For heterogeneous attribute data, how to define appropriate $dissim(\mathbf{x}_i, \mathbf{b}_l)$ for computing the dissimilarities between objects and clusters composed of heterogeneous attribute values is a more challenging problem, which has attracted much more attention than clustering algorithm. As discussed in the penultimate paragraph of Section I, our goal is to define an appropriate dissimilar metric that can exploit interattribute relationship information for quantifying the dissimilarities of any type of attribute in a homogeneous way. Following the conventional dissimilar definitions that considers the interattribute relationship, $dissim(\mathbf{x}_i, \mathbf{b}_l)$ can be written as

$$\mathrm{dissim}(\mathbf{x}_i, \mathbf{b}_l) = \sum_{r=1}^{d}\sum_{s=1}^{d} \mathrm{dissim}^{\mathrm{rs}}\left(x_i^r, b_l^r\right) \cdot w^{\mathrm{rs}} \tag{3}$$

where $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$ denotes dissimilarity between x_i^r and h_l^r measured according to the information provided by attribute A^s. The weight w^{rs} [i.e., the (r, s)th entry of weight matrix $\mathbf{W}$] here is used to control the contributions of different attributes (i.e., A^s) in forming $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$. Most existing works [15], [16], and [17] hold that if two attributes are more interdependent, then one attribute can provide a more reliable indication in computing the dissimilarities of the other attribute and vice versa. Therefore, w^{rs} is quantified as the interdependence degree of A^r and A^s, which will be discussed in detail in Section III-B.

Then, we define the clustering problem following the k-means-type algorithm as minimizing

$$z = \sum_{i=1}^{n}\sum_{l=1}^{k} q_{\mathrm{il}} \sum_{r=1}^{d}\sum_{s=1}^{d} \mathrm{dissim}^{\mathrm{rs}}\left(x_i^r, b_l^r\right) \cdot w^{\mathrm{rs}}. \tag{4}$$

Similar to most k-means-type algorithms, the minimization problem of (4) can be solved by iteratively: 1) computing the minimizer $\mathbf{Q}$ with fixing $\mathbf{B} = \hat{\mathbf{B}}$ and 2) computing the minimizer $\mathbf{B}$ with fixing $\mathbf{Q} = \hat{\mathbf{Q}}$, until convergence or a certain stopping criterion is satisfied. For such optimization processes, correctly defined dissimilarities are the vital prerequisite for computing the minimizers $\mathbf{Q}$ and $\mathbf{B}$. Therefore, how to appropriately define the dissimilarities $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$, the contribution weights $\mathbf{W}$ and how to compute the two minimizers $\mathbf{B}$ and $\mathbf{Q}$ are two focuses of this article. In the following, exact definition of $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$ is given in Section III-B, and in Section III-C, details about how to compute the two minimizers are rigorously provided.

B. Dissimilarity Metric

Since both the values of x_i^r and b_l^r in $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$ are from O^r, we should first understand the space structures corresponding to different O^rs before defining $\mathrm{dissim}^{\mathrm{rs}}(x_i^r, b_l^r)$. As discussed in Section III-A, if A^r is a nominal attribute, all its n values are taken from a finite number of possible values, which can be viewed as a discrete space $\mathcal{C}^{r\langle n\rangle}$. If A^r is a numerical attribute, the corresponding value space is surely a 1-D real space $\mathcal{R}^{r\langle u\rangle}$. However, an ordinal attribute has the characteristics of both nominal and numerical attributes, i.e., ordinal attribute values fall in a limited number of possible values as a nominal attribute, but these values are comparable in the direction of their order, which is similar to the values of a numerical attribute. Thus, the possible values of an ordinal attribute A^r can be viewed as a 1-D discrete space, which is denoted as $\mathcal{C}^{r\langle o\rangle}$. We first discuss the connections of these three types of attributes in the following and then present detailed dissimilarity definitions accordingly.

Remark 1: Connection of ordinal and numerical attributes: for an ordinal attribute, if we increase its number of possible values to approach infinity, then this ordinal attribute approximates a numerical attribute. From such a perspective, the possible values of an ordinal attribute can be viewed as a certain number of numerical values with unknown dissimilarities, and thus, these dissimilarities can be reflected by the numerical attributes correlated with the ordinal one.

Remark 2: Connection of nominal and ordinal attributes: possible values of an ordinal attribute are the grades between two contradicted concepts, e.g., the four grades {none, to_pain, to_speech, spontaneous} between the most severe "none" and the most normal "spontaneous" of "GCS-E" attribute in Table I. When considering any pair of intraattribute nominal possible values, they can be viewed as an ordinal attribute A^r with only two grades describing the two contradicted concepts. Therefore, each pair of nominal possible values can be viewed as a pair of concept-contradicted values (CCVs) during dissimilarity measurement.

Definition 1: Any pair of possible values in O^r when $r < d^{\langle n\rangle}$ (i.e., A^r is a nominal attribute), and the two values on the two ends of O^r when $d^{\langle n\rangle} + 1 < r < d^{\langle n\rangle} + d^{\langle o\rangle}$ (i.e., A^r is an ordinal attribute), are defined as a pair of **CCVs**. All the CCVs from O^r are denoted as $O^{r\langle *\rangle}$. Accordingly, there are $v^r(v^r - 1)/2$ pairs of CCVs for a nominal attribute, and only one pair of CCVs for an ordinal attribute when $v^r > 1$.

Remark 3: Connection of nominal and numerical attributes: according to Remark 2, the dissimilarity between any pair of nominal possible values can be viewed as the dissimilarity between two possible values of an ordinal attribute with $v^r = 2$, and thus, the dissimilarity can be indicated by numerical attributes according to Remark 1.

To appropriately represent the dissimilarity spaces of heterogeneous attributes, we construct graphs for the attributes according to the above-mentioned connections and the intrinsic characteristics of heterogeneous attributes. Then, the spaces of the heterogeneous attributes in A are mapped into homogeneous spaces $G = \{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^d\}$ described by nodes and edges.

Remark 4: Graph construction details: the graph $\mathcal{G}^r = \{O^r, E^r\}$ corresponding to A^r is constructed by treating values in O^r as the nodes and then linking them using edges E^r. As shown in Fig. 1, different diameters of nodes indicate the occurrence frequency of the values. For the nominal case (i.e., $r \le d^{\langle n\rangle}$), $v^r(v^r-1)/2$ edges fully connect the possible values. For the ordinal (or numerical) case (i.e., $r > d^{\langle n\rangle}$), all the v^r values are linearly arranged and are successively connected by $v^r - 1$ edges.

A universal definition of dissimilarity is that the higher the cost of transforming two subjects into each other, the more dissimilar the two subjects are. From the perspective of the constructed graph structure, the transformation can be viewed as transporting values falling from a possible value to another possible value along the edge between them. Consequently, our goal of dissimilarity measurement is converted to estimating the transformation cost between two nodes o_m^r and o_h^r according to the graph structures, which can be written as

$$\Psi^r(o_m^r, o_h^r) = \sum_{s=1}^{d} \psi^{\mathrm{rs}}\left(o_m^r, o_h^r\right) \cdot w^{\mathrm{rs}} \tag{5}$$

where $\psi^{\mathrm{rs}}(\cdot, \cdot)$ is a function that quantifies the transformation cost between two nodes o_m^r and o_h^r of $\mathcal{G}^r$ reflected by $\mathcal{G}^s$. The weight w^{rs} controls the contribution of $\mathcal{G}^s$ in forming $\Psi^r(o_m^r, o_h^r)$, which can be quantified as the interdependence of A^r and A^s as discussed in Section III-A. Moreover, r and s should satisfy $r \in \{1, 2, \ldots, d^{\langle n\rangle} + d^{\langle o\rangle}\}$ and $s \in \{1, 2, \ldots, d\}$, respectively, which is discussed in Remark 5.

Remark 5: The role of numerical attributes: as numerical attributes are with well-defined Euclidean dissimilarity space, they are only represented by the graphs to reflect dissimilarities of nominal and ordinal attributes. Accordingly, we have $r \in \{1, 2, \ldots, d^{\langle n\rangle} + d^{\langle o\rangle}\}$ and $s \in \{1, 2, \ldots, d\}$ in (5).

Definitions of $\psi^{rs}(\cdot,\cdot)$ and w^{rs} in (5) are as follows. We first define

$$\mathbf{u}^{rs}_m = \left[p\left(o^s_1\middle|o^r_m\right), p\left(o^s_2\middle|o^r_m\right), \ldots, p\left(o^s_{v^s}\middle|o^r_m\right)\right]^\top$$

which is the CPD of the values in O^s as given o^r_m. Note that

$$p\left(o^s_g\middle|o^r_m\right) = \frac{\sigma\left(X^s_g \cap X^r_m\right)}{\sigma\left(X^r_m\right)}$$

where $X^r_m = \{\mathbf{x}_i | x^r_i = o^r_m, i = 1, 2, \ldots, n\}$ is a subset of X with the rth values of all its objects equal to o^r_m, and the function $\sigma(\cdot)$ counts the cardinality of a set. By further considering the case $r = s$, $\mathbf{u}^{rs}_m$ can be rewritten as

$$\mathbf{u}^{rs}_m = \begin{cases} \Bigg[\underbrace{0, 0, \ldots,}_{m-1} 1, \underbrace{\ldots, 0, 0}_{v^r-m}\Bigg]^\top, & \text{if } r = s \\ \left[p\left(o^s_1\middle|o^r_m\right), p\left(o^s_2\middle|o^r_m\right), \ldots, p\left(o^s_{v^s}\middle|o^r_m\right)\right]^\top. & \text{if } r \neq s. \end{cases}$$

Actually, $\mathbf{u}^{rs}_m$ describes the situation of $\mathcal{G}^s$ as given o^r_m, and we denote such situation as $\mathcal{G}^{rs}_m$. Then, we will discuss how to define the cost $\psi^{rs}(o^r_m, o^r_h)$ based on $\mathbf{u}^{rs}_m$.

The cost quantification function $\psi^{rs}(\cdot,\cdot)$ should sufficiently exploit the homogeneity brought by the graph space structures. Inspired by the earth mover's distance (EMD) [35], [36], which is originally proposed for computing the cost of transforming one signature into another, we compute the cost of transforming $\mathcal{G}^{rs}_m$ and $\mathcal{G}^{rs}_h$ into each other by

$$\psi^{rs}(o^r_m, o^r_h) = \sum_{\mathfrak{g}=1}^{v^r_{mh}-1} \left|\mathbf{u}^{rs}_{\mathfrak{g}} - \mathbf{u}^{rs}_{\mathfrak{t}}\right| \cdot \mathbf{t}^{rs}_{\mathfrak{gt}} \tag{6}$$

where $\mathfrak{t} = \mathfrak{g} + 1$. $\mathbf{u}^{rs}_{\mathfrak{g}}$ and $\mathbf{u}^{rs}_{\mathfrak{t}}$ are the CPDs of the values in O^s as given the $\mathfrak{g}$th and $\mathfrak{t}$th values in O^r_{mh}, which is a set containing all the v^r_{mh} intermediate values (including o^r_m and o^r_h themselves) on the shortest path between o^r_m and o^r_h. More specifically, if A^r is a nominal attribute, every two nodes in the corresponding graph $\mathcal{G}^r$ are directly linked, and thus, we have $O^r_{mh} \equiv \{o^r_m, o^r_h\}$ and $v^r_{mh} \equiv 2$. If A^r is an ordinal or numerical attribute, all the values in O^r ordered between o^r_m to o^r_h (including themselves) are the nodes on the shortest path from o^r_m to o^r_h in the corresponding graph $\mathcal{G}^r$, and thus, we have $v^r_{mh} = |m - h| + 1$, and $O^r_{mh} = \{o^r_m, o^r_{m+1}, \ldots, o^r_h\}$ if $m < h$, or $O^r_{mh} = \{o^r_h, o^r_{h+1}, \ldots, o^r_m\}$ if $m > h$. In (6), the $v^r_{mh} - 1$ subtransformation costs on the path through the v^r_{mh} nodes are successively accumulated. The CPD difference $\mathbf{u}^{rs}_{\mathfrak{g}} - \mathbf{u}^{rs}_{\mathfrak{t}}$ describes the differences between the corresponding nodes of $\mathcal{G}^{rs}_{\mathfrak{g}}$ and $\mathcal{G}^{rs}_{\mathfrak{t}}$ that should be transported for offsetting in the graph transformation. Vector $\mathbf{t}^{rs}_{\mathfrak{gt}} = [t^{rs}_{gt1}, t^{rs}_{gt2}, \ldots, t^{rs}_{gtv^s}]^\top$ stores the minimum total edge lengths that should be taken to transport each of the differences. Given graph structures with edge lengths, obtaining $\mathbf{t}^{rs}_{\mathfrak{gt}}$ will become a very simple optimization problem in operations research. The determination of the edge lengths is discussed in the following.

Remark 6: Edge length: a prior knowledge we have is that each pair of CCVs represent two different concepts. Similar to the one-hot encoding, we set the total length of the edges on the shortest path between each pair of CCVs to an identical value θ. Another prior knowledge is that ordinal attribute values have relative order, and thus, θ for an ordinal attribute with only one pair of CCVs is equally allocated to the edges.

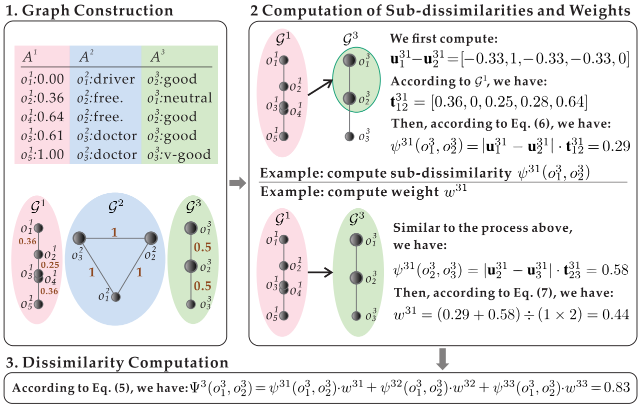
1. Graph Construction
A^1
A^2
A^3
o^1_1:0.00
o^1_2:0.36
o^1_4:0.64
o^1_3:0.61
o^1_5:1.00
o^2_1:driver
o^2_2:free.
o^2_2:free.
o^2_3:doctor
o^2_3:doctor
o^3_2:good
o^3_1:neutral
o^3_2:good
o^3_2:good
o^3_3:v-good
$\mathcal{G}^1$
o^1_1
0.36
o^1_2
0.25
o^1_3
o^1_4
0.36
o^1_5
$\mathcal{G}^2$
1
o^2_3
o^2_2
1
1
o^2_1
$\mathcal{G}^3$
o^3_1
0.5
o^3_2
0.5
o^3_3
2 Computation of Sub-dissimilarities and Weights
$\mathcal{G}^1$
$\mathcal{G}^3$
We first compute:
$\mathbf{u}^{31}_1 - \mathbf{u}^{31}_2 = [-0.33, 1, -0.33, -0.33, 0]$
According to $\mathcal{G}^1$, we have:
$\mathbf{t}^{31}_{12} = [0.36, 0, 0.25, 0.28, 0.64]$
Then, according to Eq. (6), we have:
$\psi^{31}(o^3_1, o^3_2) = |\mathbf{u}^{31}_1 - \mathbf{u}^{31}_2| \cdot \mathbf{t}^{31}_{12} = 0.29$
Example: compute sub-dissimilarity $\psi^{31}(o^3_1, o^3_2)$
Example: compute weight w^{31}
Similar to the process above, we have:
$\psi^{31}(o^3_2, o^3_3) = |\mathbf{u}^{31}_2 - \mathbf{u}^{31}_3| \cdot \mathbf{t}^{31}_{23} = 0.58$
Then, according to Eq. (7), we have:
$w^{31} = (0.29 + 0.58) \div (1 \times 2) = 0.44$
3. Dissimilarity Computation
According to Eq. (5), we have: $\Psi^3(o^3_1, o^3_2) = \psi^{31}(o^3_1, o^3_2) \cdot w^{31} + \psi^{32}(o^3_1, o^3_2) \cdot w^{32} + \psi^{33}(o^3_1, o^3_2) \cdot w^{33} = 0.83$

Fig. 2. Pipeline of dissimilarity calculation (taking the data fragment in the lower part of Table I as an example). In Step 1, we first normalize the numerical attribute A^1 and then construct graphs for the three attributes according to Remark 4. Since values of A^1 are normalized into [0, 1], which yields $\theta = 1$, we set $\theta = 1$ for all the graphs. Steps 2 and 3 are demonstrated by taking the computation of $\Psi^3(o^3_1, o^3_2)$ as an example. In Step 2, we first prepare $\mathbf{u}^{31}_1 = [0, 1, 0, 0, 0]$, $\mathbf{u}^{31}_2 = [0.33, 0, 0.33, 0.33, 0]$, and $\mathbf{u}^{31}_3 = [0, 0, 0, 0, 1]$, then we successively compute $\psi^{31}(o^3_1, o^3_2)$ and w^{31} according to (6) and (7). In the same way, we obtain $\psi^{32}(o^3_1, o^3_2)$, w^{32}, $\psi^{33}(o^3_1, o^3_2)$, and w^{33}. Finally, in Step 3, the subdissimilarities and the corresponding weights are combined according to (5) to form $\Psi^3(o^3_1, o^3_2)$.

Then, we discuss how to quantify the weight w^{rs} in (5) based on the interdependence between A^r and A^s. Note that the transformation cost $\psi^{rs}(o^r_m, o^r_h)$ is the degree of dissimilarity between the two different values o^r_m and o^r_h reflected by A^s, which partially reflects the dependence of A^s on A^r. More intuitively, if the dissimilarities between different values of A^r reflected by A^s are always higher than the dissimilarities reflected by the other attributes, A^r and A^s are considered to have a stronger interdependence. Accordingly, we derived the definition of w^{rs} as

$$w^{rs} = \frac{\sum_{\mathfrak{q}=1}^{v^{r\langle*\rangle}-1} \sum_{\mathfrak{c}=\mathfrak{q}+1}^{v^{r\langle*\rangle}} \psi^{rs}(o^r_{\mathfrak{q}}, o^r_{\mathfrak{c}})}{\dfrac{v^{r\langle*\rangle}\left(v^{r\langle*\rangle}-1\right)}{2} \cdot \left(v^r_{\mathfrak{qc}} - 1\right)} \tag{7}$$

where $o^r_{\mathfrak{q}}$ and $o^r_{\mathfrak{c}}$ are the $\mathfrak{q}$th and $\mathfrak{c}$th nodes in the set $O^{r\langle*\rangle}$, respectively, and $O^{r\langle*\rangle}$ is the set contains $v^{r\langle*\rangle}$ concept-contradicted nodes of $\mathcal{G}^r$. $v^r_{\mathfrak{qc}}$ is the number of intermediate nodes (including $o^r_{\mathfrak{q}}$ and $o^r_{\mathfrak{c}}$ themselves) on the shortest path between $o^r_{\mathfrak{q}}$ and $o^r_{\mathfrak{c}}$. Equation (7) actually quantifies the averaged transformation cost between every pair of the concept-contradicted nodes in $O^{r\langle*\rangle}$ reflected by $\mathcal{G}^s$. Dissimilarity of heterogeneous attributes are uniformly quantified as transformation cost by (5)–(7). A pipeline and a toy example are provided in Fig. 2 to further demonstrate the calculation process.

Based on the dissimilarity between two attribute values defined by (5), the overall dissimilarity between two data objects $\mathbf{x}_i$ and $\mathbf{x}_j$ can be computed by

$$\Psi(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{d} \Psi^r\left(x^r_i, x^r_j\right)^2}. \tag{8}$$

Based on (6)–(8), graph-based unified dissimilarity (GUD) suitable for ADC is, thus, formed. We prove that GUD is a metric in the following.

Lemma 1: Transformation cost defined in (6) satisfies $\psi^{rs}(o^r_m, o^r_h) \leq \psi^{rs}(o^r_m, o^r_t) + \psi^{rs}(o^r_t, o^r_h)$ for any $m, h, t \in \{1, 2, \ldots, v^r\}$, $r \in \{1, 2, \ldots, d^{\langle n\rangle} + d^{\langle o\rangle}\}$, and $s \in \{1, 2, \ldots, d\}$.

Proof: We prove Lemma 1 in the following two cases.

Case 1: A^r is a nominal attribute (i.e., $r \in \{1, 2, \ldots, d^{\langle n \rangle}\}$). In this case, O^r_{mh} only contains o^r_m and o^r_h themselves according to the graph structure because there is no intermediate nodes between o^r_m and o^r_h and we have $\upsilon^r_{mh} \equiv 2$ in (6). Since the lengths of all the edges in $\mathcal{G}^s$ are identical, we set them to 1 without affecting the analysis, and we have

$$\psi^{\mathrm{rs}}\left(o^r_m, o^r_h\right) = \frac{\left\|\left|\mathbf{u}^{\mathrm{rs}}_m - \mathbf{u}^{\mathrm{rs}}_h\right|\right\|_1}{2}. \tag{9}$$

Since $\mathbf{u}^{\mathrm{rs}}_m$ and $\mathbf{u}^{\mathrm{rs}}_h$ satisfy $|\mathbf{u}^{\mathrm{rs}}_m(f) - \mathbf{u}^{\mathrm{rs}}_h(f)| \leq |\mathbf{u}^{\mathrm{rs}}_m(f) - \mathbf{u}^{\mathrm{rs}}_t(f)| + |\mathbf{u}^{\mathrm{rs}}_t(f) - \mathbf{u}^{\mathrm{rs}}_h(f)|$ for any $f \in \{1, 2, \ldots, \upsilon^s\}$, where the operation $\mathbf{u}^{\mathrm{rs}}_m(f)$ takes the value of the fth digit of $\mathbf{u}^{\mathrm{rs}}_m$. Then, we have

$$\frac{\left\|\left|\mathbf{u}^{\mathrm{rs}}_m - \mathbf{u}^{\mathrm{rs}}_h\right|\right\|_1}{2} \leq \frac{\left\|\left|\mathbf{u}^{\mathrm{rs}}_m - \mathbf{u}^{\mathrm{rs}}_t\right|\right\|_1}{2} + \frac{\left\|\left|\mathbf{u}^{\mathrm{rs}}_t - \mathbf{u}^{\mathrm{rs}}_h\right|\right\|_1}{2}$$
$$\Rightarrow \psi^{\mathrm{rs}}\left(o^r_m, o^r_h\right) \leq \psi^{\mathrm{rs}}\left(o^r_m, o^r_t\right) + \psi^{\mathrm{rs}}\left(o^r_t, o^r_h\right).$$

Case 2: A^r is an ordinal attribute (i.e., $r \in \{d^{\langle n \rangle} + 1, d^{\langle n \rangle} + 2, \ldots, d^{\langle n \rangle} + d^{\langle o \rangle}, \}$). In this case, when g and t satisfy $|g - t| = 1$, there is no intermediate nodes between o^r_g and o^r_t, and thus, the transformation cost between two nodes of an ordinal attribute's graph structure degenerates to (9). Accordingly, (6) can be transformed into

$$\psi^{\mathrm{rs}}\left(o^r_m, o^r_h\right) = \sum_{g=\min(m,h), t=g+1}^{\max(m,h)-1} \frac{\left\|\left|\mathbf{u}^{\mathrm{rs}}_g - \mathbf{u}^{\mathrm{rs}}_t\right|\right\|_1}{2}.$$

Then, we have

$$\begin{aligned}
&\psi^{\mathrm{rs}}\left(o^r_m, o^r_h\right) = \psi^{\mathrm{rs}}\left(o^r_m, o^r_t\right) + \psi^{\mathrm{rs}}\left(o^r_t, o^r_h\right)\\
&\text{if } m \leq t \leq h \text{ or } m \geq t \geq h \text{ , and}\\
&\psi^{\mathrm{rs}}\left(o^r_m, o^r_h\right) < \psi^{\mathrm{rs}}\left(o^r_m, o^r_t\right) + \psi^{\mathrm{rs}}\left(o^r_t, o^r_h\right)\\
&\text{if } t < \min\left(m, h\right) \text{ or } t > \max\left(m, h\right).
\end{aligned}$$

□

Theorem 1: Dissimilarity measure defined in (5)–(8) is a metric.

Proof: According to (5)–(7) and Lemma 1, it is clear that the defined intraattribute dissimilarity $\Psi^r(o^r_m, o^r_h) = \sum_{s=1}^{d} \psi^{\mathrm{rs}}(o^r_m, o^r_h) \cdot w^{\mathrm{rs}}$ satisfies the following properties for any $m, h, t \in \{1, 2, \ldots, \upsilon^r\}$ and $r \in \{1, 2, \ldots, d^{\langle n \rangle} + d^{\langle o \rangle}\}$.

1) $\Psi^r(o^r_m, o^r_h) \geq 0$.
2) $o^r_m = o^r_h \Leftrightarrow \Psi^r(o^r_m, o^r_h) = 0$.
3) $\Psi^r(o^r_m, o^r_h) = \Psi^r(o^r_h, o^r_m)$.
4) $\Psi^r(o^r_m, o^r_h) \leq \Psi^r(o^r_m, o^r_t) + \Psi^r(o^r_t, o^r_h)$.

According to (8), it is clear that the following properties hold for any $i, j, l \in \{1, 2, \ldots, n\}$.

1) $\Psi(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
2) $\mathbf{x}_i = \mathbf{x}_j \Leftrightarrow \Psi(\mathbf{x}_i, \mathbf{x}_j) = 0$.
3) $\Psi(\mathbf{x}_i, \mathbf{x}_j) = \Psi(\mathbf{x}_j, \mathbf{x}_i)$.
4) $\Psi(\mathbf{x}_i, \mathbf{x}_j) \leq \Psi(\mathbf{x}_i, \mathbf{x}_l) + \Psi(\mathbf{x}_l, \mathbf{x}_j)$.

The defined dissimilarity measure satisfies all the metric properties. □

In practice, a set of dissimilarity matrices $D = \{\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^{d^{\langle n \rangle} + d^{\langle o \rangle}}\}$, where $\mathbf{D}^r$ is a $\upsilon^r \times \upsilon^r$ symmetric matrix storing intraattribute dissimilarities of A^r, can be computed before clustering. Value of the (m, h)th entry of $\mathbf{D}^r$ is $\Psi^r(o^r_m, o^r_h)$ computed by (5). With D, dissimilarities can be directly read off during clustering. However, for the case that A^s is a numerical attribute (i.e., $s \in \{d^{\langle n \rangle} + d^{\langle o \rangle} + 1, d^{\langle n \rangle} + d^{\langle o \rangle} + 2, \ldots, d\}$), a large υ^s may make the computation laborious. A fast approximation way for the computation is to discretize the value range of A^s into a small number of intervals, and then A^s can be treated in the same way as ordinal attributes for indicating the dissimilarities of the other attributes. Note that such a discretization process is optional, and exact Euclidean distances between the values of numerical attributes are still utilized for the object level dissimilarity computation in (8). Then, we analyze the time and space complexity of GUD as follows.

Theorem 2: The time complexity for computing the dissimilarity matrices in D using GUD is $O(nd^{\langle n \rangle}d + nd^{\langle o \rangle}d + \vartheta^3 d^{\langle n \rangle}d + \vartheta^3 d^{\langle o \rangle}d)$.

Proof: For convenience, we adopt $\vartheta = \max(\upsilon^1, \upsilon^2, \ldots, \upsilon^{d^{\langle n \rangle} + d^{\langle o \rangle}})$ in all subsequent time and space complexity analyses. Since the number of intraattribute dissimilarities that should be computed for a nominal attribute is $\vartheta(\vartheta - 1)/2$, which is larger than $\vartheta - 1$ of an ordinal attribute, we treat A^r as a nominal attribute in the following analysis.

CPDs should be first computed for the $(d^{\langle n \rangle} + d^{\langle o \rangle}) \times d$ pairs of attributes. Since for each pair of attributes, the n values of A^s should be scanned once, time complexity for preparing all the CPDs is $O(n(d^{\langle n \rangle} + d^{\langle o \rangle})d)$.

Then, transformation costs $\psi^{\mathrm{rs}}(o^r_m, o^r_h)$ in (6) should be computed for the $\vartheta(\vartheta - 1)/2$ pairs of possible values of $d^{\langle n \rangle} + d^{\langle o \rangle}$ attributes indicated by d attributes. If A^s is a nominal attribute, subtraction and sum operations should be performed to the ϑ digits of $\mathbf{u}^{\mathrm{rs}}_m$ and $\mathbf{u}^{\mathrm{rs}}_h$, in turn, in (6). If A^s is an ordinal attribute, subtraction and digit-by-digit sum operations should be performed to the ϑ digits of $\mathbf{u}^{\mathrm{rs}}_m$ and $\mathbf{u}^{\mathrm{rs}}_h$, in turn, in (6). That is, the nominal and ordinal cases of A^s have the same complexity. Therefore, time complexity for computing all the transformation costs is $O(\vartheta^3(d^{\langle n \rangle} + d^{\langle o \rangle})d)$.

After that, $(d^{\langle n \rangle} + d^{\langle o \rangle}) \times d$ weights w^{rs} in $\mathbf{W}$ should be computed according to (7). Each component (i.e., transformation cost) involved in the sum operation in the numerator has already been computed, and only the sum operation is needed. There are $\vartheta(\vartheta - 1)/2$ and 1 sum operations for the case that A^r is a nominal attribute and A^r is an ordinal attribute, respectively. The value of the denominator is a constant given A^r. Therefore, time complexity for computing the weights is $O(d^{\langle n \rangle}\vartheta^2 d + d^{\langle o \rangle}d)$.

Finally, transformation costs computed by (6) and weights computed by (7) are concatenated by (5) to form $\vartheta(\vartheta - 1)/2$ intraattribute dissimilarities of the $d^{\langle n \rangle} + d^{\langle o \rangle}$ nominal and ordinal attributes. Since each dissimilarity is indicated by $d - 1$ attributes, time complexity for computing the intraattribute dissimilarities is $O(\vartheta^2(d^{\langle n \rangle} + d^{\langle o \rangle})d)$.

The overall time complexity for computing D is the summation of the above-analyzed time complexities, which is $O(nd^{\langle n \rangle}d + nd^{\langle o \rangle}d + \vartheta^3 d^{\langle n \rangle}d + \vartheta^3 d^{\langle o \rangle}d)$. □

Theorem 3: The space complexity for computing the dissimilarity matrices in D using GUD is $O(nd + \vartheta^2 d^{\langle n \rangle} + \vartheta^2 d^{\langle o \rangle})$.

Proof: The space for storing the dataset X and the dissimilarity matrices D is $n \times d$ and $\vartheta^2 \times (d^{\langle n \rangle} + d^{\langle o \rangle})$, respectively. All the values in each $\vartheta \times \vartheta$ space for storing each $\mathbf{D}^r$ in D are initialized to 0. Then, we analyze the space complexity for computing D in the following.

CPDs of A^s's values given each of the ϑ values of A^r should be stored for the computation of cost $\psi^{\mathrm{rs}}(o^r_m, o^r_h)$. Since each CPD is a ϑ-bit vector, the space taken for storing

CPDs is $\vartheta \times \vartheta$, and this space is used by different attribute pairs.

There are a total of $\vartheta \times (\vartheta - 1)/2$ costs for different A^r and A^s, and the required space is $\vartheta \times \vartheta$, which can be denoted as a matrix $\mathbf{D}^{\mathrm{rs}}$. Accordingly, the weight w^{rs} can be obtained through $\mathbf{D}^{\mathrm{rs}}$. Then, we multiply all the costs in $\mathbf{D}^{\mathrm{rs}}$ with w^{rs} to form the weighted costs $\psi^{\mathrm{rs}}(o_m^r, o_h^r) \cdot w^{\mathrm{rs}}$ to replace the original costs $\psi^{\mathrm{rs}}(o_m^r, o_h^r)$ in $\mathbf{D}^{\mathrm{rs}}$. When considering the next A^s for a given A^r, the current weighted costs in $\mathbf{D}^{\mathrm{rs}}$ are added to the space for storing $\mathbf{D}^r$, and then the new weighted costs are computed and stored in $\mathbf{D}^{\mathrm{rs}}$. The above-mentioned processes are repeated until the weighted costs contributed by each A^s ($s = \{1, 2, \ldots, d\}$) have been added to the space for $\mathbf{D}^r$. At this time, we obtain the final $\mathbf{D}^r$. Since the $\vartheta \times \vartheta$ space $\mathbf{D}^{\mathrm{rs}}$ is utilized for storing the costs and weighted costs for different attribute pairs, space taken for the computation of D is $\vartheta \times \vartheta$.

The overall space complexity for computing D is $O(nd + \vartheta^2(d^{\langle n \rangle} + d^{\langle o \rangle}) + 2\vartheta^2)$, which can be simplified to $O(nd + \vartheta^2 d^{\langle n \rangle} + \vartheta^2 d^{\langle o \rangle})$. □

Since the existing distance metrics are either for categorical data or numerical data, we compare their complexity of them and the proposed GUD in the following two cases: 1) given a numerical dataset, GUD has the same time and space complexity as the commonly used Euclidean distance because GUD directly adopts Euclidean distances for numerical attributes. 2) Given a categorical dataset, we have $d^{\langle n \rangle} + d^{\langle o \rangle} = d$, and thus, the time and space complexity of GUD becomes $O(nd^2 + \vartheta^3 d^2)$ and $O(nd + \vartheta^2 d)$, respectively, which is the same as that of mainstream [17], [18], [26], [27] and the state-of-the-art [13], [15], [28] distance metrics.

C. Clustering Algorithm

In this section, a new clustering algorithm is proposed to unify the processing of heterogeneous attributes in clustering, and thus, provide an elegant solution for ADC. Furthermore, the derivation of the algorithm also ensures the consistency of the optimization process and the objective under the premise of using the newly defined GUD dissimilarity metric. We reform the objective function defined in (4) based on GUD as

$$z = \sum_{i=1}^{n} \sum_{l=1}^{k} q_{il} \Psi(\mathbf{x}_i, \mathbf{b}_l) \tag{10}$$

and the updating strategies of $\mathbf{Q}$ and $\mathbf{B}$ are given in the following.

Theorem 4: Let $\mathbf{B}$ be fixed, z is minimized iff $\mathbf{Q}$ is computed by

$$q_{il} = \begin{cases} 1, & \text{if } l = \arg\min_{y} \Psi(\mathbf{x}_i, \mathbf{b}_y) \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Proof: Since all the inner sums of (10) are nonnegative and independent, the inner sum contributed by $\mathbf{x}_i$ can be written as

$$z^{\langle \mathbf{x}_i \rangle} = \sum_{l=1}^{k} q_{il} z^{\langle \mathbf{x}_i, C_l \rangle}, \quad z^{\langle \mathbf{x}_i, C_l \rangle} = \Psi(\mathbf{x}_i, \mathbf{b}_l)$$

where $z^{\langle \mathbf{x}_i, C_l \rangle}$ is the inner sum contributed by $\mathbf{x}_i$ in the lth cluster C_l. Since $\sum_{l=1}^{k} q_{il} = 1$ and $q_{il} \in \{0, 1\}$, it is clear that $z^{\langle \mathbf{x}_i \rangle}$ is minimized if the minimum $z^{\langle \mathbf{x}_i, C_l \rangle}$ is assigned with $q_{il} = 1$ and the other $z^{\langle \mathbf{x}_i, C_l \rangle}$s are assigned with $q_{il} = 0$. The result follows (11). □

Theorem 5: Let $\mathbf{Q}$ be fixed, z is minimized iff $\mathbf{B}$ is computed by

$$b_l^r = \arg\min_{b_t^r} \sqrt{\sum_{g=1}^{\sigma(C_l)} \Psi^r\left(c_{lg}^r, b_t^r\right)^2} \tag{12}$$

where b_t^r is a possible value of A^r, c_{lg}^r is the gth value of C_l^r, and C_l^r is the set of rth values of the objects in cluster C_l.

Proof: Since all the inner sums contributed by different clusters in $C = \{C_1, C_2, \ldots, C_k\}$ are nonnegative and independent of (10), the inner sum contributed by cluster C_l can be written as

$$z^{\langle C_l \rangle} = \sum_{i}^{n} q_{il} \sqrt{\sum_{r=1}^{d} \Psi^r\left(x_i^r, b_l^r\right)^2} = \sum_{g=1}^{\sigma(C_l)} \sqrt{\sum_{r=1}^{d} \Psi^r\left(c_{lg}^r, b_l^r\right)^2}.$$

Since all the inner sums of $z^{\langle C_l \rangle}$ are nonnegative and independent, the inner sum contributed by C_l^r can be written as

$$z^{\langle C_l^r \rangle} = \sum_{g=1}^{\sigma(C_l)} \Psi^r\left(c_{lg}^r, b_l^r\right)^2.$$

Then, we prove Theorem 5 in the following two cases.

Case 1: $r \in \{1, 2, \ldots, d^{\langle n \rangle} + d^{\langle o \rangle}\}$, since the values c_{lg}^r and b_l^r are categorical values, which cannot participate in mathematical operations themselves, the optimal b_l^r can only be obtained by comparing the values of $z^{\langle C_l^r \rangle}$ with different b_l^rs among all the possible values of A^r. Then, it is clear that the optimal b_l^r is the one producing the minimum $z^{\langle C_l^r \rangle}$. The result follows (12).

Case 2: $r \in \{d^{\langle n \rangle} + d^{\langle o \rangle} + 1, d^{\langle n \rangle} + d^{\langle o \rangle} + 2, \ldots, d\}$, since the values c_{lg}^r and b_l^r can participate in mathematical operations and b_l^r can be any possible values in the value range of A^r, we have

$$z^{\langle C_l^r \rangle} = \sum_{g=1}^{\sigma(C_l)} \Psi^r\left(c_{lg}^r, b_l^r\right)^2 = \sum_{g=1}^{\sigma(C_l)} \left(c_{lg}^r - b_l^r\right)^2$$

$$\frac{\partial z^{\langle C_l^r \rangle}}{\partial b_l^r} = 2\left(\sum_{g=1}^{\sigma(C_l)} b_l^r - \sum_{g=1}^{\sigma(C_l)} c_{lg}^r\right).$$

Let $((\partial z^{\langle C_l^r \rangle})/(\partial b_l^r)) = 0$, we then have

$$\sum_{g=1}^{\sigma(C_l)} b_l^r = \sum_{g=1}^{\sigma(C_l)} c_{lg}^r \Rightarrow \sigma(C_l) b_l^r = \sum_{g=1}^{\sigma(C_l)} c_{lg}^r$$

$$\Rightarrow b_l^r = \left(\sum_{g=1}^{\sigma(C_l)} c_{lg}^r\right) \Big/ \sigma(C_l). \tag{13}$$

Equation (13) actually computes the mean of the values in C_l^r. Then, it is clear that $z^{\langle C_l^r \rangle}$ is minimized iff the mean of the values in C_l^r is assigned to b_l^r, and (13) is equivalent to (12) for the case $r \in \{d^{\langle n \rangle} + d^{\langle o \rangle} + 1, d^{\langle n \rangle} + d^{\langle o \rangle} + 2, \ldots, d\}$. The result follows (12). □

Based on the above-mentioned analysis, we iteratively update $\mathbf{Q}$ and $\mathbf{B}$ in two steps as follows: 1) fix $\mathbf{Q}$, update the former $d^{\langle n \rangle} + d^{\langle o \rangle}$ columns of $\mathbf{B}$ according to (12) and 2) fix $\mathbf{B}$, update $\mathbf{Q}$ according to (11). Since infinite number of possible values of numerical attributes make the value of b_l^r unobtainable according to (12), the latter $d^{\langle u \rangle}$ columns of $\mathbf{B}$ are obtained according to (13). These two steps repeat until

Algorithm 1 ADC Clustering Algorithm

```
Input: Data set $S$, number of clusters $k$.
Output: Partition $\mathbf{Q}$.
Step 0: Compute dissimilarity matrices $D$ by (5);
    Initialize the time-step by $\tau = 0$; Randomly initialize each row of $\mathbf{B}^{(\tau)}$, and compute $\mathbf{Q}^{(\tau)}$ and $z^{\langle\tau\rangle}$ accordingly;
Step 1: Fix $\mathbf{Q} = \mathbf{Q}^{(\tau)}$, compute $\mathbf{B}$ by (12) and (13),
    obtain $\mathbf{B}^{(\tau+1)}$, and compute $z^{\langle\mathbf{B}^{(\tau+1)}\rangle}$ based on $\mathbf{Q}^{(\tau)}$ and $\mathbf{B}^{(\tau+1)}$; If $z^{\langle\mathbf{B}^{(\tau+1)}\rangle} \neq z^{\langle\tau\rangle}$, go to Step 2; Otherwise, stop and Output $\mathbf{Q}^{(\tau)}$;
Step 2: Fix $\mathbf{B}^{(\tau+1)}$, compute $\mathbf{Q}^{(\tau)}$ by (11), obtain
    $\mathbf{Q}^{(\tau+1)}$, and compute $z^{\langle\tau+1\rangle}$ based on $\mathbf{Q}^{\langle\tau+1\rangle}$ and $\mathbf{B}^{\langle\tau+1\rangle}$;
    If $z^{\langle\tau+1\rangle} \neq z^{\langle\mathbf{B}^{(\tau+1)}\rangle}$, update the time-step by $\tau = \tau + 1$, go to Step 1; Otherwise, stop and Output $\mathbf{Q}^{(\tau)}$;
```

the value of z remains unchanged. The corresponding clustering algorithm is summarized in Algorithm 1. Meanwhile, its convergence, time, and space complexity are analyzed as follows.

Theorem 6: ADC algorithm converges to a local minimal solution in a finite number of iterations.

Proof: We show that a possible partition $\mathbf{Q}$ appears at most once by implementing the ADC algorithm before it stops. To prove this, we first assume that this is not true and then provide contradiction proof in the following. Assume $\mathbf{Q}^{(\tau 1)} = \mathbf{Q}^{(\tau 2)}$, where $\tau 1 \neq \tau 2$. Note that given $\mathbf{Q}^{(\tau 1)}$ and $\mathbf{Q}^{(\tau 2)}$, we can compute the minimizers $\mathbf{B}^{(\tau 1+1)}$ and $\mathbf{B}^{(\tau 2+1)}$, respectively, according to Step 1 of Algorithm 1. Since $\mathbf{Q}^{(\tau 1)} = \mathbf{Q}^{(\tau 2)}$, it is clear that $\mathbf{B}^{(\tau 1+1)} = \mathbf{B}^{(\tau 2+1)}$. Therefore, we have

$$z^{\langle\mathbf{B}^{(\tau 1+1)}\rangle} = z^{\langle\mathbf{B}^{(\tau 2+1)}\rangle}. \tag{14}$$

However, the values of z generated by the algorithm are strictly decreasing, which has been proven in Theorems 4 and 5. Hence, (14) is not true. Since there are only a finite number of possible partitions of the dataset S, then the algorithm will reach a local minimal solution in a finite number of iterations. The result follows. □

Theorem 7: Given D, the time complexity of ADC algorithm is $O(\eta(k(d^{\langle n\rangle} + d^{\langle o\rangle})\vartheta n + kd^{\langle u\rangle}n + kdn))$.

Proof: We still adopt $\vartheta = \max(v^1, v^2, \ldots, v^{d^{\langle n\rangle}+d^{\langle o\rangle}})$ for the analysis. In Step 1 of Algorithm 1, there are $k \times d$ values in $\mathbf{B}$ to be computed. For the rth column of $\mathbf{B}$, there are k values to be computed. Then, we analyze the time complexity of Step 1 in the following two cases.

Case 1: $r \in \{1, 2, \ldots, d^{\langle n\rangle} + d^{\langle o\rangle}\}$, b_l^rs in the rth column of $\mathbf{B}$ are computed according to (12). For each b_l^r, there are v^r candidates $O^r = \{o_1^r, o_2^r, \ldots, o_{v^r}^r\}$ to be considered. For each candidate o_m^r, total dissimilarity between o_m^r and C_l^r with $\sigma(C_l)$ values, i.e., $(\sum_{g=1}^{\sigma(C_l)} \Psi^r(c_{lg}^r, o_m^r)^2)^{1/2}$, should be computed. Then, the value in O^r producing the minimum total dissimilarity is assigned to b_l^r. Since we have $\sum_{l=1}^{k} \sigma(C_l) = n$, the time complexity is $O(k(d^{\langle n\rangle} + d^{\langle o\rangle})\vartheta n)$.

Case 2: $r \in \{d^{\langle n\rangle} + d^{\langle o\rangle} + 1, d^{\langle n\rangle} + d^{\langle o\rangle} + 2, \ldots, d\}$, b_l^rs in the rth column of $\mathbf{B}$ are computed according to (13), where $\sigma(C_l)$ values in C_l^r are involved to obtain their mean and is assigned to b_l^r. Hence, the time complexity is $O(kd^{\langle u\rangle}n)$.

In Step 2 of Algorithm 1, for each data object $\mathbf{x}_i$, dissimilarity between it and each row $\mathbf{b}_l$ of $\mathbf{B}$ should be computed. Since dissimilarity matrices of nominal and ordinal attributes (i.e., D) has been computed, overall dissimilarity $\Psi(\mathbf{x}_i, \mathbf{b}_l)$ reflected by the $d^{\langle n\rangle}$ nominal attributes and $d^{\langle o\rangle}$ ordinal attributes can be directly read off, and $\Psi(\mathbf{x}_i, \mathbf{b}_l)$ reflected by the $d^{\langle u\rangle}$ numerical attributes can be obtained directly by subtracting corresponding values in $\mathbf{x}_i$ and $\mathbf{b}_l$. After the above-mentioned computation, $\mathbf{b}_l$ that yields the minimum dissimilarity is selected, then the (i, l)th value (i.e., q_{il}) in $\mathbf{Q}$ is set at 1, and the other values in the ith row of $\mathbf{Q}$ are set at 0. Since there are n objects and k rows in $\mathbf{B}$ in total, the time complexity is $O(kdn)$.

We use η to indicate the total number of iterations required to make Step 1 and Step 2 converge. Then, the time complexity of ADC algorithm is $O(\eta(k(d^{\langle n\rangle} + d^{\langle o\rangle})\vartheta n + kd^{\langle u\rangle}n + kdn))$. □

Theorem 8: Given D, the space complexity of ADC algorithm is $O(nd + nk + kd)$.

Proof: During the computation of ADC, an $n \times d$ space, an $n \times k$ space, and a $k \times d$ space are required to store X, $\mathbf{Q}$, and $\mathbf{B}$, respectively. □

Since ϑ is usually a very small value ($\vartheta \ll n$) for real datasets, the time complexity of ADC can be simplified to $O(\eta kdn)$, which has the same order of magnitude as that of conventional k-means-type clustering algorithms. As for space complexity, ADC does not introduce or generate additional values that need to be stored compared to the conventional k-means-type algorithms. Consequently, they have the same space complexity $O(nd + nk + kd)$.

IV. Experiments

A. Experimental Settings

Five types of experiments have been conducted to comprehensively evaluate the proposed method. Experimental designs are introduced in the following.

- *Comparative Study of the Clustering Performance:* to demonstrate the superiority of the proposed clustering approach, we compare it with the conventional, representative, and state-of-the-art counterparts, respectively, on different types of datasets using different validity indices.
- *Ablation Study of the Proposed Method:* to more specifically show the effectiveness of the proposed ADC clustering approach, we compare different ablated versions of ADC formed corresponding to its components and the ways it handles different types of attributes.
- *Significance study of the comparative results:* To statistically illustrate the superiority of the proposed clustering approach, we implement significance tests to the clustering performance produced in the first experiment "Comparative study of the clustering performance."
- *Intuitive Performance Comparison:* To intuitively show the effectiveness of the proposed GUD metric, we encode the datasets using the GUD measured dissimilarities, conduct dimensionality reduction and visualize the represented data and the measured dissimilarities to provide an impression of GUD.
- *Convergence and Efficiency Evaluation:* Values of the objective function and execution time per iteration are recorded to illustrate the convergence and efficiency of ADC. We also evaluate the efficiency of a large-scale synthetic dataset with the different sampling rates and numbers of possible values. The results of the convergence and efficiency evaluation can be found in the supplementary material.

For all the experiments involving the comparison of clustering performance, each compared approach is executed 50 times and the averaged performance is reported.

In the experiments, the proposed ADC is compared with 15 counterparts, including eight representative approaches in the literature and seven variants of ADC. Conventional approaches, including k-means [37] + NC, k-modes [38] clustering algorithm for pure categorical data, and k-prototypes [8] clustering algorithm for datasets composed of both numerical and categorical attributes are selected as the benchmark counterparts. Representative approaches, including the entropy-based LSM [11] and context-based distance metric (CBDM) [17], which are two representative similarity measures proposed for categorical data, are both selected as counterparts. Both of them are combined with k-modes to form two clustering approaches for comparison. The state-of-the-art clustering approaches, includes iterative clustering learning based on object-cluster similarity metric (OCIL) [10] proposed for dataset composed of numerical and categorical attributes, and unified distance metric (UDM)-based clustering (UDMC) [13] and HD-based clustering (HDC) [15] proposed for dataset composed of nominal and ordinal attributes, have also been chosen as counterparts. Note that we set the parameters (if any) of the above-mentioned counterparts at the values suggested by the corresponding papers. In addition, seven variations of ADC, called ADC$^{\text{I}}$, ADC$^{\text{II}}$, ADC$^{\text{III}}$, ADC$^{\text{IV}}$, O2N, O2U, and ON2U are generated for the ablation studies. Details of the seven ADC variants are provided in Section IV-C.

Three validity indices are chosen for evaluating the performance. CA [39] is a conventional and popular validity index, which computes the matching rate based on the best permutation mapping between the obtained clusters and the true classes. Thus, the value range of CA is [0, 1]. Adjusted rand index (ARI) [40], [41], [42] is a powerful and popular index, which is a random labeling independent version of the RI, where RI quantifies the agreement between the obtained clusters and the true classes as the percentage of object pairs that are assigned in the same or different clusters in the obtained clusters and the clusters obtained according to the true data label. ARI has a value close to 0 for random labeling, and thus, the value range of ARI is $[-1, 1]$. The maximum value of 1 indicates a perfect label matching and vice versa. We also conduct the Bonferroni–Dunn (BD) test [43] to the clustering performance of the compared approaches and compute the critical difference (CD) interval to statistically illustrate the superiority of the proposed approach.

TABLE IV

Statistics of the 30 Datasets. $d^{\langle u\rangle}$, $d^{\langle n\rangle}$, and $d^{\langle o\rangle}$ Indicate the Number of Numerical, Nominal, and Ordinal Attributes, Respectively. n and k^* Indicate the Number of Data Objects and the True Number of Clusters, Respectively

No.	Data Set	Abbrev.	$d^{\langle \mathrm{u}\rangle}$	$d^{\langle \mathrm{n}\rangle}$	$d^{\langle \mathrm{o}\rangle}$	n	k^*
1	Inflammations Diagnosis	DS	1	5	0	120	2
2	Heart Failure	HF	7	5	0	299	2
3	Autism-Adolescent	AA	7	2	0	104	2
4	Dermatology	DT	1	1	32	366	6
5	Australia Credit	AC	6	1	7	690	2
6	Contraceptive Choice	CC	2	1	6	1,473	3
7	Common Toad	CT	2	6	6	189	2
8	Tree Frog	TF	2	6	6	189	2
9	Mammographic	MM	1	2	2	961	2
10	German Credit	GC	7	8	5	1,000	2
11	Adult Income	AI	5	8	1	48,842	2
12	Forest Covertype	FC	10	2	0	581,012	4
13	Census Income	CI	8	33	0	299,258	3
14	KDD Cup	KC	25	3	0	494,021	9
15	MIMIC-III Clinical	MC	6	1	0	58,545	13
16	Breast Cancer	BC	0	5	4	286	2
17	Hayes-Roth	HR	0	2	2	132	3
18	Lenses	LS	0	2	2	24	3
19	Lymphography	LG	0	15	3	148	4
20	Assistant Evaluation	AE	0	2	2	72	3
21	Fruit Evaluation	FT	0	2	3	100	5
22	Soybean	SB	0	35	0	47	4
23	Solar Flare	SF	0	9	0	323	6
24	Congressional Voting	VT	0	16	0	435	2
25	Mushroom	MR	0	21	0	8,124	2
26	Employee Rejection	ER	0	0	4	1,000	9
27	Employee Selection	ES	0	0	4	488	9
28	Lecturer Evaluation	LE	0	0	4	1,000	5
29	Social Workers	SW	0	0	10	1,000	4
30	Internship Questionnaires	IQ	0	0	3	90	2

To ensure a more comprehensive evaluation, various real datasets from different domains (including medicine, finance, biology, sociology, and so on) have been chosen for the experiments. Statistics of the datasets are summarized in Table IV. Datasets 1–14, 16–19, and 22–25 are public datasets collected from the University of California, Irvine (UCI) machine learning repository[3]. Datasets 20, 21, and 30 are collected from Shenzhen University, Shenzhen, China, an advertising company, and the Education University of Hong Kong, Hong Kong, respectively. Dataset 20 has 72 data objects corresponding to the questionnaire answers of 72 students. The four attributes are nominal "gender" and "language," and ordinal "professional" and "helpful." The label describes the course the student is taking. Dataset 21 has 100 data objects corresponding to the business survey records of fruit advertisement (abbreviated as "ad" hereinafter). The five attributes are nominal "fruit type" and "ad type," and ordinal "ad volume," "vendor level," and "fruit price." The label describes the overall evaluation results of the ads. Both datasets 20 and 21 are collected to study the cluster effect of data objects reflected by the heterogeneous attributes, while dataset 30 is collected to study the cluster effect reflected by ordinal attributes. The 90 objects in dataset 30 correspond to the questionnaire answers of 90 intern students. The three ordinal attributes are "recognition," "willingness," and "education level." The label describes the student's final internship mode. Dataset 15 collected from the PhysioNet[4] and datasets 26–29 collected from the Weka website[5] are all public real datasets. For dataset 15, the six numerical attributes are "heart rate," "respiration rate," "systolic blood pressure," "diastolic blood pressure," "Sao2," and "age," the nominal attribute is "gender," and the label is the "overall GCS score." See Table I for the meaning of "Sao2" and "GCS." All the datasets are preprocessed by removing objects with missing value(s). In all the experiments, the number of clusters k is set to $k = k^*$.

B. Clustering Performance Evaluation

It is well acknowledged that the heterogeneity between numerical and categorical attributes is more awkward than that between the two subtypes of categorical attributes, i.e., nominal and ordinal attributes. Therefore, we first evaluate the performance of ADC on the heterogeneous attribute datasets composed of numerical and categorical attributes (i.e., datasets 1–15 in Table IV). Among the 15 datasets, datasets 4–15 are more challenging because datasets 4–11 are composed of all three types of attributes, i.e., numerical, nominal, and ordinal attributes, and datasets 11–15 are large-scale datasets. The

[3] https://archive.ics.uci.edu/ml/datasets.php

[4] https://physionet.org/content/mimiciii/1.4/

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

TABLE V

CA PERFORMANCE ON 15 MIXED DATASETS COMPOSED OF NUMERICAL AND CATEGORICAL ATTRIBUTES

Data	k-prototypes	k-means + NC	OCIL	ADC
DS	0.6463±0.09	0.6763±0.20	0.7173±0.13	**0.8380±0.07**
HF	0.6019±0.07	0.5377±0.02	0.5373±0.03	**0.6443±0.00**
AA	0.5377±0.03	0.5471±0.03	0.5221±0.03	**0.5823±0.04**
DT	0.5313±0.07	0.8030±0.09	0.6901±0.11	**0.8240±0.09**
AC	0.7939±0.04	0.7233±0.16	0.7216±0.15	**0.7942±0.00**
CC	0.3983±0.01	0.4054±0.01	0.3858±0.01	**0.4367±0.02**
CT	0.5524±0.03	0.5088±0.02	0.5196±0.04	**0.5950±0.01**
TF	0.5185±0.02	0.5594±0.01	**0.5723±0.01**	0.5652±0.00
MM	0.8205±0.00	0.7409±0.07	0.8067±0.07	**0.8313±0.00**
GC	0.5350±0.03	0.5798±0.05	0.5299±0.03	**0.6134±0.05**
AI	0.6545±0.01	0.6713±0.09	0.6584±0.08	**0.6903±0.02**
FC	0.3899±0.03	0.3963±0.03	**0.4158±0.03**	0.4140±0.02
CI	0.5065±0.07	0.5021±0.05	0.5268±0.00	**0.5365±0.05**
KC	**0.8210±0.04**	0.8017±0.03	0.8089±0.03	0.8199±0.03
MC	0.1302±0.02	0.1304±0.02	**0.1305±0.02**	0.1304±0.01
Ave. Rank	3.0000	2.9000	2.8000	1.3000

TABLE VI

ARI PERFORMANCE ON 15 MIXED DATASETS COMPOSED OF NUMERICAL AND CATEGORICAL ATTRIBUTES

Data	k-prototypes	k-means + NC	OCIL	ADC
DS	0.1100±0.16	0.1990±0.43	0.2471±0.23	**0.4715±0.13**
HF	0.0537±0.06	-0.0013±0.01	0.0016±0.01	**0.0788±0.00**
AA	-0.0008±0.01	-0.0005±0.01	-0.0074±0.01	**0.0226±0.02**
DT	0.3780±0.08	0.7683±0.10	0.6133±0.13	**0.7939±0.12**
AC	**0.3494±0.06**	0.3010±0.24	0.2875±0.23	0.3453±0.00
CC	0.0187±0.00	0.0186±0.00	0.0090±0.01	**0.0377±0.01**
CT	0.0078±0.01	-0.0175±0.01	-0.0188±0.00	**0.0184±0.01**
TF	-0.0029±0.00	0.0011±0.00	**0.0045±0.00**	0.0037±0.00
MM	0.4101±0.00	0.2506±0.12	0.3972±0.10	**0.4384±0.00**
GC	0.0040±0.01	0.0167±0.02	-0.0012±0.01	**0.0570±0.03**
AI	0.0908±0.02	0.1351±0.09	0.1184±0.08	**0.1419±0.02**
FC	0.0435±0.02	0.0529±0.03	**0.0631±0.02**	0.0615±0.02
CI	0.0372±0.03	0.0197±0.04	0.0524±0.00	**0.0573±0.04**
KC	**0.8665±0.00**	0.8563±0.00	0.8616±0.00	0.8640±0.00
MC	0.0018±0.00	0.0020±0.00	0.0022±0.00	**0.0025±0.00**
Ave. Rank	2.8667	3.0000	2.8667	1.2667

clustering performance of ADC and all the counterparts that apply to datasets 1–15 are evaluated by CA and ARI, and the corresponding results are compared in Tables V and VI. The best and the second-best results on each dataset are highlighted using **boldface** and underline, respectively. The row of "Ave. Rank" in Tables V and VI computes the average rank of the performance of the compared approaches on all the datasets and will be utilized to conduct the BD significance test. For the case that two approaches are with the same rank, we follow the common way to add 0.5 to their ranks. Since most of the existing clustering approaches are inapplicable to the challenging case that both numerical and categorical attributes exist in a dataset, we compare the other counterparts on datasets 16–30 in Section IV-E for completeness.

It can be observed from Tables V and VI that, except for the TF, FC, KC, and MC datasets, ADC outperforms all the counterparts. Even on TF, FC, KC, and MC, ADC still performs the second best, with only a tiny gap compared with the best-performing one. Another interesting finding is that the clustering performance of the compared methods on the four large-scale datasets with a larger number of true clusters, i.e., FC, CI, KC, and MC, is not worse than their performance on the other datasets. It is because a larger number of samples may provide richer statistical information to more clearly describe the distribution of clusters.

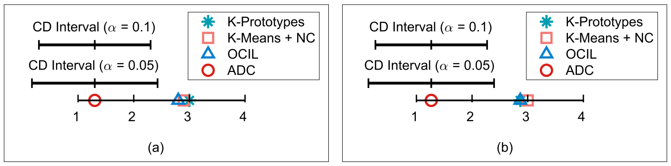

Fig. 3. BD test in (a) and (b) are based on the average performance ranks of counterparts shown in Tables V and VI. (a) BD Test on the CA performance in Table V. (b) BD Test on the ARI performance in Table VI.

To further demonstrate the superiority of ADC, BD tests are conducted on the CA and ARI performance of the compared approaches. To make the test results easy to observe, we compute the CD interval for the two-tailed BD test and visualize the test results in Fig. 3. CD intervals for the two-tailed BD tests at 0.95 confidence interval ($\alpha = 0.05$) and 0.90 confidence interval ($\alpha = 0.1$) are 1.1285 and 1.0031, respectively, for comparing four approaches on 15 datasets. In Fig. 3, a counterpart that ranks outside the right boundary of a CD interval is believed to perform significantly worse than ADC. Therefore, the BD test results in Fig. 3 show that ADC outperforms all its counterparts.

C. Ablation Study

We conduct ablation studies based on the clustering performance evaluated by the discriminative ARI index. To evaluate the effectiveness of the unified updating scheme of $\mathbf{B}$ derived in Theorem 5, we compare ADC with its variation ADC$^{\text{I}}$, which adopts the prototype selection strategy of the conventional k-prototypes clustering algorithm. To evaluate the effectiveness of the interattribute dependence-based attributes weighting in (5) and (7), we further set all the values in $\mathbf{W}$ to 1 for the GUD-based dissimilarity measurement of ADC$^{\text{I}}$, and form the variation ADC$^{\text{II}}$. To evaluate the reasonableness of exploiting interattribute relationship for computing the dissimilarities, we also generate ADC$^{\text{III}}$ based on ADC$^{\text{II}}$ by setting $s = r$ in (5), which makes that the dissimilarities of an attribute A^r can only be indicated by A^r itself. For completeness, we further make ADC$^{\text{III}}$ use only the conventional Euclidean and Hamming distances, thus forming ADC$^{\text{IV}}$. The clustering performance of ADC and its four variations are compared in Fig. 4.

It can be observed that ADC outperforms its four variations in general, which illustrates the effectiveness of ADC. More specifically, ADC outperforms ADC$^{\text{I}}$ on seven datasets and has a similar performance as ADC$^{\text{I}}$ on the rest. This indicates that the updating of $\mathbf{B}$ is effective. ADC$^{\text{I}}$ outperforms ADC$^{\text{II}}$ on nine out of the 15 datasets illustrating that the attributes weighing scheme is effective for the dissimilarity measurement of GUD. ADC$^{\text{II}}$ outperforms ADC$^{\text{III}}$ on ten out of the 15 datasets illustrating that GUD can effectively exploit the information provided by the interattribute relationship for reasonable dissimilarity measurement. Moreover, ADC$^{\text{III}}$ performs no worse than ADC$^{\text{IV}}$ on 14 out of the 15 datasets, which justifies the use of graph-based dissimilarities.

To validate the reasonableness of ADC in dealing with heterogeneous attributes, we compare the clustering performance of ADC and three different versions of it that treat ordinal and nominal attributes in different ways. The version that treats ordinal attributes as nominal ones is denoted as O2N. The two versions that adopt NC to treat ordinal attributes as Numerical ones, and both ordinal and nominal attributes as numerical ones, are denoted as O2U and ON2U, respectively. Corresponding results are demonstrated in Fig. 5.

On the former three and the latter four datasets, i.e., DS, HF, AA, FC, CI, KC, and MC, the performance of ADC is

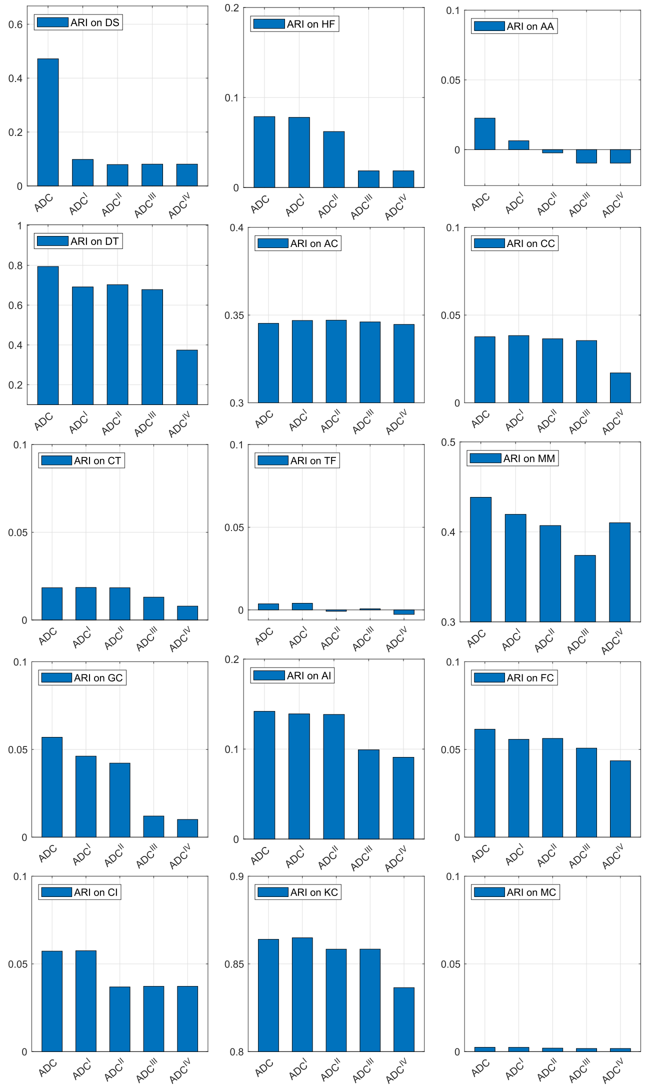

Fig. 4. Comparison of the clustering performance of ADC and its four ablated versions (i.e., ADC^{I}, ADC^{II}, ADC^{III}, and ADC^{IV}).

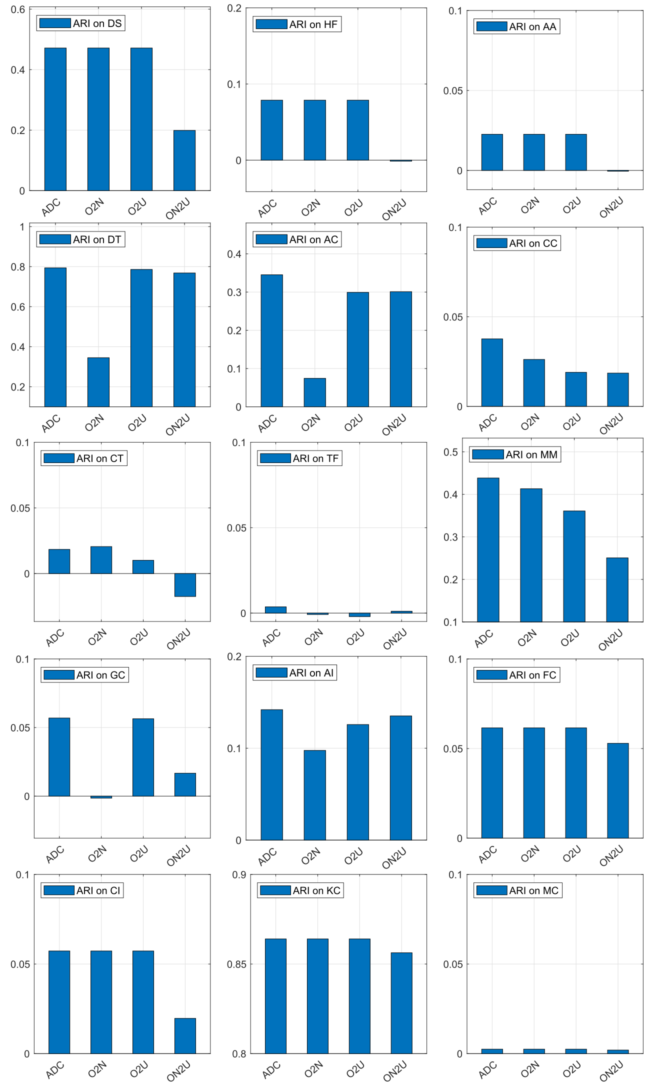

Fig. 5. Comparison of the clustering performance of ADC and its three versions that treat ordinal and nominal attributes in different ways.

the same as O2N and O2U, because ADC is equivalent to them when there is no ordinal attribute in datasets. On the remaining eight datasets, ADC performs better than O2N and O2U in general, which demonstrates the reasonableness of considering the orders of possible values, and the correctness of the dissimilarities defined for ordinal attributes, respectively. Moreover, the results show that ON2U performs worse than O2N and O2U in general. This is because ON2U involves more imprudent conversions between heterogeneous attributes.

D. Intuitive Results

To provide an intuitive impression of the dissimilarities defined by the proposed GUD metric, Fig. 6 shows the dissimilarities between attribute values of the MM dataset composed of two nominal, two ordinal, and one numerical attribute. As discussed in Remarks 1–3 and demonstrated in Fig. 1, all the $v^r(v^r)/2$ dissimilarities among v^r values of a nominal attribute A^r cannot be exactly visualize in a plane like the linearly arranged dissimilarities of ordinal and numerical attributes. Therefore, we visualize the dissimilarities from the perspective of each nominal attribute value. Since A^1 and A^2 have four and five possible values, respectively, the visualization of A^1 and A^2 are separated into four and five subfigures, i.e., $\{A^1 : o^1_1, A^1 : o^1_2, \ldots, A^1 : o^1_4\}$ and $\{A^2 : o^2_1, A^2 : o^2_2, \ldots, A^2 : o^2_5\}$, respectively. The center points of these nine subfigures are $o^1_1, o^1_2, \ldots, o^1_4, o^2_1, o^2_2, \ldots, o^2_5$, respectively. The diameters of the points indicate the frequency of such value in the attribute, and the lengths of the links indicate the value of the dissimilarity.

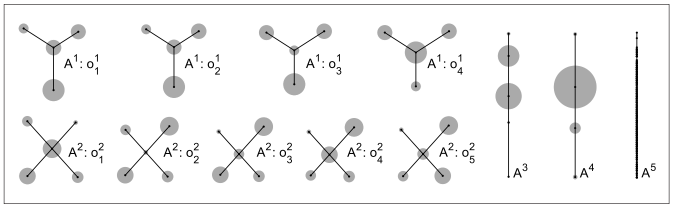

Fig. 6. Dissimilarities among the values of the attributes of MM dataset, which is composed of two nominal attributes (i.e., A^1 and A^2), two ordinal attributes (i.e., A^3 and A^4), and one numerical attribute (i.e., A^5).

T-distributed stochastic neighbor embedding (T-SNE) [44] dimensionality reduction is also adopted for intuitively illustrating the effectiveness of GUD. We first compute the dissimilarities among values of each nonnumerical attribute using GUD and then encode the attribute values by vectorizing them using the dissimilarities. Then, the encoded dataset is

TABLE VII

CA PERFORMANCE ON SIX MIXED CATEGORICAL DATASETS COMPOSED OF NOMINAL AND ORDINAL ATTRIBUTES

Data	k-modes	k-modes + LSM	k-modes + CBDM	OCIL	UDMC	HDC	ADC
BC	0.5754±0.07	0.5268±0.03	0.6198±0.08	0.5989±0.10	0.5608±0.07	0.6348±0.09	**0.6362±0.09**
HR	0.3848±0.02	0.3935±0.03	0.4045±0.04	0.3718±0.05	0.4032±0.03	0.4795±0.05	**0.4809±0.05**
LS	0.5300±0.08	0.5358±0.08	-	0.5367±0.08	0.5317±0.07	**0.5900±0.12**	**0.5900±0.12**
LG	0.5914±0.08	0.6121±0.05	0.5930±0.08	0.6154±0.07	0.6014±0.08	0.6917±0.05	**0.6928±0.05**
AE	0.5194±0.06	0.5403±0.06	0.5581±0.08	0.5175±0.08	0.5658±0.05	0.6147±0.08	**0.6219±0.08**
FT	0.4640±0.06	0.3756±0.03	0.5066±0.04	0.4900±0.04	0.4684±0.05	0.5682±0.06	**0.5710±0.06**
Ave. Rank	6.0000	5.3333	4.3333	4.6667	4.6667	1.9167	1.0833

TABLE VIII

ARI PERFORMANCE ON SIX MIXED CATEGORICAL DATASETS COMPOSED OF NOMINAL AND ORDINAL ATTRIBUTES

Data	k-modes	k-modes + LSM	k-modes + CBDM	OCIL	UDMC	HDC	ADC
BC	0.0275±0.06	0.0017±0.01	0.0645±0.07	0.0598±0.08	0.0242±0.05	**0.0785±0.10**	0.0761±0.09
HR	-0.0039±0.01	0.0046±0.03	0.0058±0.02	-0.0051±0.02	0.0073±0.02	0.0601±0.03	**0.0603±0.03**
LS	0.0791±0.11	0.1017±0.12	-	0.0966±0.10	0.0929±0.11	**0.2519±0.20**	**0.2519±0.20**
LG	0.0543±0.08	0.0532±0.04	0.0494±0.07	0.0657±0.08	0.0602±0.07	0.1505±0.06	**0.1523±0.07**
AE	0.1034±0.05	0.1077±0.05	0.1461±0.08	0.1127±0.09	0.1734±0.06	0.2611±0.08	**0.2920±0.07**
FT	0.1865±0.06	0.0755±0.04	0.2663±0.04	0.2331±0.05	0.1996±0.05	0.3553±0.07	**0.3615±0.06**
Ave. Rank	5.8333	5.6667	4.6667	4.5000	4.3333	1.7500	1.2500

TABLE IX

CA PERFORMANCE ON FOUR PURE NOMINAL AND FIVE PURE ORDINAL DATASETS

Data	k-modes	k-modes + LSM	k-modes + CBDM	OCIL	UDMC	HDC	ADC
SB	0.7668±0.14	0.7745±0.15	0.8013±0.16	**0.8154±0.17**	0.7817±0.15	0.8128±0.16	0.8128±0.15
SF	0.4737±0.06	0.4275±0.05	0.4355±0.04	0.4708±0.05	0.4539±0.05	**0.5030±0.05**	0.4921±0.05
VT	0.8622±0.01	0.8684±0.00	0.8754±0.00	0.8706±0.04	0.8652±0.00	0.8680±0.04	**0.8759±0.00**
MR	0.7671±0.12	0.7883±0.10	0.7829±0.13	0.7979±0.09	0.6170±0.06	0.7533±0.12	**0.8014±0.11**
ER	0.1898±0.01	**0.2216±0.01**	0.1873±0.01	0.1866±0.01	0.2024±0.01	0.2078±0.01	0.2081±0.01
ES	0.3646±0.04	0.3944±0.02	0.3950±0.04	0.3739±0.04	0.3786±0.03	0.4020±0.04	**0.4089±0.04**
LE	0.3266±0.03	0.3567±0.02	0.3121±0.03	0.3227±0.04	0.3589±0.03	**0.3609±0.03**	0.3589±0.03
SW	0.3760±0.03	**0.4071±0.03**	0.3765±0.02	0.3842±0.02	0.4010±0.03	0.3878±0.03	0.3885±0.03
IQ	0.5871±0.07	0.6482±0.08	0.5309±0.03	0.5631±0.06	0.6713±0.08	0.6689±0.06	**0.6809±0.04**
Ave. Rank	5.6667	3.7778	5.0000	4.4444	4.3333	3.0000	1.7778

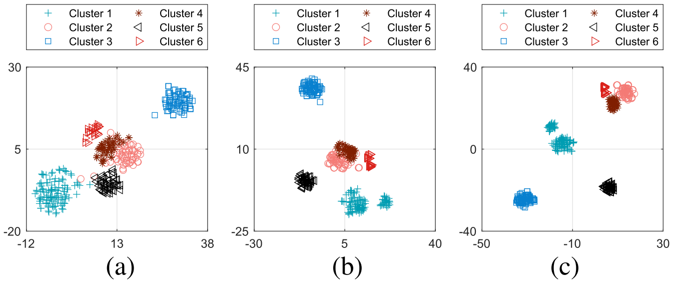

Fig. 7. t-SNE visualization of the DT dataset represented by HE, NC, and GUD. Different markers indicate true clusters of data objects. (a) HE@DT. (b) NC@DT. (c) GUD@DT.

processed by t-SNE for dimensionality reduction and visualization. The NC encoding strategy and the metric formed by combining the Hamming and Euclidean (HE) distance metrics are compared as they are commonly used for mixed data. The visualization results on the DT dataset are shown in Fig. 7(a)–(c), and the data objects belonging to different true clusters are annotated with different markers.

From Fig. 7(a), it can be seen that the true clusters 1, 2, 4, and 5 of the DT dataset are not well-separated by HE. The reason may be that HE cannot exploit the order information of the 32 ordinal attributes of DT. Since NC encodes such information, the true clusters are more separable in Fig. 7(b). However, the true clusters 2 and 4 are still overlapped. For GUD, it can be observed from Fig. 7(c) that the true clusters are more separable. Although the true clusters 2 and 4 overlap heavily in Fig. 7(a) and (b), they can still be well separated by GUD. The reason why GUD outperforms NC is that GUD not only exploits the order information of ordinal attributes but can also reasonably extract and exploit the information provided by the interdependence relationship among the attributes, even if the attributes are heterogeneous.

E. Clustering Performance on Categorical Datasets

Although ADC is proposed for heterogeneous attribute data, we claim that ADC is also competent for categorical data, including mixed categorical data composed of both nominal and ordinal attributes (datasets 16–21), pure nominal data (datasets 22–25), and pure ordinal data (datasets 26–30).

Comparative results on the mixed categorical datasets are shown in Tables VII and VIII. It can be observed that ADC outperforms all its counterparts in general, which proves the effectiveness of ADC in processing the heterogeneity of nominal and ordinal attributes. HDC also has a competitive clustering performance. It is because HDC and ADC adopt a similar graph-based dissimilarity modeling idea. However, HDC is only proposed for categorical data, which does not

TABLE X
ARI PERFORMANCE ON FOUR PURE NOMINAL AND FIVE PURE ORDINAL DATASETS

Data	k-modes	k-modes + LSM	k-modes + CBDM	OCIL	UDMC	HDC	ADC
SB	0.6680±0.18	0.6872±0.21	0.7324±0.21	**0.7592±0.22**	0.7083±0.19	0.7528±0.20	0.7567±0.20
SF	0.2141±0.07	0.1532±0.05	0.1634±0.04	0.2204±0.06	0.1887±0.06	**0.2761±0.07**	0.2370±0.06
VT	0.5236±0.02	0.5420±0.01	0.5626±0.01	0.5530±0.08	0.5325±0.01	0.5453±0.08	**0.5641±0.01**
MR	0.3375±0.22	0.3738±0.21	0.3829±0.25	0.4039±0.18	0.0705±0.07	0.3094±0.23	**0.4067±0.22**
ER	0.0103±0.01	**0.0300±0.01**	0.0105±0.01	0.0122±0.00	0.0257±0.01	0.0276±0.01	0.0278±0.00
ES	0.1555±0.03	0.2270±0.02	0.1992±0.03	0.1844±0.03	0.2109±0.02	0.2288±0.03	**0.2322±0.03**
LE	0.0322±0.02	0.0642±0.02	0.0324±0.02	0.0406±0.02	0.0644±0.02	**0.0740±0.02**	0.0738±0.02
SW	0.0466±0.02	0.0646±0.02	0.0564±0.01	0.0611±0.01	**0.0711±0.02**	0.0580±0.01	0.0572±0.02
IQ	0.0225±0.06	0.0980±0.09	-0.0035±0.01	0.0185±0.05	0.1074±0.11	0.1061±0.08	**0.1183±0.07**
Ave. Rank	6.2222	4.0000	5.0000	3.7778	4.1111	3.0000	1.8889

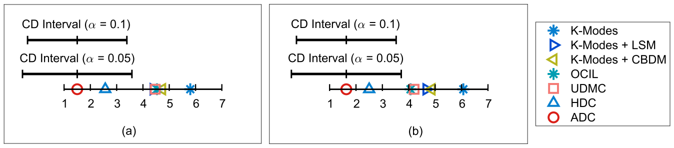

Fig. 8. BD test in (a) and (b) are based on the average performance ranks of counterparts shown in Tables VII and X. (a) BD Test on the results in Tables VII and IX. (b) BD Test on the results in Tables VIII and X.

apply to the case where there are numerical attributes. Moreover, the reason why ADC still outperforms HDC is that ADC adopts the rigorously derived updating strategy of **B**, while HDC does not. For completeness, clustering performance on pure nominal and ordinal attributes are also reported in Tables IX and X. For the datasets composed of only one type of categorical attribute, the efficacy of ADC is suppressed as ADC focuses on the processing of heterogeneous attributes. This is the reason why the superiority of ADC in the pure datasets is not as significant as its superiority in the mixed categorical datasets. Nevertheless, ADC is still very competitive compared with its counterparts on the nine pure datasets. The clustering performance of k-modes + CBDM on the lenses (LS) dataset is not reported because the attributes of LS are independent of each other, which makes CBDM fail in dissimilarity measurement.

We also conduct BD significance test to the results shown in Tables VII–X, and visualize the test results in Fig. 8. CD intervals for the two-tailed BD tests at 0.95 ($\alpha = 0.05$) and 0.90 ($\alpha = 0.1$) confidence intervals are 2.0809 and 1.8884, respectively, for comparing seven approaches on 15 datasets. It can be observed that all the counterparts excepting HDC rank outside the right boundaries of CD intervals centered on ADC, which proves the effectiveness of ADC in categorical data clustering.

V. CONCLUSION

In this article, we have proposed the GUD dissimilarity metric and ADC clustering algorithm for heterogeneous attribute data clustering. Based on the graph structures that homogeneously represent the value spaces of heterogeneous attributes, GUD is formed to reasonably quantify the dissimilarities between attribute values by sufficiently exploiting the heterogeneous information. ADC clustering algorithm, adopting GUD as its dissimilarity metric, has also been rigorously derived with a convergence guarantee. Since the GUD metric and ADC algorithm are both proposed under the guidance of the homogeneous graph space, awkward gaps among heterogeneous attributes have been novelly circumvented throughout the clustering process, thereby ensuring more accurate cluster detection. It turns out that ADC is capable of conducting cluster analysis of any-type-attributed data. Moreover, ADC is parameter-free, easy to implement, and efficient. Given dissimilarities computed by GUD, the time complexity of ADC is in the same order of magnitude as the time complexity of the conventional k-means-type clustering algorithms. Comprehensive experimental evaluations have illustrated the promising efficacy of ADC. This article has currently addressed the clustering of data objects described by the values from heterogeneous attributes. Our future work will focus on the heterogeneity issue in more complex data, including multimodal data and data with concept drifts.

REFERENCES

[1] R. V. Marinescu *et al.*, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," *NeuroImage*, vol. 192, pp. 166–177, May 2019.

[2] H. Xu, C. Ma, K. Xu, E. Chaima, and J. Lian, "Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China," *J. Hydrol.*, vol. 563, pp. 975–986, Aug. 2018.

[3] H. Zhang, X. Zhan, and B. Li, "GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation," *Nature Commun.*, vol. 12, no. 1, pp. 1–11, Dec. 2021.

[4] A. Agresti, *Categorical Data Analysis*. Hoboken, NJ, USA: Wiley, 2003.

[5] V. E. Johnson and J. H. Albert, *Ordinal Data Modeling*. Cham, Switzerland: Springer, 2006.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[7] M. Alamuri, B. R. Surampudi, and A. Negi, "A survey of distance/similarity measures for categorical data," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2014, pp. 1907–1914.

[8] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific–Asia Conf. Knowl. Discovery Data Mining*, 1997, pp. 21–34.

[9] M. A. Ben Haj Kacem, C.-E. Ben N'cir, and N. Essoussi, "MapReduce-based k-prototypes clustering method for big data," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2015, pp. 1–7.

[10] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognit.*, vol. 46, no. 8, pp. 2228–2238, 2013.

[11] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.

[12] Y. Zhang, Y.-M. Cheung, and K. C. Tan, "A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 39–52, Jan. 2020.

[13] Y. Zhang and Y.-M. Cheung, "A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 758–771, Feb. 2022.

[14] C. Zhu, Q. Zhang, L. Cao, and A. Abrahamyan, "Mix2Vec: Unsupervised mixed data representation," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2020, pp. 118–127.

[15] Y. Zhang and Y.-M. Cheung, "Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3560–3576, Jul. 2022.

[16] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in *Proc. 8th Int. Symp. Intell. Data Anal.*, 2009, pp. 83–94.

[17] D. Ienco, R. G. Pensa, and R. Meo, "From context to distance: Learning dissimilarity for categorical data clustering," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–25, Mar. 2012.

[18] H. Jia, Y.-M. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1065–1079, May 2016.

[19] S. Jian, G. Pang, L. Cao, K. Lu, and H. Gao, "CURE: Flexible categorical data representation by hierarchical coupling learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 853–866, Jun. 2018.

[20] C. Zhu, L. Cao, and J. Yin, "Unsupervised heterogeneous coupling learning for categorical representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 549–553, Jul. 2022.

[21] Y. Zhang, Y.-M. Cheung, and A. Zeng, "Het2Hom: Representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 3758–3765.

[22] S. Li, C. Gentile, and A. Karatzoglou, "Graph clustering bandits for recommendation," 2016, *arXiv:1605.00596*.

[23] J. Liang, J. Cui, J. Wang, and W. Wei, "Graph-based semi-supervised learning via improving the quality of the graph dynamically," *Mach. Learn.*, vol. 110, no. 6, pp. 1345–1388, Jun. 2021.

[24] L. Yu, L. Sun, B. Du, C. Liu, W. Lv, and H. Xiong, "Heterogeneous graph representation learning with relation awareness," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2022, doi: 10.1109/TKDE.2022.3160208.

[25] P. Arabie, N. D. Baier, C. F. Critchley, and M. Keynes, *Studies in Classification, Data Analysis, and Knowledge Organization*. Cham, Switzerland: Springer, 2006.

[26] S. Q. Le and T. B. Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2549–2557, 2005.

[27] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognit. Lett.*, vol. 28, no. 1, pp. 110–118, 2007.

[28] S. Jian, L. Cao, K. Lu, and H. Gao, "Unsupervised coupled metric similarity for non-IID categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1810–1823, Sep. 2018.

[29] Y. Zhang and Y.-M. Cheung, "Exploiting order information embedded in ordered categories for ordinal data clustering," in *Proc. 24th Int. Symp. Methodol. Intell. Syst.*, 2018, pp. 247–257.

[30] Y. Zhang and Y.-M. Cheung, "An ordinal data clustering algorithm with automated distance learning," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 6869–6876.

[31] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, Oct. 2016.

[32] S. Jian, L. Cao, G. Pang, K. Lu, and H. Gao, "Embedding-based representation of categorical data by hierarchical value coupling learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1937–1943.

[33] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," in *Proc. 2019 Int. Conf. Learn. Representation*, 2019, pp. 1–24.

[34] Kosko, "Unsupervised learning in noise," in *Proc. 34th Int. Joint Conf. Neural Netw.*, 1989, pp. 517–526.

[35] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[36] J. Xu, B. Lei, Y. Gu, M. Winslett, G. Yu, and Z. Zhang, "Efficient similarity join based on Earth Mover's distance using MapReduce," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2148–2162, Aug. 2015.

[37] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behav. Sci.*, vol. 12, no. 2, pp. 153–155, Mar. 1967.

[38] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[39] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[40] J. M. Santos and M. Embrechts, "On the use of the adjusted Rand index as a metric for evaluating supervised classification," in *Proc. 19th Int. Conf. Artif. Neural Netw.*, 2009, pp. 175–184.

[41] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[42] J. A. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3049–3076, 2017.

[43] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Yiqun Zhang (Member, IEEE) received the B.Eng. degree from the South China University of Technology, Guangzhou, China, in 2013, and the M.S. and Ph.D. degrees from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2014 and 2019, respectively.

He is currently a Lecturer with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou. His research interests include machine learning, data mining, and pattern recognition and their applications.

Dr. Zhang serves as a Reviewer for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, and *Neurocomputing*, to name a few.

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, in 2000.

He is currently the Chair Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning and visual computing and their applications.

Prof. Cheung is a fellow of the American Association for the Advancement of Science (AAAS), the Institution of Engineering and Technology (IET), and the British Computer Society (BCS). He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, and *Neurocomputing*, among others. More details can be found at: https://www.comp.hkbu.edu.hk/∼ymc.