Pattern Recognition 138 (2023) 109353

Contents lists available at ScienceDirect

Pattern Recognition

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

A generalized multi-aspect distance metric for mixed-type data clustering

Elahe Mousavi[a], Mohammadreza Sehhati[b,*]

[a] Department of Bioelectrics and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
[b] Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

Article history:
Received 9 December 2021
Revised 20 November 2022
Accepted 21 January 2023
Available online 24 January 2023

Keywords:
Clustering
Mixed data
Ordinal and nominal attribute
Inter-dependency
Intra-attribute information
Mutual information

ABSTRACT

Distance calculation is straightforward when working with pure categorical or pure numerical data sets. Defining a unified distance to improve the clustering performance for a mixed data set composed of nominal, ordinal, and numerical attributes is very challenging due to the attributes' different natures. In this study, we proposed a new measure of distance for a mixed-type data set that regards inter-attribute information and intra-attribute information depending on the type of attributes. In this regard, entropy and Jensen–Shannon divergence concepts were used to exploit the inter-attribute information of categorical-categorical and categorical-numerical attributes, respectively. Also, a modified version of Mahalanobis distance was proposed to consider the intra- and inter-attribute information of numerical attributes. We also introduced a unified framework based on mutual information to control attributes' contribution to distance measurement. The proposed distance in conjunction with spectral clustering was extensively evaluated concerning various categorical, numerical, and mixed-type benchmark data sets, and the results demonstrated the efficacy of the proposed method.

1. Introduction

Most of the clustering methods in the field of measurement scales work either with pure numerical data or pure categorical data, while by expanding research areas, observations could be described by a combination of both. Clustering this kind of data, known as a mixed-type dataset, is a challenging task mainly because of the differences in the intrinsic nature of these types of attributes [1]. Furthermore, the categorical attributes can also be a combination of nominal and ordinal attributes. Hence, interpreting their information and considering the distinction between nominal and ordinal attributes need more attention [2,3].

Although attribute conversion is one of the most straightforward approaches for clustering mixed data, it may result in information loss. Thus, designing specialized methods for clustering mixed data is more recommended [4–6]. Distance-based methods for clustering mixed data are the main focus of this study. In the following, we explore mixed-distance-based methods and, by considering the aspects mentioned for a distance measure of categorical attributes [7], propose a distance metric for mixed-type data.

The first aspect of defining a distance is exploiting the intra-attribute information, which refers to each attribute's distribution and statistical information. Unlike the Hamming distance, which merely considers the match of the categorical attribute categories, it has been proven in several studies, such as Goodall's [6] and Lin's similarity [8], that employing the distributional information of each attribute and its categories can better define the objects' distance.

The second aspect, inter-attribute information, points to the role of each attribute in the distance measurement of other attributes, especially in defining the categories' distances. In the KMCMD algorithm [9], the inter-attribute information for distance measurements of categorical attributes is considered through the conditional probability of attributes. The KMCMD results revealed the importance of utilizing attribute dependency. Nonetheless, ignoring intra-attribute information in KMCMD resulted in failed clustering for data sets with independent attributes. This point emphasizes the necessity of simultaneously utilizing intra- and inter-attribute information in distance definition.

By highlighting the ordering properties of ordinal attributes, the third aspect mentions the importance of distinguishing between nominal and ordinal attributes. In a recent study on categorical data clustering, the authors proposed a distance based on the entropy concept for combinations of nominal and ordinal attributes [2]. Following that, they utilized joint probabilities of attributes to

* Corresponding author.
E-mail address: mr.sehhati@amt.mui.ac.ir (M. Sehhati).

https://doi.org/10.1016/j.patcog.2023.109353
0031-3203/

integrate the inter- and intra-attribute information in the definition of distance [7].

The fourth aspect points to the distribution and properties of numerical attributes. As the fifth aspect, attribute weighting refers to considering a weight for each attribute according to its offered information. Considering the relevance/interdependency of attributes and providing a unified framework for applying the information of all types of attributes are the two last aspects of distance definition that can help prevent information loss.

While none of the existing mixed-type distances encompass all the mentioned aspects simultaneously, in the current study, by extending the framework of [7] for mixed-type data sets, we proposed the Generalized Unified Distance Metric for Mixed-type data (GUDMM), which includes all of the above-mentioned aspects.

In GUDMM, to quantify the intra- and inter-information of all attributes, a unified framework based on the joint probabilities of attributes was proposed. To calculate the distance between two category values in a categorical attribute, we utilized the joint pdf of each category of the target attributes and other attributes, including both numerical and categorical attributes. To apply the interdependency of attributes besides inter- and intra-attribute information for numerical attributes, we proposed a modified Mahalanobis distance for the numerical part, such that it can provide a unified framework for both numerical and categorical attributes. Similarly, the interdependency of categorical attributes is also embedded in the categorical part. Since we used the concept of entropy for the distance definition of categorical attributes, the distinction between nominal and ordinal attributes could also be preserved.

The performance of the proposed method has been experimentally evaluated on different data sets regarding the existing techniques.

The main contributions of this paper can be summarized as follows.

1. A unified framework for the distance definition of mixed-type data is proposed by considering all the aspects mentioned above for distance. In the proposed framework, the distance between pair-wise categories was calculated by using both the inter-attribute information of categorical attributes and the inter-attribute information of numerical and categorical attributes.
2. A new method for considering the interdependency of numerical attributes in addition to intra- and inter-attribute information is proposed.
3. To control how much different types of attributes depend on each other, a unified method that takes into account all categorical-categorical, categorical-numerical, and numerical-numerical relationships is introduced.
4. Clustering of mixed-type attributes has been performed based on the proposed distance measure in combination with spectral clustering.

The rest of this paper is organized as follows: In Section 2, the existing distance-based clustering methods and interdependency measures for mixed data are reviewed. Section 3 introduces the details of the proposed method. The experimental results are presented and discussed in Section 4. Finally, we state a conclusion in Section 5.

2. Overview of related work

In general, three strategies have been proposed to solve the problem of clustering mixed-type data [4]. The first strategy involves discretizing numerical attributes and applying the categorical clustering methods [5]. The potential problem of discretization is the loss of information. Furthermore, finding the optimal cut point for discretization is also controversial. The second strategy is coding categorical attributes using existing numerical clustering techniques [10]. Coding of categorical attributes is commonly performed through dummy coding for nominal attributes and integer coding for ordinal attributes. Dummy coding obscures the proximity concept, and integer coding is performed based on the presumption of equidistant categories. It should be noted that this assumption is not valid in many situations. These two strategies also ignore the statistical information of categories and attribute [7].

In the third strategy to overcome the mentioned limitations, it is suggested to preserve the intrinsic nature of each attribute. Two of the well-known schemes introduced to accomplish this goal are (I) utilizing statistical models that can deal with the issue of different types of attributes by considering different distributions for different types of attributes, and (II) applying distance measures specifically designed for mixed-type data. Model-based techniques are, however, data-dependent due to assumptions about attribute distributions. Furthermore, enforcing attribute dependency, as demonstrated by the mixture of copulas, either between categorical attributes or between numerical and categorical attributes, requires estimating unknown parameters, which becomes computationally inefficient as the number of attributes grows [11]. Thus, with a greater focus on mixed-type distances, the following subsection provides a review of mixed-type clustering methods from the standpoint of the distance aspects mentioned above, and Then, by focusing on the interdependence of attributes, various methods of defining the relevance coefficients are reviewed.

2.1. Existing mixed-type distance measures and clustering methods

One of the earliest studies for mixed data clustering is the *K*-prototypes algorithm defined based on a hybrid distance for numerical and categorical attributes. In this method, the objective function was made by the combination of Hamming and Euclidean distances with centers defined based on mode and mean for categorical and numerical attributes, respectively [12]. It has been indicated that mode is not an appropriate representative for categorical attributes [1] and the Hamming distance ignores characteristics of the attribute distributions. It also does not consider any discrimination between different categories. Some modifications to the *K*-prototype have been proposed to mitigate its drawbacks [13]. In [14], the distribution center for the categorical part has been proposed based on the frequency of occurrences of each category in a cluster. Besides, this algorithm assumed a weight for each attribute inversely related to the within-cluster distance. Other studies also investigated the proposition of a probabilistic center for categorical parts based on kernel density estimation and also utilized the information theory concept as the dissimilarity measure [15–17]. Although these partitioning methods utilized the distribution information of categorical attributes for defining the distance between each cluster and its center, the dependency and inter-attribute information were not considered.

To consider the intra-attribute information in the distance definition of mixed data, Li and Biswas proposed a similarity-based agglomerative clustering (SBAC) based on Goodall similarity [6]. Goodall similarity gives extra weight to uncommon categories by considering the distribution of attributes. Another example of utilizing intra-attribute information is the distance hierarchy, which considers a hierarchy of concepts at different levels of abstraction to define the distance between categories of categorical attributes [18]. CAVE is an incremental algorithm for mixed data clustering that utilizes variance as the similarity measure for numeric attributes and a weighted version of entropy for categorical attributes [19]. Probability distribution, variance, and entropy are all concepts that employ intra-attribute information.

Utilizing the intra- and inter-attribute information, Wang et al. proposed a representation-based method for clustering mixed data. Based on the multiplication of intra- and inter-coupling similarities found by the occurrence and co-occurrence of attribute values, the representation of categorical attributes is formed, and the inter-coupling of the categorical and numerical parts is done by discretizing the numerical part. Furthermore, the intra- and inter-attribute information of numerical attributes is computed based on the Pearson correlation [20]. In addition to the discretization problem of the numerical part, no explicit definition for the ordinal attribute was given in this method.

To highlight the importance of attribute weights, Modha and Spangler proposed a convex *K*-means algorithm that uses a weighted combination of the squared Euclidean for numerical parts and the cosine distance for dummy-coded categorical parts. Although the optimal weights of attributes are determined through a brute-force procedure, in this method, no interdependency or inter-attribute information has been considered in distance measurements [21]. Foss et al. presented the KAMILA method to balance the weight of continuous vs. categorical attributes [22,23]. KAMILA avoids restrictive parametric assumptions in a model formulation and leverages the properties of Gaussian-multinomial mixture models integrated with the *K*-means algorithm. However, due to the complexity of parameter estimation, the interdependency of attributes was ignored in the model.

Providing a distribution-based centroid for mixed data, Li et al. introduced the MCFCIW algorithm, in which the concepts of intra-cluster homogeneity and inter-cluster heterogeneity are followed by utilizing information entropy and Hellinger distance. Furthermore, the weight of each attribute in each cluster is defined in a comprehensive framework utilizing the distribution of each attribute [24]. Based on the concept of object-cluster similarity, Cheung and Jia have proposed the OCIL algorithm [25]. In an online-learning procedure, every object-cluster similarity has been computed by Euclidean distance for the numerical part and probability of occurrence for the categorical part. In this algorithm, a weight is also defined based on its entropy for each categorical attribute. Another study [26], also proposed the WOCIL method, which follows the OCIL but focuses on the computation of attribute weights in each cluster as a subspace clustering method. The proposed framework used inter-cluster distance and intra-cluster similarity to calculate attribute weights. Since WOCIL needs initial points to run, they also proposed an oriented-initialization method to provide more stable results (WOCIL + OI) [26]. None of MCFCIW, OCIL, or WOCIL considered the properties of ordinal attributes and considered no inter-attribute information in the definitions of object-center similarity.

Due to the important role of distance measurement for the categorical attributes in mixed data clustering, several categorical clustering methods are also reviewed in the following.

In an association-based dissimilarity for categorical data (ABDM) proposed by Le and Ho [27], the interdependency of attributes has been exploited to calculate the distance of two categories. The conditional distributions of all attributes concerning two investigated categories are estimated, and the distance of these distributions is reported as the distance of the categories. In this algorithm, Kullback–Leibler divergence is used for the calculation of distribution distance. Following association-based dissimilarity, the context concept has been proposed to find a subset of the more correlated attributes [28]. Context refers to the sets of attributes identified as more relevant attributes based on criteria such as symmetrical uncertainty or χ^2-test. Once the context of each attribute is found, the distance of categories is calculated only based on its context [29]. The two latter distances are formulated based on the interdependency of attributes, and when the attributes are completely independent, the distance of different categories would be zero, which is incorrect. To circumvent this problem, Jia et al. have proposed the categorical distance metric (CDM) which exploits the occurrence frequency of categories in addition to the co-occurrence frequencies of categories with other attributes [30].

While the information-theoretic dissimilarity was proposed for ordinal attributes [8], Zhang et al. proposed a new entropy-based distance metric (EBDM) that considers both ordinal and nominal attributes and the distinction between them [2]. In EBDM, the sum of the entropy of the categories was used to calculate the distance between two nominal attribute categories, whereas the cumulative entropy of all categories within the range of the two evaluated categories was used to apply the ordering nature of ordinal attributes. Furthermore, in calculating the final distance between two objects, each attribute was also weighted by its normalized entropy.

Following the EBDM, the authors proposed the unified distance metric (UDM) to impose intra- and inter-attribute information on the distance of categorical attributes [7]. Unlike the previous methods that considered the conditional probability of attributes and categories, in UDM, the joint probabilities were used to bring up the intra-attribute information along with the inter-attribute information. To control attributes' contributions to computing the distance of other attributes, the authors have also proposed a new measure of dependency based on the concordance and discordance concepts for the relevancy of different attributes (ordinal-ordinal, nominal-ordinal, and nominal-nominal).

Although some of the reviewed clustering methods are not precisely focused on the definition of distance for mixed data, however considering the total framework of clustering and according to the distance characteristics mentioned in the introduction, a brief comparison of the methods and our proposed algorithm is indicated in Table 1.

2.2. Existing interdependence measures

As mentioned in the previously introduced aspects of distance, considering the interdependency of attributes could be helpful in measuring the distance of objects. This relevance coefficient can be determined using correlation coefficients, but it is worth mentioning that the type of attribute plays an essential role in the definition of this measurement. Since our problem is composed of different attribute types (continuous, nominal, and ordinal), different association definitions are required for six cases: nominal-nominal, nominal-ordinal, nominal-continuous, ordinal-ordinal, ordinal-continuous, and continuous-continuous. Some of them are reviewed in the following.

One of the correlation coefficients is Pearson's correlation, which is suitable for attributes with a Gaussian distribution. For attributes with non-Gaussian distributions, rank correlation methods should be used. In rank correlation methods, the ordinal relationship between attributes is quantified instead of their values [31]. Some well-known rank correlation methods are as follows: Spearman's Rank Correlation, Kendall's Rank Correlation, Goodman & Kruskal's Rank Correlation, and Somers' Rank Correlation [32].

For two nominal attributes, the one general measure of association is Goodman & Kruskal's lambda, λ. While λ is a simple and easily interpretable measure, for the attributes with skewed distribution, it may not be appropriate [33]. For capturing the association of nominal and continuous attributes, the point-biserial correlation can be calculated by considering the continuous attributes and all pairs of categories of the nominal attribute [34].

Another method to measure the relatedness of attributes is the Mutual information (MI) measure, which is known as the amount of shared information between attributes. MI quantifies how different the joint distribution of two attributes is from the product of the marginal distribution of attributes. It has several advan-

Table 1
Comparison of different clustering methods from the perspective of various aspects of distance, including the intra- and inter-attribute information, the ordering properties of ordinal attributes, information of numerical attributes, attribute weighting, inter-dependency of attributes, and a unified framework for different types of attributes.

Method	Intra-attribute	Inter-attribute	Ordinal	Numerical	Weight	Interdependecy	Unified
K-prototype [12]				✓			
SBAC [6], CAVE [19]	✓			✓	✓		
KMCMD [9]		✓		✓	✓		
Wang et al. [20]	✓	✓		✓		✓	
OCIL [25], MCFCIW [24]	✓			✓	✓		
WOCIL [26]	✓			✓	✓		
ABDM [27]		✓				✓	
CDM [30]	✓	✓			✓	✓	
EBDM [2]	✓		✓		✓		
UDM [7]	✓	✓	✓		✓	✓	✓
Proposed	✓	✓	✓	✓	✓	✓	✓

Table 2
Frequently used symbols.

Symbol	Meaning	Symbol	Meaning
$\mathbf{X}$	An $N * d$ data matrix. Each row indicates a data object, and each column represents an attribute.	$\mathbf{D_c}(.,.)$	Distance between categorical parts of two data objects.
N	Number of data objects in $\mathbf{X}$.	$\mathbf{R(A_r, A_s)}$	The relevancy coefficient of A_r and A_s.
d	Number of all attribute in $\mathbf{X}$.	s_{ij}	The similarity of data objects $\mathbf{x_i}$ and $\mathbf{x_j}$.
d_u	Number of all numerical attributes in $\mathbf{X}$.	$a_{r,m}$	The mth category of the attribute A_r.
d_c	Number of all categorical attributes in $\mathbf{X}$.	$p(a_{r,mn}, A_s)$	Joint probability distributions of A_s and mth and nth categories of A_r.
d_{ord}	Number of all ordinal attributes in $\mathbf{X}$.	$E(a_{r,mn}, A_s)$	Joint Entropy of A_s and mth and nth categories of A_r.
d_{nom}	Number of all nominal attributes in $\mathbf{X}$.	$\varphi_{A_s}(a_{r,m}, a_{r,n})$	Distance between mth and nth categories of A_r according to numerical attribute A_s.
A_r	The rth attribute in $\mathbf{X}, 1 \le r \le d$.	$\phi_{A_s}(a_{r,m}, a_{r,n})$	Distance between mth and nth categories of A_r according to categorical attribute A_s.
$\mathbf{L}$	Laplacian matrix.	$D_{JS}(.\|\|.)$	Jensen–Shannon distance of two distributions.
$\mathbf{x_i}$	The ith data object in $\mathbf{X}$.	$\psi(.)$	Digamma function.
$\mathbf{x_i^u}$	The numerical part of ith data object.	$\delta_{A_r=m}$	Occurrence time of the value m in A_r.
$\mathbf{x_i^u}$	The categorical part of ith data object.	I^{uu}	Mutual information of two numerical attributes.
$\mathbf{D}(.,.)$	Distance between two data objects.	I^{cc}	Mutual information of two categorical attributes.
$\mathbf{D_u}(.,.)$	Distance between numerical part of two data objects.	I^{cu}	Mutual information of one numerical and one categorical attributes.

tages over the other association measures. While Pearson correlation only captures linear relationships and rank-based correlations are almost defined based on the concordance and discordance concepts (which refer to the monotonicity behavior of two attributes), MI detects any type of relationship between attributes, linear or non-linear. MI is also insensitive to the size of the data sets and has a straightforward interpretation [35]. Considering the advantages mentioned above and having a unified definition for different types of attributes, we utilized MI as the measure of association in our method.

3. Method

3.1. Preliminaries

Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ denotes N data objects, represented by d mixed-type attributes $A_1, A_2, \ldots, A_d$. The first d_u attributes are numerical, and the rest d_c attributes are categorical. In more detail, the d_c categorical attributes can be composed of d_{ord} ordinal attributes and d_{nom} nominal attributes, where $d = d_u + d_{ord} + d_{nom}$. Considering finite value domains for categorical attributes, $dom(A_r)$ with v_d categories can be represented with $dom(A_r) = \{a_{r1}, a_{r2}, \ldots, a_{rv_d}\}$. The main difference between ordinal and nominal attributes is that the ordinal categories are naturally ordered. For convenience, $\mathbf{X}$ could be sorted as $\mathbf{X} = [\mathbf{X^u}, \mathbf{X^{ord}}, \mathbf{X^{nom}}]$, where $\mathbf{X^u}$, $\mathbf{X^{ord}}$, and $\mathbf{X^{nom}}$ indicate the numerical, ordinal, and nominal parts of the data, respectively. Frequently used symbols are summarized in Table 2.

3.2. Spectral clustering

Given a set of N data objects $\mathbf{x_1}, \ldots, \mathbf{x_N}$ and a pre-specified definition of similarity s_{ij} between all pairs of $\mathbf{x_i}$ and $\mathbf{x_j}$, the main goal of clustering is to find groups of objects that are similar to each other while being dissimilar to other groups.

One of the well-known representations for similarities between data objects, is the similarity graph, $\mathbf{G}$, where each data object $\mathbf{x_i}$ indicate a vertex in the graph and the connection of two vertices, called edge, is weighted by s_{ij}. By constructing the similarity graph, the problem of clustering would be equal to finding the partitions such that the edges between them have very low weights and the edges within each partition have high weights. The aforementioned strategy served as the foundation for the spectral clustering proposed in Shi and Malik [36], which is summarized below.

Given the similarity matrix, $\mathbf{S} = (s_{ij})_{i,j=1,\ldots,N}$, the degree matrix of the graph, $\mathbf{D_g}$, is defined as a diagonal matrix with the degrees $d_i = \sum_{j=1}^{N} s_{ij}$, for each vertex on its diagonal. The main tool in spectral clustering is a graph Laplacian matrix, that is defined as $\mathbf{L} = \mathbf{D_g} - \mathbf{S}$. Computing the first K generalized eigenvectors of the generalized eigenvector problem $\mathbf{Lu} = \lambda \mathbf{D_g u}$ leads to the construction of matrix $\mathbf{U} \in R^{N \times K}$ whose columns are eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K$. By performing k-means clustering on the rows of $\mathbf{U}$ indicated by $\mathbf{z}_i, i = 1, \ldots, N$, clusters $Z_1, Z_2, \ldots, Z_K$ are achieved and the final clusters of data objects $\mathbf{x}_i, i = 1, \ldots, N$, are defined by clusters $C_1, \ldots, C_K$ with $C_k = \{\mathbf{x}_i | \mathbf{z}_i \in Z_k\}$.

Considering the original data $\mathbf{x_i} \in R^d$, similarities typically come from the Gaussian kernel, $s_{ij} = \exp(-||\mathbf{x_i} - \mathbf{x_j}||^2/\sigma^2)$, and make the complete graph $\mathbf{G}$. The definition of spectral clustering based

Table 3
An example mixed-type dataset.

Sample no.	Pain location (PL)	Pain frequency (PF)	Fast blood sugar (FBS)	Weight (W)
x_1	right	sometimes	120	54
x_2	right	usually	130	60
x_3	left	always	110	65
x_4	left	sometimes	120	73
x_5	midlle	usually	115	77
x_6	right	sometimes	130	63

on similarity is one of its advantages which allows us to apply it to a variety of non-vector data and also use new distances in the Gaussian kernel instead of the Euclidean distance. In the following sections, with a focus on the definition of distance for mixed-type data sets, we explain our proposed clustering method.

3.3. Distance definition

To build a unified framework for handling mixed-type data, we present a new distance metric using the dependency between different attributes and the distribution of each attribute. For each data object $\mathbf{x_i} = [\mathbf{x_i^u}, \mathbf{x_i^c}]$, the distance of the numerical part is considered as a whole, and the distance of the categorical part is calculated individually [25]. In this regard, by emphasizing on the categorical part to measure the distance value between two samples, the contribution of each categorical attribute is assumed to be equal to all numerical attributes. So the proposed distance measure would have two parts which can be weighted by $d_c + 1$ attributes, and denoted as follow:

$$\mathbf{D}(\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{d_c+1}\mathbf{D_u}(\mathbf{x_i^u}, \mathbf{x_j^u}) + \frac{d_c}{d_c+1}\mathbf{D_c}(\mathbf{x_i^c}, \mathbf{x_j^c}), \tag{1}$$

where:

$$\mathbf{D_c}(\mathbf{x_i^c}, \mathbf{x_j^c}) = \sum_{r=1}^{d_c} dist(x_{ir}^c, x_{jr}^c) \tag{2}$$

During the calculation of the distances between samples, the definitions would be illustrated by a given mixed dataset indicated in Table 3. The data is made up of two numerical attributes: fasting blood sugar (FBS) and weight (W), as well as one ordinal and one nominal attribute: the frequency of chest pain (PF) and pain location (PL). Sometimes, usually, and always are the three categories of the PF, while right, middle, and left are the three categories of the PL.

3.3.1. Distance metric for categorical attributes

Based on the concept of unified distance metric (UDM) proposed for categorical data in Zhang and Cheung [7], we introduced a new distance for mixed-type data. In order to define the distance for a categorical attribute, the first step is defining the exact distance between its pairwise categories. We denote the mth category of the attribute A_r by $a_{r,m}$, where $r \in \{1, 2, \ldots, d_c\}$ and $m \in \{1, 2, \ldots, v_r\}$.

To apply the inter-attribute information in distance calculation of the categorical part, the distance between mth and nth categories of A_r could be measured according to the distribution of A_s denoted by $\phi_{A_s}(a_{r,m}, a_{r,n})$. Furthermore, to determine the interdependence degree of A_r and A_s, which controls the contribution of A_s in the distance calculations of A_r, $\phi_{A_s}(a_{r,m}, a_{r,n})$ should be multiplied by the coefficient of $R(A_r, A_s)$. The definition of $\phi_{A_s}(a_{r,m}, a_{r,n})$ is adopted by one of the concepts of information theory that states objects with more different information are usually more dissimilar to each other.

To calculate the distance of $a_{r,m}$ and $a_{r,n}$ according to A_s, based on the intrinsic nature of A_s, meaning categorical or numerical, we introduce two definitions:

Definition 1. Given two categorical attributes A_r and A_s, the entropy-based distance between two categories m and n of A_r according to A_s is defined as Zhang and Cheung [7]:

$$\phi_{A_s}(a_{rm}, a_{rn}) = \begin{cases} E(a_{r,mn}, A_s), & \text{if } m \neq n \\ 0 & \text{if } m = n \end{cases}, \tag{3}$$

where $E(a_{r,mn}, A_s) = -\sum_{u=1}^{v_s} p(a_{r,mn}, a_{s,u}) \log_2 p(a_{r,mn}, a_{s,u})$.

To avoid the double-counting of the common information of $p(a_{r,m}, a_{s,u})$ and $p(a_{r,n}, a_{s,u})$, the joint probability $p(a_{r,mn}, a_{s,u})$ is defined for the calculation of entropy which is $p(a_{r,m}, a_{s,u}) + p(a_{r,n}, a_{s,u})$. Each $p(a_{r,m}, a_{s,u})$ is the joint probability of the co-occurrence of the mth category of attribute A_r and uth category of attribute A_s and can be defined by $p(a_{r,m}, a_{s,u}) = \delta_{rm \wedge su}(\mathbf{X})/N$, where $\delta_{rm \wedge su}(\mathbf{X})$ is the number of objects in X with their $A_r = m$ and $A_s = u$. The reason for the utilization of joint probability instead of a conditional probability is that when A_r is completely independent of all the other attributes $A_s, s = 1, \ldots, d$, the final $\Phi(a_{rm}, a_{rn}) = \sum_s \phi_{A_s}(a_{rm}, a_{rn})$, would be equal to zero which is incorrect.

Since by increasing the number of categories of a categorical attribute, its entropy also increases, it has been recommended to normalize Eq. (3). The maximum entropy of a categorical attribute is achieved when the probabilities of occurrence of all the categories are the same. So, for the attribute A_s by v_s categories, $S_{A_s} = -\log_2(v_s)$ is the standard entropy. Dividing $E(a_{r,mn}, A_s)$ by S_{A_s} or calculating the entropy of Eq. (3) in base v_s, the mentioned problem would be solved and, the definition of ϕ_{A_s} is redefined as:

$$\phi_{A_s}(a_{rm}, a_{rn}) = \begin{cases} \dfrac{E(a_{r,mn}, A_s)}{S_{A_s}}, & \text{if } m \neq n \\ 0 & \text{if } m = n \end{cases}, \tag{4}$$

As mentioned in Zhang and Cheung [7], considering the ordering concept between categories of ordinal attributes, more information could be achieved from ordinal attributes, which could be applied to the definition of the distance between their categories. In other words, by summing the entropy over the range of two categories of an ordinal attribute, the ordering nature of their categories is also preserved, and the definition of distance is redefined as:

$$\phi_{A_s}(a_{rm}, a_{rn}) = \begin{cases} \dfrac{\sum_{g=\min(m,n)}^{\max(m,n)-1} E(a_{r,g(g+1)}, A_s)}{S_{A_s}}, & \text{if } m \neq n,\ r\text{: ordinal} \\ \dfrac{E(a_{r,mn}, A_s)}{S_{A_s}}, & \text{if } m \neq n,\ r\text{: nominal} \\ 0 & \text{if } m = n \end{cases} \tag{5}$$

Example 1. For the dataset mentioned in Table 3, for every categorical attribute, the distances between its categories should be calculated. The variation of PF, W, and FBS that coincides with the PL categories must be considered when calculating the distances between each of the two PL categories (right-middle, right-left, and middle-left). In the following, for instance, the distance between the left and right categories according to the other *categorical attributes* (i.e., PF) is calculated. Thus, first, the joint pdf of each category of PL and PF would be calculated. After that, according to the pdfs' variation of PF regarding the PL categories and considering the nominal nature of PL, differences between PL categories could be calculated utilizing Eq. (5) as follows:

$$\phi_{A_{s=PF}}(A_{rm} = \text{left}, A_{rn} = \text{right}) = \sum_{u=1}^{v_s} \frac{E(A_{rmn}, A_{s_u})}{S_{A_s}}$$

where, v_s according to the categories af PF {sometime, usually, always} equals three and $S_{A_s} = -\log(3)$, and $E(A_{rmn}, A_s = \text{PF})$ equals to:

$$\begin{aligned} E(A_{rmn}, A_s = \text{PF}) &= -\sum_{u=1}^{v_s} p(A_{rmn}, A_{su}) \log p(A_{rmn}, A_{su}) \\ &= -\sum_{u=1}^{v_s} p(A_{rm} = \text{left}, A_{su}) \log p(A_{rm} = \text{left}, A_{su}) \\ &\quad -\sum_{u=1}^{v_s} p(A_{rm} = \text{right}, A_{su}) \log p(A_{rm} = \text{right}, A_{su}) \\ &= -(1/6 \log(1/6) + 0 + 1/6 \log(1/6)) \\ &\quad -(1/3 \log(1/3) + 1/6 \log(1/6) + 0) \end{aligned}$$

The above-mentioned formula for calculation of distance between different categorizes of a categorical attribute should be performed based on all of the other attributes, i.e., $s = 1, \ldots, d$. The described $\phi_{A_s}(a_{rm}, a_{rn})$ is based on the assumption of the categorical nature of A_s. In the following, we propose the calculation of $\phi_{A_s}(a_{rm}, a_{rn})$ in case A_s is a numerical attribute.

Definition 2. Given a categorical attributes A_r and a numerical attribute A_s, the entropy-based distance between two categories m and n of A_r according to A_s is defined as Zhang and Cheung [7]:

$$\varphi_{A_s}(a_{r,m}, a_{r,n}) = D_{JS}(p(a_{r,m}, A_s) || p(a_{r,n}, A_s)) \tag{6}$$

where D_{JS} is the Jensen–Shannon Distance (JSD).

The start of information theory was based on the introduction of Shannon entropy, i.e., entropy for the categorical cases. For a numerical random variable $\mathbf{x}$ with density $f(x)$, the differential entropy $h(\mathbf{x})$ is defined as:

$$h(\mathbf{x}) = \int_S f(x) \log(\frac{1}{f(x)}) dx \tag{7}$$

where S shows the support set of $\mathbf{x}$. In the numerical case, however, this definition is not scale-invariant [37]. It could not help to define the distance between two categories $a_{r,m}$ and $a_{r,n}$ based on the numerical attribute A_s indicated by $\varphi_{A_s}(a_{r,m}, a_{r,n})$. In this situation, relative entropy between distributions could be useful. The relative entropy, also known as Kullback–Leibler (KL) divergence for the calculation of distance between two probability density functions (pdf) f and g, is defined as:

$$D_{KL}(f \| g) = \int_S f(x) \log_2 \frac{f(x)}{g(x)} dx, \tag{8}$$

Since we want to use the entropy concept to define the distance and $D_{KL}(f||g)$ is not symmetric, we use one of its symmetrization, which is also a metric, called JSD defined by Nielsen [38]:

$$\begin{aligned} D_{JS}(f \| g) &= \sqrt{\tfrac{1}{2}\left(D_{KL}\left(f \middle\| \tfrac{(f+g)}{2}\right) + D_{KL}\left(\tfrac{(f+g)}{2}\right) \| g\right)} \\ &= \left(\tfrac{1}{2} \int_S \left(f(x) \log_2 \tfrac{2f(x)}{f(x)+g(x)} + g(x) \log_2 \tfrac{2g(x)}{f(x)+g(x)}\right) dx\right)^{1/2} \end{aligned} \tag{9}$$

The D_{JS} is symmetric and bounded for two probability distributions, such that $0 \le D_{JS}(f||g) \le 1$. During the utilization of D_{JS}, it is important to notice that $f(x)$ and $g(x)$ should be measured on the same set of measurements.

To calculate $\varphi_{A_s}(a_{r,m}, a_{r,n})$ for a numerical attribute A_s the JSD of two joint probability distributions $p(a_{r,m}, A_s)$ and $p(a_{r,n}, A_s)$ should be calculated. But, $p(a_{r,m}, A_s)$ and $p(a_{r,n}, A_s)$ are unknown and should be estimated from the data. Here, we use the kernel density estimation (KDE) method [39] to solve the problem, summarized in the following.

Given a sequence of N data points $\{x_1, x_2, \ldots, x_N\}$ from a probability density function $f(x)$, KDE estimates $f(x)$ at a particular point x, by performing a weighted average on the given data points by kernel $K(.)$ with bandwidth parameter h as follows:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{10}$$

The only condition on the kernel function $K(x)$ is normalization: $\int_{-\infty}^{+\infty} K(x) = 1$. We use the Gaussian kernel with the introduced optimal bandwidth mentioned in Silverman [39], equal to: $h_{opt} = 1.06\sigma n^{-1/5}$, where σ is the standard deviation of given data points.

Utilizing the explained KDE and normalizing the estimated density, estimation of $p(a_{r,m}, A_s)$ and $p(a_{r,n}, A_s)$ could be performed. By sampling over the full range of A_s, the values of $p(a_{r,m}, A_s)$ and $p(a_{r,n}, A_s)$ on the same set of measurement are obtained, so that the calculation of distance $\varphi_{A_s}(a_{r,m}, a_{r,n})$ using D_{JS} can be practicable and calculated by (6).

In fact, by measuring the variation of $A_{s=1,..,d}$ coincided with the variation of A_r from the category m to n, the distance between these two categories would be estimated from the view of other attributes.

Example 2. In the following of Example 1, it is required to calculate the joint pdf of each category of PL and *numerical attributes* (i.e., FBS and, W) based on the Definition 2. The estimation of the joint pdf of FBS and each of the categories of PL (left, middle, and right) could be performed utilizing the kernel density estimation method mentioned in the Eq. (10). An example of these pdfs indicated in Fig. 1. By sampling along the estimated joint density functions of $f(As = \text{FBS}, A_r = \text{right})$, $f(As = \text{FBS}, A_r = \text{middle})$, and $f(As = \text{FBS}, A_r = \text{left})$, pair-wise distances of categories according to $A_s = \text{FBS}$ could be calculated by JSD mentioned in Eq. (6). It is worth mentioning that the number of samples should not be small for density estimation. Thus, for the six samples in this example, the estimation of densities is not possible, and the figures are just indicated schematically. This step should be performed for other numerical attributes, meaning W.

Based on the pairwise distances between the different categories of a categorical attribute, the distance for a categorical attribute $\mathbf{D_c}(\mathbf{x_i}, \mathbf{x_j})$ is defined by combining Eqs. (5) and (6); however, an interdependency coefficient is required to controls the contribution of A_s in distance definition of attribute A_r that gives more weight to more dependent attributes. In Section 3.3.3 R would be introduced in details. Thus, the final distance between two objects for a categorical attribute is defined as:

$$\begin{aligned} &dist(x_{ir}^c, x_{jr}^c) \\ &= \frac{1}{d}\left(\sum_{s}^{d_c} R(A_r, A_s) \phi_{A_s}(a_{ri}, a_{rj}) + \sum_{s}^{d_u} R(A_r, A_s) \varphi_{A_s}(a_{ri}, a_{rj})\right) \end{aligned} \tag{11}$$

where a_{ri} and a_{rj} indicate the value of attribute A_r for samples $\mathbf{x_i}$ and $\mathbf{x_j}$, respectively.

3.3.2. Distance metric for numerical attributes

One of the main purposes of this paper was to investigate the effect of applying the relationship of attributes and their distribution on the distance calculations of mixed-type data. To achieve this goal, for the numerical part of the distance in Eq. (1), we proposed a modified version of Mahalanobis distance (MD). MD is one of the most well-known distance measures for numerical attributes that takes into account the probabilistic information of the data. In order to calculate the MD for two samples with d_u numerical attributes, the first step is to calculate the covariance matrix S. Indeed, to apply the inter-attribute information to the distance calculation, we can consider the 2-dimensional pdfs of two-by-two attributes, so S includes the variances of all the pdfs. MD is defined as:

$$MD(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^T M (\mathbf{x_i} - \mathbf{x_j})} \tag{12}$$

where M indicates the inverse covariance matrix.

Similar to the distance definition of categorical attributes, the interdependency degree of attributes in forming the distance of

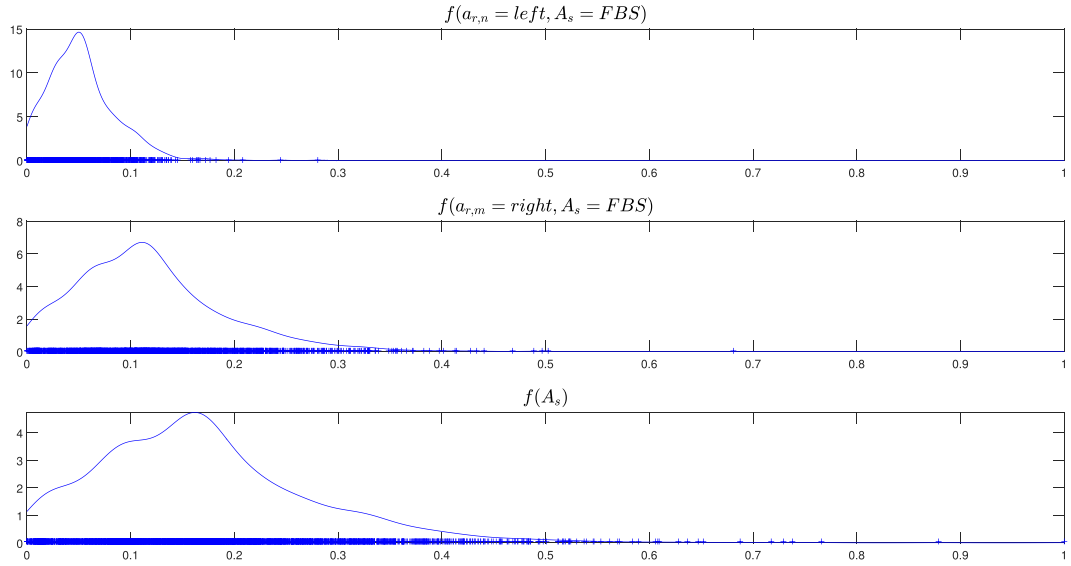

Fig. 1. Comparison of variations in estimated density function $f(A_s = \text{FBS})$ corresponding to the variation of two categories $n = \text{left}$ and $m = \text{right}$ of $A_r = \text{PL}$. The markers "+" on the horizontal axes show the samples of each distribution.

one other variable could be determined through a relevancy coefficient. In other words, when the interdependency of two attributes A_r and A_s is not strong, applying the information offered by A_s on the distance calculation of the A_r is not reliable and can be ignored or given a weight according to their relevancy. By considering this idea, a modified version of MD is proposed as follow:

$$MD^*(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^T \Delta (\mathbf{x_i} - \mathbf{x_j})} \tag{13}$$

where,

$$\Delta = R \odot M \tag{14}$$

The $\odot$ indicates the Hadamard product of M and R and, R is the relevance matrix between pairwise attributes.

In the next subsection, the definition of relevancy in a unified concept for both the categorical and numerical attributes will be described.

3.3.3. Definition of relevance coefficient

Based on the concept of mutual information (MI), we have defined the relevance of two attributes A_r and A_s noted by $R(A_r, A_s)$. $R(A_r, A_s)$ was utilized as a relevance coefficient in Eqs. (11) and (14) for categorical and numerical part of the distances between samples, respectively. However, similar to the distance definition, the intrinsic nature of attributes plays an important role in the definition of the relevance coefficient. So, considering the type of variable, we explain the MI-based relevance coefficient in the following three ways:

Definition 3. Given two categorical attributes, A_r and A_s, the MI-based relevance coefficient can be defined by Eshima [40]:

$$I^{cc}(A_r, A_s) = \sum_{g=1}^{v_r} \sum_{u=1}^{v_s} p(A_r = g, A_s = u) \log_2 \frac{p(A_r = g, A_s = u)}{p(A_r = g) p(A_s = u)} \tag{15}$$

where all of the probabilities are calculated on the whole of the dataset $\mathbf{X}$.

When A_r, A_s are both discrete or in our case categorical, the probabilities can be estimated conveniently by counting the number of times each pair occurs in the data, for example, $p(A_r = g, A_s = u) = \delta_{A_r = g \wedge A_s = u}(\mathbf{X})/N$.

Definition 4. Given a categorical attribute A_r and a numerical attribute A_s, the MI-based relevance coefficient can be defined by averaging on the MI of all samples, $\{I_i, i = 1, \ldots, N\}$ as follow [35]:

$$I^{cu}(A_r, A_s) = \langle\, I_i \,\rangle = \psi(N) - \langle \psi(N_{r_i}) \rangle + \psi(k) - \langle \psi(m_i) \rangle \tag{16}$$

in which, m_i is the number of samples with A_s in the distance of d_{knn} around sample $\mathbf{x_i}$, d_{knn} is the distance between kth nearest neighbor to sample $\mathbf{x_i}$ among those N_{ri} data points whose value of the categorical attribute equals to a_{ri}, $\{\mathbf{X}(A_s | A_r = a_{ri})\}$, and $\psi(.)$ is the digamma function.

Definition 5. Given two numerical attributes A_r and A_s, the MI-based relevance coefficient can be defined based on two parameters N_{r_i} and N_{s_i} for all samples, $\mathbf{X} = \{\mathbf{x_i}, i = 1, \ldots, N\}$, as follow [41]:

$$I^{uu}(A_r, A_s) = \psi(k) - \langle \psi(N_{r_i} + 1) + \psi(N_{s_i} + 1) \rangle + \psi(N). \tag{17}$$

in which N_{r_i} and N_{s_i} are the number of data points $\mathbf{X}(A_r)$ and the number of data points $\mathbf{X}(A_s)$ within the distance $\epsilon_i/2$, respectively. The ϵ_i is $\max\{\epsilon_{s_i}, \epsilon_{r_i}\}$, where $\epsilon_{r_i}/2$ and $\epsilon_{s_i}/2$ denotes the distance from $\mathbf{x_i}(A_r)$ and the distance from $\mathbf{x_i}(A_s)$ to their kth neighbor, respectively.

Since numerical attributes can be discrete or continuous, it is recommended to consider the discrete attribute as categorical in the calculation of MI [35,41].

In order to normalize the values of MI-based relevance coefficients the estimated relevance coefficients has been divided by the maximum MI of each attribute which equals $H(A_r) = MI(A_r, A_r)$, so that the relevancy of each attribute to itself equals 1. The final MI-based relevance matrix R could be defined as:

$$R(A_r, A_s) = \begin{cases} I^{cc}(A_r, A_s)/I^{cc}(A_r, A_r), & \text{if } A_r \& A_s: \text{ categorical} \\ I^{cu}(A_r, A_s)/I^{cu}(A_r, A_r), & \text{if } A_r: \text{ categorical}, A_s: \text{ numerical} \\ I^{uu}(A_r, A_s)/I^{uu}(A_r, A_r), & \text{if } A_r \& A_s: \text{ numerical} \end{cases} \tag{18}$$

Example 3. After completing the pair-wise distance calculation of categories of PL, the relevance degree of attributes (R) must be calculated. Hence, according to the Definitions 3, and 4 and considering the type of attributes, $R(A_s = \text{PF}, A_r = \text{PL})$, $R(A_s = \text{FBS}, A_r = \text{PL})$, and $R(A_s = \text{W}, A_r = \text{PL})$ could be calculated by Eq. (18).

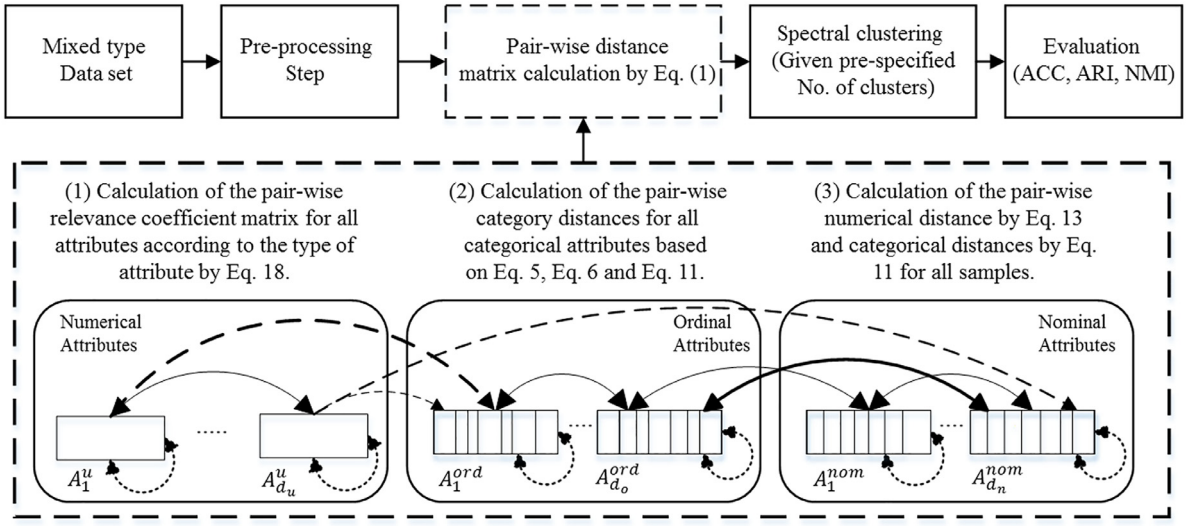

Fig. 2. Flowchart of GUDMM-S clustering. Solid, dashed, and dotted lines represent the inter-relationship of two categorical attributes or two numerical attributes, a categorical and a numerical attribute, and intra-attribute information, respectively. The thickness of lines, schematically, points to the strength of the interdependency of attributes.

Example 4. Considering the FBS and W as numerical attributes, the calculation of the covariance matrix of numerical attributes and the relevance of numerical attributes according to Definition 5 are required to calculate the other part of the distance, which refers to the numerical attributes. So, after the calculation of $R(A_s = \text{FBS}, A_r = \text{W})$ by Eq. (18), the calculation of the total numerical part of the distance between each two samples will be performed by Eq. (13).

During distance calculation between two samples, for the categorical part, according to Eq. (11), we sum across all attributes and consider the joint entropy of attributes. However, when $s = r$, the inter-attribute turns into the intra-attribute information, which considers the pdf of each attribute by itself. The same happens for numerical attributes when the numerical part is multiplied by the Δ using Eq. (14). Multiplication of the numerical part of the data by the diagonal elements points to the distributional information of each numerical attribute, i.e., intra-attribute, while multiplication by non-diagonal elements refers to inter-attribute information.

It is straightforward to demonstrate that the proposed distance, namely GUDMM satisfies the four following properties of a distance metric:

1. Non-negativity: $0 \leq \mathbf{D}(\mathbf{x_i}, \mathbf{x_j})$.
2. Symmetry: $\mathbf{D}(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{D}(\mathbf{x_j}, \mathbf{x_i})$.
3. Identity of indiscernibles: $\mathbf{D}(\mathbf{x_i}, \mathbf{x_j}) = 0 \Leftrightarrow \mathbf{x_i} = \mathbf{x_j}$.
4. Triangle Inequality: $\mathbf{D}(\mathbf{x_i}, \mathbf{x_j}) \leq \mathbf{D}(\mathbf{x_i}, \mathbf{x_k}) + \mathbf{D}(\mathbf{x_k}, \mathbf{x_j})$.

3.4. Parameter setting of spectral clustering

One important point about the spectral clustering method is its dependency on the scaling parameter σ when the Gaussian similarity function is used. σ controls the decaying rate of points' similarity by distance. Various strategies have been proposed for determining the optimal σ. By examining a range of pre-specified values for σ, it can be choose according to the results of clustering evaluated by an internal validity metric. Furthermore, by considering a local scaling parameter σ_i for each data object $\mathbf{x_i}$, it can be selected according to the distance of $\mathbf{x_i}$ to its k-nearest neighbors, so that the Gaussian similarity function form as, $s(\mathbf{x_i}, \mathbf{x_j}) = \exp(-d^2(\mathbf{x_i}, \mathbf{x_j})/\sigma_i\sigma_j)$ [42].

3.5. The proposed GUDMM-S algorithm

Algorithm 1 contains the pseudo-code for the GUDMM-S algorithm. The GUDMM-S takes a mixed dataset X and a predefined parameter K as input. For each dataset X, the relevancy matrix R is calculated once using Eq. (18). Then the pair-wise distances of categories for all categorical attributes using Eq. (11), and Δ for numerical attributes using Eq. (14) are calculated. The pair-wise distance matrix for all samples could be computed using Eq. (1). Utilizing the Gaussian function, distances could be transformed into the similarity matrix and considered as the input of spectral clustering. Figure 2 indicates the flowchart of the proposed method and a simple representation of the intra- and inter-attribute information in distance measurement.

Algorithm 1 Spectral clustering based on generalized unified distance metric for mixed-type data (GUDMM-S).

Input: dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, number of clusters K, and scaling parameter σ.
Output: cluster label $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$.

for $r = 1$ to d **do**
  for $s = 1$ to d **do**
    Calculate $R(A_r, A_s)$ according to Eq. (18).
for $r = d_u + 1$ to d **do**
  for $m = 1$ to $\nu_r - 1$ **do**
    for $n = m + 1$ to ν_r **do**
      Calculate $dist(x^c_{r,m}, x^c_{r,n})$ according to Eq. (11).
      $dist(x^c_{r,n}, x^c_{r,m}) = dist(x^c_{r,m}, x^c_{r,n})$
Calculate Δ matrix by Eq. (14).
Calculate pairwise distance $\mathbf{D}(\mathbf{x_i}, \mathbf{x_j})$ between data objects by Eq. (1).
Calculating similarity matrix $\mathbf{S}$ using Gaussian function by computed $\mathbf{D}$, and pre-specified σ.
Performing spectral clustering on $\mathbf{S}$.
return cluster label $\mathbf{y}$.

4. Experiments

In order to investigate the performance of GUDMM-S for mixed-type data, we compared the obtained clustering results with

Table 4
Description of the utilized data sets.

Mixed data set				Categorical dataset.				Numerical data set			
Dataset	N	$d_u + d_c(d_o + d_n)$	K	Dataset	N	$d_c(d_o + d_n)$	K	Dataset	N	d_u	K
Heart	303	$7 + 6(2 + 4)$	2	Soybean	47	$35(0 + 35)$	4	Iris	150	4	3
Credit	653	$9 + 6(1 + 5)$	2	Voting	435	$16(0 + 16)$	2	Wine	178	13	3
Dermatology	366	$1 + 32(32 + 0)$	6	WBCD	699	$9(9 + 0)$	2	Ionosphere	351	33	2
German	1000	$13 + 7(2 + 5)$	2	Car	1728	$6(6 + 0)$	4	Sonar	208	60	2
Adult	4000	$8 + 6(0 + 6)$	2	Lym	148	$18(3 + 15)$	4	Yeast	1484	8	10
A-Credit	690	$8 + 6(2 + 4)$	2	Mass	824	$4(2 + 2)$	2	Ecoli	336	6	8

the existing counterparts on various mixed, pure categorical and pure numerical data sets. The source code of the proposed method is available in https://github.com/Elmsvi-Git/GUDMM-S.git.

4.1. Utilized data sets

To provide a comprehensive framework for investigating the clustering performance of GUDMM-S, we utilized 18 real and benchmark data sets, including six mixed, six pure categorical, and six pure numerical types of data sets from the UCI machine learning data repository. The mixed-type dataset includes heart disease (abbreviated as Heart), Credit approval (abbreviated as Credit), German Credit (abbreviated as German), Dermatology, Adult, and Australian Credit Approval (abbreviated as A-Credit).

Pure categorical data sets were selected from those commonly used for the evaluation of categorical data clustering methods. The categorical data sets are Soybean-small (abbreviated as Soybean), Congressional Voting Records (abbreviated as Vote), Breast Cancer Wisconsin (abbreviated as WBCD), Car, Lymphography (abbreviated as Lym), and Mammographic Mass (abbreviated as Mass). The first two data sets have only nominal attributes; the third and fourth ones have only ordinal attributes; and the last two include both nominal and ordinal attributes. Pure numerical data sets are Iris, Wine, Ionosphere, Sonar, Yeast, and Ecoli. The detailed information of the data sets is summarized in Table 4.

4.2. Performance evaluation measures

To evaluate the performance of the proposed method and provide a comprehensive comparison to other state-of-the-art methods, we used three external evaluation measures, including Clustering Accuracy (CA), Rand Index (RI), Normalized Mutual Information (NMI). Furthermore, we used two internal validation measures to determine the optimal value of the Gaussian similarity function parameter σ, including Davies-Bouldin (D), and Xie-Beni (XB) indices.

4.3. Experimental setting

4.3.1. Preprocessing

To scale the numerical attributes in range [0,1], we normalized their values by $(x_{ij} - x_{i,\min})/(x_{i,\min} - x_{i,\max})$. Except for Mass data which we utilized the clean dataset provided by Zhang and Cheung [7], other data sets had no or a small number of missing values. In the latter cases, missing values of categorical and numerical attributes were replaced by the mode and mean of attributes, respectively. However, there are other methods for handling missing values, which propose imputing missing values into the clustering of mixed data [43]. The details of the spectral clustering setting is provided in Supplementary Materials 1.

4.4. Experiments on mixed-type data sets

Firstly, to evaluate the clustering performance of the GUDMM-S algorithm on mixed-type data, we compared it with *K*-prototype [12], KMCMD [9], OCIL [25], WOCIL [26], and WOCIL-OI [26] algorithms. Since *K*-prototype, KMCMD, OCIL and WOCIL need random initialization, each algorithm has been executed 50 times and the results are reported statistically. Although Spectral clustering is almost stable, due to the randomized nature of *k*-means in the last step of spectral clustering and the added noise during the calculation of MI for numerical attributes, we also repeated our algorithm ten times. The distance regulation parameter γ of the *K*-prototype algorithm, as it was proposed in Ahmad and Dey [9], was set to $\sigma/2$, where σ is the average standard deviation of numerical attributes. The number of intervals in the KMCMD algorithm is set to 5 or the average number of categories of categorical attributes. For KMCMD, and *K*-prototype algorithms we normalized the numerical attributes by $(x_{ij} - x_{i,\min})/(x_{i,\min} - x_{i,\max})$. For OCIL, WOCIL, and WOCIL-OI, as previously proposed, the numerical part is normalized by $(x_{ij} - m_j)/\sigma_j$, where m_j and σ_j show the mean and standard deviation of the jth numerical attribute.

Table 5 shows the clustering results in terms of CA, RI, and NMI. From the results, it can be observed that the GUDMM-S algorithm outperforms the WOCIL + OI, WOCIL, OCIL, KMCMD, and *K*-prototype algorithms in most cases, so that in four cases it gets the first rank and in two cases the second rank is achieved. Since the main idea of GUDMM-S is based on considering the interdependency of attributes in distance calculation, it can be observed that considering the dependency could improve the clustering results. Furthermore, the stability of the GUDMM-S results is notable, making the discovered labels reliable. As shown in Table 4, we used mixed data sets made up of distributions with varying numbers of numerical, ordinal, and nominal attributes, and our algorithm performed well in all cases.

4.5. Experiments on pure categorical data sets

Table 6 indicates the performance of GUDMM-S in comparison with five methods for clustering mixed-type data. It can be observed that the performance of GUDMM-S for categorical data is also remarkable, so that it almost achieves the first or second rank, and for cases such as Vote and WBCD, where the second score has been achieved, the difference with the first rank is trivial. Furthermore, although the KMCMD algorithm outperformed GUDMM-S for the Lym dataset, for the Car dataset, whose attributes are completely independent of each other, the calculation of distance and final clustering has failed. It occurred as the result of conditional probability utilization in the calculation of categorical distance. For more investigation, we compared the GUDMM-S algorithm with three state-of-the-art categorical data clustering methods, including CDM [30], EBDM [2], and UDM [7].

All of these methods are focused on the modification of distance measurements. Hence, we combined them with the spectral clustering method to provide a fair comparison. According to the results in Fig. 3, GUDMM outperformed other distance measures in most of the cases. Compared to UDM and EBDM, it outperformed them for Car and Lym data sets and showed similar performance for the rest. Although both UDM and GUDMM use the same pro-

Table 5
Clustering performance of six clustering algorithms on six mixed-type data sets. The first rank and second rank are indicated by boldface and underline, respectively.

Index	Dataset	GUDMM-S	WOCIL + OI	WOCIL	OCIL	KMCMD	*K*-prototype
CA	Heart	**0.8291 ± 0.0076**	0.8086 ± 0	0.7043 ± 0.0754	0.8284 ± 0.0059	0.8238 ± 0.0076	0.7874 ± 0.0656
	Credit	**0.8413 ± 0.0066**	0.5452 ± 0	0.6697 ± 0.1435	0.7798 ± 0.1086	0.7447 ± 0.154	0.7641 ± 0.0822
	Dermatology	**0.9672 ± 0.0017**	0.4918 ± 0	0.5042 ± 0.0929	0.7057 ± 0.0793	0.7228 ± 0.1274	0.5735 ± 0.1015
	German	**0.6938 ± 0.0086**	0.675 ± 0	0.5531 ± 0.049	0.5214 ± 0.0009	0.6448 ± 0.0335	0.5565 ± 0.0361
	Adults	0.7078 ± 0	0.5105 ± 0	0.5841 ± 0.0899	0.5105 ± 0	**0.721 ± 0.0034**	0.6447 ± 0.0314
	A-Credit	0.8311 ± 0.0021	**0.8551 ± 0**	0.7368 ± 0.1386	0.7892 ± 0.1163	0.7801 ± 0.0481	0.8028 ± 0.0424
RI	Heart	**0.7157 ± 0.01**	0.6894 ± 0	0.5934 ± 0.0552	0.7148 ± 0.0278	0.7089 ± 0.0098	0.6727 ± 0.0494
	Credit	**0.7327 ± 0.0091**	0.5033 ± 0	0.5996 ± 0.1182	0.6796 ± 0.0855	0.6667 ± 0.1232	0.6461 ± 0.0651
	Dermatology	**0.9815 ± 0.0012**	0.7433 ± 0	0.6945 ± 0.1036	0.8821 ± 0.0381	0.8897 ± 0.0574	0.8177 ± 0.0488
	German	**0.5748 ± 0.0066**	0.5608 ± 0	0.51 ± 0.0171	0.5004 ± 0.0001	0.5437 ± 0.0204	0.5085 ± 0.0096
	Adults	0.5862 ± 0	0.5001 ± 0	0.5302 ± 0.0392	0.5001 ± 0	**0.5976 ± 0.0032**	0.5438 ± 0.0163
	A-Credit	0.7189 ± 0.0027	**0.7518 ± 0**	0.65 ± 0.1085	0.6939 ± 0.0851	0.661 ± 0.0443	0.6865 ± 0.0378
NMI	Heart	**0.3595 ± 0.0196**	0.293 ± 0	0.1459 ± 0.0811	0.3313 ± 0.0409	0.3279 ± 0.0153	0.2663 ± 0.0762
	Credit	**0.3728 ± 0.0151**	0.023 ± 0	0.1779 ± 0.1997	0.2834 ± 0.1353	0.2754 ± 0.1867	0.2358 ± 0.0925
	Dermatology	**0.9348 ± 0.0041**	0.5870 ± 0	0.5701 ± 0.0922	0.7645 ± 0.0415	0.8194 ± 0.0727	0.5616 ± 0.0987
	German	**0.0178 ± 0.0044**	0.009 ± 0	0.0039 ± 0.0044	0.001 ± 0	0.0058 ± 0.0077	0.0072 ± 0.0064
	Adults	0.141 ± 0	0.036 ± 0	0.0566 ± 0.0611	0.036 ± 0	**0.1823 ± 0.0259**	0.123 ± 0.0368
	A-Credit	0.3512 ± 0.0047	**0.428 ± 0**	0.2505 ± 0.1854	0.3056 ± 0.1344	0.2989 ± 0.0725	0.2895 ± 0.0612

Table 6
Clustering performance of six clustering algorithm on six categorical dataset.

Index	Dataset	GUDMM-S	WOCIL + OI	WOCIL	OCIL	KMCMD	*K*-prototype
CA	Soybean	**1 ± 0**	**1 ± 0**	0.8128 ± 0.1667	0.8353 ± 0.1644	0.8906 ± 0.1469	0.8357 ± 0.1515
	Vote	0.8775 ± 0.0011	**0.8799 ± 0**	0.8663 ± 0.0512	0.8668 ± 0.0672	0.8769 ± 0.0032	0.8656 ± 0.0098
	WBCD	0.9634 ± 0	0.8736 ± 0	0.8781 ± 0.0621	0.9128 ± 0.0012	**0.9642 ± 0.0007**	0.812 ± 0.1674
	Car	0.3727 ± 0.0201	0.3848 ± 0	**0.3986 ± 0.0576**	0.3683 ± 0.05	–	0.3735 ± 0.0391
	Lym	0.5984 ± 0.0062	0.5068 ± 0	0.4942 ± 0.0701	0.4993 ± 0.0332	**0.6232 ± 0.0541**	0.4423 ± 0.0529
	Mass	**0.8289 ± 0**	0.8265 ± 0	0.815 ± 0.0623	0.8008 ± 0.083	0.818 ± 0	0.8125 ± 0.0396
RI	Soybean	**1 ± 0**	**1 ± 0**	0.8908 ± 0.0939	0.9067 ± 0.2002	0.9413 ± 0.0798	0.9051 ± 0.0762
	Vote	0.7844 ± 0.0016	**0.7884 ± 0**	0.773 ± 0.0391	0.7782 ± 0.0565	0.7836 ± 0.0049	0.7669 ± 0.0143
	WBCD	0.9294 ± 0	0.7786 ± 0	0.7934 ± 0.0661	0.8406 ± 0.0021	**0.9308 ± 0.0013**	0.7503 ± 0.1656
	Car	**0.5176 ± 0.0175**	0.4913 ± 0	0.514 ± 0.0306	0.5071 ± 0.0263	–	0.4946 ± 0.0152
	Lym	**0.6507 ± 0.0037**	0.5582 ± 0	0.5771 ± 0.0366	0.6093 ± 0.0237	0.6394 ± 0.0252	0.5619 ± 0.0229
	Mass	**0.716 ± 0**	0.7128 ± 0	0.7058 ± 0.0421	0.6945 ± 0.0574	0.7018 ± 0	0.6981 ± 0.0281
NMI	Soybean	**1 ± 0**	**1 ± 0**	0.842 ± 0.1562	0.8666 ± 0.1125	0.9252 ± 0.1007	0.8479 ± 0.1191
	Vote	0.4925 ± 0.0023	**0.503 ± 0**	0.4743 ± 0.0678	0.4807 ± 0.0972	0.4912 ± 0.007	0.4675 ± 0.0313
	WBCD	0.76 ± 0	0.484 ± 0	0.5097 ± 0.109	0.5782 ± 0.0044	**0.7642 ± 0.004**	0.4471 ± 0.2643
	Car	0.1438 ± 0.0375	**0.144 ± 0**	0.1263 ± 0.0775	0.1095 ± 0.061	–	0.0513 ± 0.0206
	Lym	**0.3012 ± 0.0053**	0.126 ± 0	0.1812 ± 0.0624	0.228 ± 0.0418	0.28 ± 0.0394	0.15 ± 0.0441
	Mass	0.344 ± 0	**0.351 ± 0**	0.3338 ± 0.067	0.3112 ± 0.0916	0.328 ± 0	0.3183 ± 0.0453

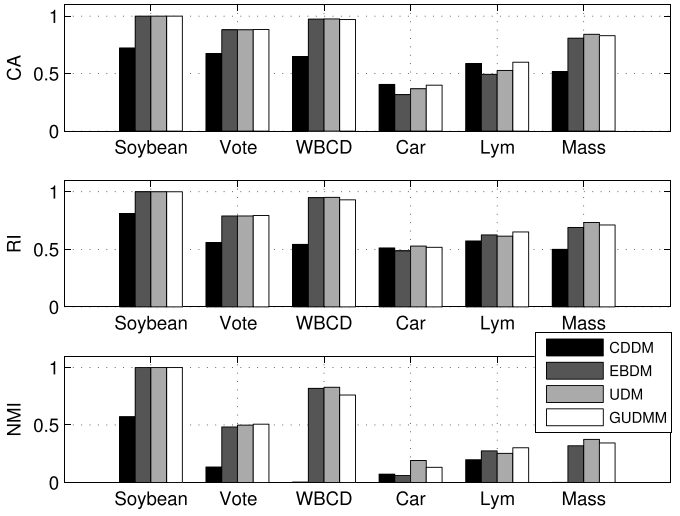

Fig. 3. Clustering performance of four distance metrics combined with spectral clustering on six categorical data sets.

cedure to define the category distance and distinguish between ordinal and nominal attributes (see Eq. (5)), the main difference between these two is how the relevance matrix R is calculated. UDM uses the concordance and discordance concepts to form R and distinguishes between nominal and ordinal attributes in the definitions of relevance coefficients, but GUDMM is based on MI, which imposes no discrimination between ordinal and nominal attributes in this step. Rank-based dependency coefficients only measure the monotonicity behavior of two attributes, but MI detects any relationship between attributes, whether linear or non-linear. Furthermore, MI is also insensitive to the size of the data sets and can provide a straightforward interpretation [35].

From the purely categorical datasets, Soybean and Vote are pure nominal data sets; WBCD and Car are pure ordinal data sets, and Lym and Mass are composed of ordinal and nominal attributes. Various combinations of ordinal and nominal attributes were used in these experiments to evaluate the GUDMM's performance compared to three state-of-the-art categorical distances. The results indicated that although the GUDMM is developed especially for mixed-type data, it also has an acceptable performance in various compositions of purely categorical datasets.

Table 7
Clustering performance of six clustering algorithm on six numerical data set.

Index	Dataset	GUDMM-S	WOCIL + OI	WOCIL	OCIL	KMCMD	*K*-prototype
CA	Iris	**0.9733 ± 0**	0.8667 ± 0	0.7798 ± 0.126	0.7692 ± 0.1208	0.8267 ± 0.12	0.8233 ± 0.1217
	Wine	**0.9719 ± 0**	0.9607 ± 0	0.9271 ± 0.0319	0.9329 ± 0.0946	0.9524 ± 0.0028	0.948 ± 0.0067
	Ionosphere	**0.7521 ± 0**	0.7123 ± 0	0.7093 ± 0.005	0.7094 ± 0.0014	0.7103 ± 0.0096	0.7065 ± 0.0156
	Sonar	0.5337 ± 0	0.5385 ± 0	**0.5641 ± 0.03**	0.5451 ± 0.0287	0.5469 ± 0.0323	0.5459 ± 0.0154
	Yeast	0.4103 ± 0.0005	**0.4474 ± 0**	0.4042 ± 0.0266	0.4074 ± 0.024	0.3672 ± 0.0205	0.3663 ± 0.0281
	Ecoli	**0.7914 ± 0.0366**	0.5625 ± 0	0.0.6073 ± 0.0446	0.6167 ± 0.0364	0.5563 ± 0.056	0.5523 ± 0.0656
RI	Iris	**0.9656 ± 0**	0.8568 ± 0	0.8142 ± 0.0566	0.8064 ± 0.052	0.8436 ± 0.0602	0.8415 ± 0.0601
	Wine	**0.9608 ± 0**	0.9457 ± 0	0.9062 ± 0.0375	0.9267 ± 0.0689	0.9379 ± 0.0038	0.9314 ± 0.0084
	Ionosphere	**0.6261 ± 0**	0.5889 ± 0	0.5865 ± 0.0041	0.5865 ± 0.0012	0.5874 ± 0.0068	0.5845 ± 0.0111
	Sonar	0.4999 ± 0	0.50063 ± 0	**0.5076 ± 0.0085**	0.5033 ± 0.0061	0.5041 ± 0.0081	0.5023 ± 0.0026
	Yeast	0.7467 ± 0.0002	0.746 ± 0	0.6685 ± 0.0836	0.7334 ± 0.0003	0.7447 ± 0.0044	**0.7489 ± 0.0033**
	Ecoli	**0.877 ± 0.0389**	0.7978 ± 0	0.8152 ± 0.0268	0.813 ± 0.0172	0.8027 ± 0.0198	0.8021 ± 0.0226
NMI	Iris	**0.901 ± 0**	0.697 ± 0	0.6614 ± 0.0462	0.6478 ± 0.0301	0.712 ± 0.06	0.7028 ± 0.0577
	Wine	**0.896 ± 0**	0.616 ± 0	0.7833 ± 0.0664	0.8341 ± 0.1073	0.8405 ± 0.013	0.8312 ± 0.0179
	Ionosphere	**0.232 ± 0**	0.135 ± 0	0.1304 ± 0.0075	0.13 ± 0.0024	0.1318 ± 0.0176	0.1245 ± 0.0286
	Sonar	0.004 ± 0	0.018 ± 0	**0.0214 ± 0.015**	0.0142 ± 0.0101	0.0132 ± 0.0118	0.0065 ± 0.0047
	Yeast	0.2807 ± 0.0006	**0.307 ± 0**	0.2402 ± 0.0379	0.2975 ± 0.0063	0.2586 ± 0.0113	0.27 ± 0.0124
	Ecoli	**0.689 ± 0.042**	0.542 ± 0	0.5708 ± 0.0286	0.6018 ± 0.0131	0.5991 ± 0.0328	0.5901 ± 0.0298

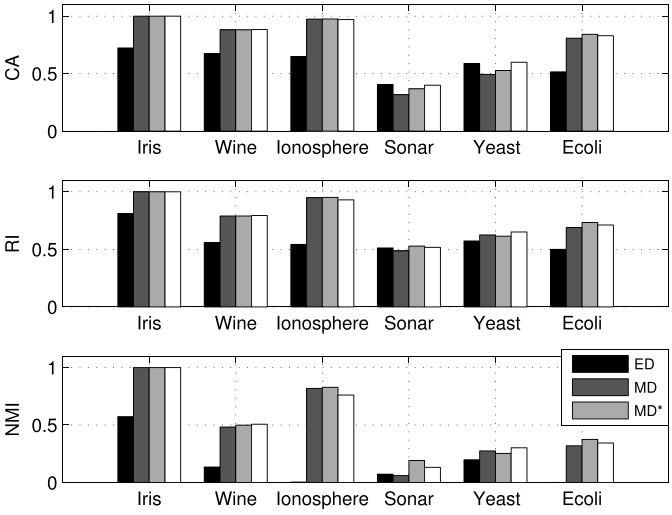

Fig. 4. Clustering performance of three distance metrics combined with spectral clustering on six categorical data sets. ED, MD, and MD* indicate Euclidean, Mahalanobis, and the proposed modified Mahalanobis distances, respectively.

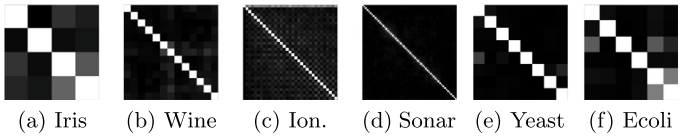

Fig. 5. Relevancy matrix of pure numerical data sets.

4.6. Experiment on pure numerical data sets

For the last experiments, we evaluated the performance of GUDMM-S on six pure numerical data sets, and the results are summarized in Table 7. It can be observed that, in most cases, GUDMM-S outperforms others. Especially for Ecoli, the performance of GUDMM-S is considerably higher than others.

To further investigate the efficiency of the proposed MD* for the numerical part of the data, we compared its performance with Euclidean and Mahalanobis distance in the framework of spectral clustering. Figure 4 indicates the results of evaluating these metrics on six pure numerical data sets. It can be observed that the performance of the MD* in Iris, Ionosphere, and Ecoli are better than the two others, and for Wine, Sonar, and Yeast, its performance is as good as clustering by Euclidean distance. Although MD imposes the intra and inter-attribute information to the distance calculation, MD has the lowest score in most data sets. But in MD*, where MD has been modified to control the imported inter-attribute information according to their association, the clustering results have improved. In fact, MD* considers more information in the distance calculation, while less unrelated attributes information is entered in the distance calculation. When the attributes relationships are significant, utilizing the relationship information of attributes could improve the clustering, while imposing the inter-attribute information with the negligible association on the distance is pointless and resulted in the same performance for Euclidean and MD*. Figure 5 indicates the relevancy matrices of the investigated data sets. Demonstrated sub-figures show the existence of associations between some of the attributes in Iris, Ionosphere, and Ecoli data sets, where exploiting inter-attribute information has improved their clustering performances.

To explore further the effectiveness of our different contributions on the performance of clustering, we designed five experiments and compared the results. The details of the experiments are explained in the Supplementary Materials 2.1. Furthermore, to examine the performance of GUDMM in combination with a partitioning algorithm, we also embedded GUDMM in the framework of *k*-prototype and reported the results in Supplementary Materials 2.2.

Supplementary Materials 3, includes the computational cost of GUDMM. The execution times of different algorithms on different data sets and the scalability test of the GUDMM-S and other methods, based on sub-sampling of the Adults data set from 2000 to 18000 samples, are summarized in Supplementary Materials 4 and 5, respectively.

5. Conclusion

This article has focused on the improvement of distance calculation for clustering of mixed-type data sets. Intra and inter-attribute information by utilizing the statistical information of each attribute and its relationship with other attributes has imposed on the calculation of distance. In addition to considering the distributions of other categorical attributes, we also provided a comprehensive method for importing numerical attribute information for the calculation of the distance between categories of a categorical attribute. The proposed modified version of the Maha-

lanobis distance could also impose the inter-attribute information of numerical attributes according to their strength of association. Considering the different nature of attributes, we proposed employing mutual information as a unified framework for relevancy computation. By imposing the relevancy coefficients, entering unnecessary information into the distance calculation was prevented. The proposed distance could be especially useful for cases with high dependency between attributes and be embedded in various distance-based clustering frameworks. By combining the proposed distance with spectral clustering, we examined the clustering of numerous benchmark data sets. The results indicated the superiority of the proposed method in most of the investigated mixed, pure categorical, and numerical data types. As a limitation, we utilized the spectral clustering algorithm in our framework using the developed distance matrix to avoid the constraints of prototype definition that exist in partitioning methods for clustering of mixed types of data. Since considering the similarity of values with respect to attribute dependency and aggregating such dependencies in the definition of a pair-wise distance matrix may incur a high computational cost for high-dimensional data, for future work, we will explore providing a global aggregation framework for GUDMM with unsupervised learning to improve both clustering performance and efficiency.

Declaration of Competing Interest

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

Data availability

I have shared the link to source code and the utilized dataset in the main manuscript.

Acknowledgments

The study was funded by Isfahan University of Medical Sciences (Number 3981000).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2023.109353.

References

[1] A. Ahmad, S.S. Khan, Survey of state-of-the-art mixed data clustering algorithms, IEEE Access 7 (2019) 31883–31902.
[2] Y. Zhang, Y.-M. Cheung, K.C. Tan, A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering, IEEE Trans. Neural Netw. Learn. Syst. 31 (1) (2019) 39–52.
[3] F. Yuan, Y. Yang, T. Yuan, A dissimilarity measure for mixed nominal and ordinal attribute data in *k*-modes algorithm, Appl. Intell. 50 (5) (2020) 1498–1509.
[4] A.H. Foss, M. Markatou, B. Ray, Distance metrics and clustering methods for mixed-type data, Int. Stat. Rev. 87 (1) (2019) 80–109.
[5] Z. He, X. Xu, S. Deng, Scalable algorithms for clustering large datasets with mixed type attributes, Int. J. Intell. Syst. 20 (10) (2005) 1077–1089.
[6] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, IEEE Trans. Knowl. Data Eng. 14 (4) (2002) 673–690.
[7] Y. Zhang, Y.-M. Cheung, A new distance metric exploiting heterogeneous inter-attribute relationship for ordinal-and-nominal-attribute data clustering, IEEE Trans. Cybern. 52 (2) (2020) 758–771.
[8] D. Lin, et al., An information-theoretic definition of similarity, in: ICML, vol. 98, 1998, pp. 296–304.
[9] A. Ahmad, L. Dey, A *k*-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2) (2007) 503–527.
[10] H. Ralambondrainy, A conceptual version of the *k*-means algorithm, Pattern Recognit. Lett. 16 (11) (1995) 1147–1157.
[11] I. Kosmidis, D. Karlis, Model-based clustering using copulas with applications, Stat. Comput. 26 (5) (2016) 1079–1099.
[12] Z. Huang, Clustering large data sets with mixed numeric and categorical values, in: Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD), Citeseer, 1997, pp. 21–34.
[13] J.Z. Huang, M.K. Ng, H. Rong, Z. Li, Automated variable weighting in *k*-means type clustering, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 657–668.
[14] J. Ji, T. Bai, C. Zhou, C. Ma, Z. Wang, An improved *k*-prototypes clustering algorithm for mixed numeric and categorical data, Neurocomputing 120 (2013) 590–596.
[15] T.-H.T. Nguyen, D.-T. Dinh, S. Sriboonchitta, V.-N. Huynh, A method for *k*-means-like clustering of categorical data, J. Ambient Intell. Humaniz Comput. 10 (2019) 1–11.
[16] D.-T. Dinh, T. Fujinami, V.-N. Huynh, Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient, in: International Symposium on Knowledge and Systems Sciences, Springer, 2019, pp. 1–17.
[17] D.-T. Dinh, V.-N. Huynh, *k*-PbC: an improved cluster center initialization for categorical data clustering, Appl. Intell. 50 (8) (2020) 2610–2632.
[18] C.-C. Hsu, C.-L. Chen, Y.-W. Su, Hierarchical clustering of mixed data based on distance hierarchy, Inf. Sci. 177 (20) (2007) 4474–4492.
[19] C.-C. Hsu, Y.-C. Chen, Mining of mixed data with application to catalog marketing, Expert Syst. Appl. 32 (1) (2007) 12–23.
[20] C. Wang, C.-H. Chi, W. Zhou, R. Wong, Coupled interdependent attribute analysis on mixed data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, 2015.
[21] D.S. Modha, W.S. Spangler, Feature weighting in *k*-means clustering, Mach. Learn. 52 (3) (2003) 217–237.
[22] A. Foss, M. Markatou, B. Ray, A. Heching, A semiparametric method for clustering mixed data, Mach. Learn. 105 (3) (2016) 419–458.
[23] A.H. Foss, M. Markatou, Kamila: clustering mixed-type data in R and hadoop, J. Stat. Softw. 83 (1) (2018) 1–44.
[24] X. Li, Z. Wu, Z. Zhao, F. Ding, D. He, A mixed data clustering algorithm with noise-filtered distribution centroid and iterative weight adjustment strategy, Inf. Sci. 577 (2021) 697–721.
[25] Y.-m. Cheung, H. Jia, Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number, Pattern Recognit. 46 (8) (2013) 2228–2238.
[26] H. Jia, Y.-M. Cheung, Subspace clustering of categorical and numerical data with an unknown number of clusters, IEEE Trans. Neural Netw. Learn. Syst. 29 (8) (2017) 3308–3325.
[27] S.Q. Le, T.B. Ho, An association-based dissimilarity measure for categorical data, Pattern Recognit. Lett. 26 (16) (2005) 2549–2557.
[28] D. Ienco, R.G. Pensa, R. Meo, From context to distance: learning dissimilarity for categorical data clustering, ACM Trans. Knowl. Discov. Data (TKDD) 6 (1) (2012) 1–25.
[29] Z. Khorshidpour, S. Hashemi, A. Hamzeh, CBDL: context-based distance learning for categorical attributes, Int. J. Intell. Syst. 26 (11) (2011) 1076–1100.
[30] H. Jia, Y.-m. Cheung, J. Liu, A new distance metric for unsupervised learning of categorical data, IEEE Trans. Neural Netw. Learn. Syst. 27 (5) (2015) 1065–1079.
[31] J. Brownlee, Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python, Machine Learning Mastery, 2018.
[32] A. Agresti, Analysis of Ordinal Categorical Data, vol. 656, John Wiley & Sons, 2010.
[33] T.O. Kvålseth, Measuring association between nominal categorical variables: an alternative to the Goodman–Kruskal lambda, J. Appl. Stat. 45 (6) (2018) 1118–1132.
[34] H. Khamis, Measures of association: how to choose? J. Diagn. Med. Sonogr. 24 (3) (2008) 155–162.
[35] B.C. Ross, Mutual information between discrete and continuous data sets, PLoS One 9 (2) (2014) e87357.
[36] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[37] C. Marsh, Introduction to continuous entropy, Department of Computer Science, Princeton University (2013).
[38] F. Nielsen, On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid, Entropy 22 (2) (2020) 221.
[39] B.W. Silverman, Density Estimation for Statistics and Data Analysis, vol. 26, CRC press, 1986.
[40] N. Eshima, Statistical Data Analysis and Entropy, Springer, 2020.
[41] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[42] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, Adv. Neural Inf. Process. Syst. 17 (2004) 1601–1608.
[43] D.-T. Dinh, V.-N. Huynh, S. Sriboonchitta, Clustering mixed numerical and categorical data with missing values, Inf. Sci. 571 (2021) 418–442.

Elahe Mousavi received the B.Sc. degree in electrical engineering in 2011, and the M.Sc. degree in biomedical engineering from the Tarbiat Modares University in 2013. Currently, she is a Ph.D. student of biomedical engineering at the Isfahan University of Medical Science. Her research interests include statistical modeling and clustering.

Mohammadreza Sehhati received his Ph.D. degree in Biomedical Engineering in 2015 from Isfahan University of Medical Sciences. He is currently an assistant professor at the Isfahan University of Medical Sciences. His research interests are focused on Biological Data Mining and constructing predictive models with medical application using machine learning techniques.