Unsupervised Coupled Metric Similarity for Non-IID Categorical Data

Songlei Jian, Longbing Cao, *Senior Member, IEEE,* Kai Lu, and Hang Gao

Abstract—Appropriate similarity measures always play a critical role in data analytics, learning and processing. Measuring the intrinsic similarity of categorical data for unsupervised learning has not been substantially addressed, and even less effort has been made for the similarity analysis of categorical data that is not independent and identically distributed (non-IID). In this work, a Coupled Metric Similarity (CMS) is defined for unsupervised learning which captures the value-to-attribute-to-object heterogeneous interactions. In particular, both intra- and inter-attribute similarities as well as attribute-to-object similarities are learned by considering the intrinsic heterogeneous coupling relationships in categorical data. We prove the validity of CMS w.r.t. metric properties and conditions, and show that it is suitable for both independent and coupled data. CMS is incorporated into spectral clustering and k-modes and compared with other state-of-the-art similarity measures that are not necessarily metrics. The experimental results and theoretical analysis show the effectiveness of CMS, which significantly outperforms other similarity measures on most data sets.

Index Terms—Unsupervised similarity measure, distance metric, non-IID data, categorical data, clustering.

1 Introduction

Measuring similarity or distance between objects is fundamental for effective data analytics, including tasks in data mining [1], machine learning [2], [3], image processing [4], computer vision [5], and information retrieval tasks [6], and in particular for complex data. A critical issue in data analytics and learning is determining whether learned outcomes are reliable and the underlying models genuinely capture intrinsic data characteristics.

Similarity measures for nominal (categorical) and numerical data are usually distinct. Most existing work on similarity learning focuses on numerical data, such as the commonly used Euclidean and Manhattan distances. For example, the similarity measures built on geometric analogies capture certain relations between numerical data values. As an example, the distance between age 12 and age 15 is smaller than that between age 12 and age 30, thus people aged 12 and 15 are more similar in terms of age.

The similarity (or distance) of categorical data is not as straightforward as it is for numerical data, however,since the different values of a categorical attribute may not be inherently ordered or comparable. Although there may be no inherent order in categorical data, other factors like matching statistics and frequency distribution exist and thus indicate similarity. Among existing work, the most straightforward and widely used distance metric is Hamming distance [7]. The corresponding similarity measure is *Matching* measure, which uses 0 and 1 to distinguish the similarity between distinct and identical categorical values. Other similarity measures take the frequency distribution of different attribute values into account, such as *Inverse Occurrence Frequency (IOF)* and *Occurrence Frequency (OF)* [8]. These similarity measures only capture the characteristics within an attribute but ignore the relationships between attributes.

- *Songlei Jian is with Science and Technology on Parallel and Distributed Processing Laboratory and College of Computer, National University of Defense Technology, China. Songlei Jian is also visiting the Advanced Analytics Institute, University of Technology Sydney, Australia. E-mail: songlei.jian@uts.edu.au*
- *Longbing Cao is with Advanced Analytics Institute, University of Technology Sydney, Australia.*
- *Kai Lu, Hang Gao are with the Laboratory of Science and Technology on Parallel and Distributed Processing and College of Computer, National University of Defense Technology, China.*

Manuscript received April 19, 2005; revised August 26, 2015.

1.1 Motivating toy example

We illustrate the problem with the existing work and the inherent challenges in analyzing the similarity of categorical data by taking the staff data of a lab in Table 1 as an example. The staff data consists of four categorical attributes: *Sex*, *Education*, *Occupation* and *Marriage*. From the *matching* perspective, the similarity between *Staff 1* and *Staff 2* is the same as that between *Staff 1* and *Staff 3*, because they are both 0.5. However, from both education and occupation perspectives, professors and assistant professors should be more similar than professors and students in the lab.

A further analysis of relationships in categorical data shows the need to consider the frequency distribution of an attribute value and the co-occurrences between attributes. From the OF perspective, two values of an attribute are similar if they present analogous frequency distributions [8]. For example, the similarity between Professor and Assistance Professor is greater than that between Professor and Student, because the occurrence frequency in the staff data of Professor and Assistant Professor is the same. Although the attribute values can disclose more information than simple matching, value frequency-based similarity is not sufficient. For example, the similarity between the education levels Doctor and Master is the same as that between Doctor and Bachelor. This is because the frequency distribution only captures the count statistics of attribute values, ignoring the coupling relationships within and between attributes. The co-occurrence of attribute values induced on other attributes

TABLE 1
An Example: The Staff Data

Staff	Sex	Education	Occupation	Marriage
Staff 1	F	Doctor	Professor	Married
Staff 2	M	Doctor	Assistant Professor	Married
Staff 3	F	Master	Student	Married
Staff 4	M	Master	Student	Single
Staff 5	F	Bachelor	Student	Single
Staff 6	M	Bachelor	Student	Single

is more comparable [9], [10], and complements the accuracy of frequency-based value similarity. By incorporating co-occurrence based attribute similarity, the pair Doctor and Master is more similar than Doctor and Bachelor, because the former pair co-occurs with the same occupation and marriage, while the latter does not.

1.2 Major issues and contributions

The above example shows that it is often much more complicated to define the similarity of categorical data, especially when data is embedded with complex relationships [11], [12]. Big data applications become increasingly important and popular, heterogeneous and hierarchical coupling relationships [11] are embedded into categorical attributes, values and objects, which make it even more difficult to measure similarity or dissimilarity. For instance, the market dynamics in a stock market may have many related reasons and factors, such as psychological, economic, social, organizational, political, cultural or even military. Data presenting explicit and/or implicit couplings and heterogeneity is not independent and identically distributed (i.e., non-IID) [13], [14], [15]. Such data does not follow the classic IID assumption, which has formed the basis of forming most existing similarity measures.

Non-IID data learning has attracted increasing attention and is usually built on non-IID samples in the relevant communities [14], [15]. Several model-based approaches have been proposed to address non-IID samples, such as analyzing non-IID textual data by higher order Naive Bayes [16], classification with non-IID samples [17], developing chromatic PAC-Bayes bounds for non-IID data [18], and learning from dependent observations [19].

Real-life data non-IIDness [13] is embodied in terms of various coupling relationships and heterogeneities between values, between attributes and between objects, forming the value-to-object hierarchical non-IID data [13], [20]. A fundamental learning task is to enhance the robustness and generalization [21] of learning metrics and models in such non-IID data. As an example, the *matching* and *OF*-based measures cannot fully capture the genuine similarity of non-IID categorical data as they only capture particular aspects.

In recent years, learning for hierarchical non-IID data has been recognized as a foundational issue in complex data analytics, with such typical work as Coupled Object Similarity (COS) [10], [22], which involves the couplings within and between attributes before object similarity is defined. Other related work to address non-IID characteristics [20] that incorporates couplings into various types of learning tasks includes coupled clustering [10], coupled KNN for classification [23], term coupling-based document analysis [24], coupled keyword queries [25], coupled matrix factorization by item and user couplings into recommender systems [26], understanding relationships between patterns for pattern relation analysis [27], [28], and analyzing image couplings [29], [30].

Many of the existing similarity measures in the above work are not metric-based and do not provide a solid theoretical foundation that can satisfy the metric system and support such properties as positivity, reflexivity, commutativity and triangle inequality, as will be discussed in Section 3. A distance or similarity metric reflects the results in the metric space which is a topological space. Accordingly, all definitions and theorems about general topological spaces can be applied to metric spaces to support metric properties and operations. For example, the classical k-means clustering algorithm is based on common metric space Euclidean distance. With a similarity metric, we can define the distance between data objects and the distance between data objects and data sets (such as clusters) in metric space in a similar way as in Euclidean space. Therefore, it is critical to develop appropriate metric-based similarity measures.

In this paper, we build on the idea of incorporating heterogeneous and hierarchical value-to-object coupling relationships [20] into learning systems and we propose a *coupled metric similarity* (CMS) metric for non-IID categorical data. CMS integrates the frequency-based intra-attribute similarity with the co-occurrence based inter-attribute similarity before object similarity is measured. The intra-attribute similarity captures the frequency distribution and the coupling relationships between values in an attribute. The inter-attribute similarity aggregates the attribute dependency between values of different attributes by considering intersection of their co-occurrence conditional probability. The coupled metric similarity integrates intra-attribute similarity with inter-attribute similarity by balancing their contributions. Importantly, we prove that CMS is a valid similarity metric in terms of satisfying properties of positivity, reflexivity, commutativity, and triangle inequality.

Our main contributions are detailed below:

- We propose a coupled metric similarity measure for unsupervised learning of non-IID categorical data which captures both the intra-attribute similarity of attribute values and the inter-attribute similarity coupling relationships between attributes. The intra- and inter-attribute similarities are further synergized to form a coupled metric similarity(CMS) which measures the similarity between non-IID objects.
- Five theorems are proposed and proved to ensure the validity of CMS as a metric.
- By introducing a control parameter, a flexible combination function is defined so that CMS can be used for both IID and non-IID data. This shows the flexibility of CMS.
- The proposed CMS metric is compared with the state-of-the-art similarity measures when they are incorporated into both distance-based clustering and similarity-based clustering algorithms on nineteen UCI benchmark data sets. Evaluation and empiri-

cal analysis of the resultant statistically significant outcomes are provided to understand why CMS works well from the perspective of both similarity constituents and data characteristic.

The remainder of the paper is organized as follows. In Section 2, we discuss the related work. The problem formulation and statement are specified in Section 3. Section 4 introduces the CMS measure. The proof of CMS validity and theoretical analyses of properties are given in Section 5. We demonstrate the efficiency and effectiveness of our similarity measure with experiments and analysis in Section 6, and discuss the underlying mechanisms of CMS in Section 7. Lastly, conclusions and future work are discussed in Section 8.

2 Related work

The similarity learning of categorical data has attracted increasing attention in recent years [8], [31], [32]. Compared to numeric data similarity learning, learning categorical data similarity is more complicated, and limited research outcomes have been reported. The matching-based measures are typical for categorical data. A matching-based measure simply assigns the similarity as 1 if the values of an attribute for two objects are identical; otherwise it assigns 0. However, such simple matching-based measures often result in misleading learning outcomes as discussed in the Introduction, and they disregard the hidden similarity between categorical values [33]. Further, the inverse occurrence frequency (IOF) and occurrence frequency (OF) based measures take occurrence frequency distribution into account. IOF is related to the concept of inverse document frequency, which was designed for text mining [34] and assigns lower similarity to mismatches on more frequent values, and vice versa. An OF measure gives the opposite weight of the IOF measure for mismatches.

Intensive studies have been conducted on learning the similarity between two categorical values in supervised learning [35], [36], [37]. A classic similarity measure in supervised learning is the Value Distance Matrix (VDM) and the Modified Value Distance Matrix (MVDM) [38] based on class labels. Both methods measure the distance between two numeric attribute values in a multi-dimensional attribute space for supervised learning and modify the distance with a weighting scheme. Wilson and Martineza designed a Heterogeneous Value Difference Metric (HVDM) [39] to cater for categorical attributes.

An increasing number of researchers have also paid attention to similarity analysis for unsupervised learning [40], [41]. A key point is that the attribute value similarity is also dependent on other attributes [8], [11]. Typical efforts in this area applied the Pearson and Jaccard coefficients between values [22], [31]. The Pearson correlation coefficient only reflects the strength of linear dependence [42] within numeric data. The Jaccard similarity coefficient statistically compares the similarity and diversity of sample sets and is widely used in data mining tasks [43].

A variety of techniques for categorical data similarity learning have been explored. Believing that the attribute and object similarities are interdependent, Das and Mannila [44] presented the Iterated Contextual Distances (ICD) algorithm. ICD considers and iterates attribute similarity, sub-relation similarity, and row similarity; however, it faces a number of issues including the selection of starting points, database scan times, iterations, and convergence. Ahmad and Dey [9] proposed a distance-based measure in terms of value co-occurrences. Their work considers the overall distribution of two attribute values in a data set along with their co-occurrences with the values of other attributes. While their similarity measure seems to achieve high accuracy, it only considers value co-occurrence, and does not cater for value-to-object hierarchical similarity; in addition, computation is costly. No theoretical foundation and analysis were provided to prove whether metric properties are satisfied.

Building on the concepts of intra- and inter-behavior coupling relationships [20] and coupled behavior similarity for coupled behavior analysis (CBA) [11], we proposed coupled object similarity (COS) [10], [22] for learning categorical data similarity. COS is based on the belief that object similarity, attribute similarity and value similarity form an interactive and inter-dependent hierarchical system which cannot be ignored in the similarity definition for complex data. Accordingly, COS captures the Intra-coupled Attribute Value Similarity (IaAVS), the Inter-coupled Attribute Value Similarity (IeAVS) and their integration to learn object similarity. Experiments show that COS achieves significant improvement over existing similarity measures in clustering categorical data. The CBA and COS methods have been applied in classification [23], recommender systems [26], text mining [24], keyword query [25], and video processing [30]. However, COS is not a metric-based similarity, and no theoretical foundation and analysis have been provided to verify its metric properties and detemine why it works better. In addition, the execution of the algorithm is very costly.

To address the above relevant issues, this work takes a step forward by proposing the concept and similarity learning system Coupled Metric Similarity (CMS). CMS learns object similarity by proposing hierarchical similarity measures that capture both horizontal and vertical coupling relationships between the values of an attribute, between attributes, and between objects. CMS ensures that these value-to-attribute-to-object similarity measures satisfy metric properties with sound theoretical design and proof and this foundation makes CMS applicable to distance-based algorithms.

3 Problem formulation

In this section, we first discuss the necessary conditions for a valid distance-based function and a metric similarity measure. Further preliminaries are provided to establish the foundation for proposing new similarity metrics. These will form the theoretical foundation for the concept of CMS, which will be discussed in Section 5.

3.1 Metric properties

A metric space is an ordered pair (M, δ) where M is a set and δ is a metric on M, i.e., a function: $d : M \times M \to \mathbb{R}$ so that, for any $u_x, u_y, u_z \in M$, the following properties hold [45]:

1) non-negativity: $\delta(u_x, u_y) \geq 0$
2) reflexivity: $\delta(u_x, u_y) = 0 \Leftrightarrow u_x = u_y$
3) commutativity: $\delta(u_x, u_y) = \delta(u_y, u_x)$
4) triangle inequality: $\delta(u_x, u_z) \leq \delta(u_x, u_y) + \delta(u_y, u_z)$

The function $\delta()$ is called a *distance function* (here simply called *distance*).

A valid similarity metric needs to satisfy the above properties. Given the above function $\delta()$, the following mapping function [8] is applied to convert it to a distance-based similarity (or dissimilarity-based similarity):

$$s(u_x, u_y) = \frac{1}{1 + \delta(u_x, u_y)}, \tag{1}$$

where $s(u_x, u_y)$ is the similarity between two data points u_x and u_y, and $\delta(u_x, u_y)$ is the distance u_x and u_y.

With the above mapping function, we can deduce the following conditions that a metric-based similarity measure should hold:

1) positivity: $0 < s(u_x, u_y) \leq 1$
2) reflexivity: $s(u_x, u_y) = 1 \Leftrightarrow$ then u_x is exactly the same as u_y
3) commutativity: $s(u_x, u_y) = s(u_y, u_x)$
4) triangle inequality: $\frac{1}{s(u_x,u_y)} + \frac{1}{s(u_y,u_z)} \geq 1 + \frac{1}{s(u_x,u_z)}$

3.2 Problem statement

Assume a data set DB consists of a number of data objects U that are described by a set of attributes A. DB can be organized as an information table $S =< U, A, V >$, where $U = \{u_1, ..., u_n\}$ is composed of a non-empty finite set of data objects; $A = \{a_1, ..., a_m\}$ is a finite set of attributes; $V = \cup_{j=1}^{m} V_j$ consists of sets of values of all attributes, in which V_j is the set of values of attribute a_j (where $1 \leq j \leq m$).

For better readability and easy illustrations of CMS-based similarity calculations in the following sections, the information table shown in Table 2 is used as an example. Symbols A_1 and A_2 represent two distinct values of attribute a_1; B_1 and B_2 represent two distinct values of attribute a_2; and C_1 and C_2 represent two distinct values of attribute a_3. Table 2 thus consists of six objects $\{u_1, ..., u_6\}$ and four attributes $\{a_1, a_2, a_3, a_4\}$, and the value set of attribute a_2 is $V_2 = \{B_1, B_2\}$.

We assume that the similarity between two objects u_x and u_y ($u_x, u_y \in U$) is the summation of the similarities between attribute values v_j^x, v_j^y (v_j^x, $v_j^y \in V_j$) for any attribute a_j, where $j \in [1, m]$), and v_j^x and v_j^y indicate the respective attribute values of objects u_x and u_y on the attribute a_j. For instance, $v_1^2 = A_2$ and $v_2^1 = B_1$. We propose some basic concepts below which will form the foundation to facilitating the introduction of the CMS measure, a new object similarity metric for categorical data in Section 4.

Definition 1 (Conditional Probability of Attribute Values). Given the value v_k of attribute a_k ($a_k \in A$), and the value v_j^x ($v_j^x \in V_j$) of object u_x on attribute a_j, then the conditional probability of v_k with respect to v_j^x is $p(v_k|v_j^x)$, defined as:

$$p(v_k|v_j^x) = \frac{|I(v_j^x, v_k)|}{|I(v_j^x)|}, \tag{2}$$

TABLE 2
Toy Example: A Car Data Set

U \ A	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_2
u_3	A_1	B_2	C_1
u_4	A_1	B_2	C_2
u_5	A_2	B_2	C_1
u_6	A_1	B_1	C_2

TABLE 3
List of Main Notations

Notation	Explanation
$\{u_1, ..., u_n\}$	The set of n objects
$\{a_1, ..., a_m\}$	The set of m attributes
v_j^x, v_j^y	Specific values of attribute a_j for objects u_x, u_y
v_k	Any value of attribute a_k
$I(v_j^x)$	The set of the objects whose value of attribute a_j is v_j^x
$I(v_j^x, v_k)$	The set of the objects whose value of attribute a_j is v_j^x and value of attribute a_k is v_k
V_j^I	The set of values of attribute a_j for all objects in the object set I
$\|I\|$	The size of set I, i.e., the number of objects in I
$\delta(u_i, u_j)$	The similarity between objects u_i and u_j
$p(v_k\|v_j^x)$	The conditional probability of v_k w.r.t. v_j^x

where $I(v_j^x, v_k)$ denotes the set of the objects u_x whose attribute value of a_j is v_j^x and attribute value of a_k is v_k, $I(v_j^x)$ denotes the set of the objects whose attribute value of a_j is v_j^x, $|\dot{|}$ denotes the number of elements in the contained set.

For example, the values for two objects u_1 and u_2 on attribute a_1 are $v_1^1 = A_1$ and $v_1^2 = A_2$, hence $I(v_1^1) = I(A_1) = \{u_1, u_3, u_6\}$, for the value B_1 of attribute a_2, $I(A_1, B_1) = \{u_1, u_6\}$, $p(B_1|A_1) = |I(A_1, B_1)|/|I(A_1)| = 2/3$.

The above notations and definitions form the foundation of the CMS which will be presented in the following section.

The main notations in this paper are listed in Table 3.

4 Coupled Metric Similarity Measures

In this section, we discuss the learning framework, working mechanism, and key components that form the Coupled Metric Similarity (CMS) measures.

4.1 The learning framework

Coupled metric similarity aims to capture the object similarity by considering and integrating both intra- and inter-attribute similarities as well as object similarities. Fig. 1 illustrates its learning framework, working mechanism, and corresponding key similarity measures. Coupled metric similarity is built on an information table for categorical data as discussed in Session 3.2. It first captures the value couplings in terms of intra-attribute similarities ($s_{Ia}^j(v_j^x, v_j^y)$),

followed by the feature couplings in terms of inter-attribute similarities ($s_{Ie}^{k|j}(v_j^x, v_j^y)$). The intra- and inter-attribute similarities are then integrated into coupled metric attribute value similarity $s^j(v_j^x, v_j^y)$, which are further aggregated to measure object similarities, i.e., the coupled metric similarity ($s(u_x, u_y)$).

The intra-attribute similarity captures the value co-occurrence distribution within an attribute. For example, in Table 2, the intra-attribute similarity between attribute values A_1 and A_2 is related to the frequency of the values A_1 and A_2, which are both 3. The inter-attribute similarity between values A_1 and A_2 of attribute a_1 depends on the attribute values of other two attributes (a_2 and a_3). Coupled metric similarity combines the intra-attribute similarity and the inter-attribute similarity in a flexible way to form the object similarity. We ensure the metric validity of the proposed intra-attribute similarity, the inter-attribute similarity, and the CMS.

4.2 Intra-attribute similarity

According to [31], the discrepancy in attribute value occurrence times reflects the value similarity in terms of frequency distribution. The similarity between two objects is related to their commonality. Accordingly, the intra-attribute similarity considers the relationship between the frequency of the attribute values of an attribute, defined as follows.

Definition 2 (Intra-attribute Similarity). The intra-attribute similarity between two attribute values v_j^x, v_j^y of objects u_x and u_y on attribute a_j is $s_{Ia}^j(v_j^x, v_j^y)$, defined as follows:

$$s_{Ia}^j(v_j^x, v_j^y) = \begin{cases} 1 & \text{if } v_j^x = v_j^y, \\ \dfrac{\log p \cdot \log q}{\log(p \cdot q) + \log p \cdot \log q} & \text{otherwise} \end{cases} \quad (3)$$

where log represents the natural logarithm, p denotes $|I(v_j^x)| + 1$, and q denotes $|I(v_j^y)| + 1$. $I(v_j^x)$ is the set of objects whose values of attribute a_j are v_j^x. Similarly, $I(v_j^y)$ is the set of objects whose values of attribute a_j is v_j^y.

According to metric similarity conditions defined in Section 3, if the attribute values are identical, the similarity between them should be 1. When the attribute values are not identical, their occurrence frequency indicates their similarity. Equation (3) is designed to satisfy the following three principles.

- The maximum similarity between two attribute values is reached when the values are identical.
- The greater similarity is assigned to the attribute value pair which shares approximately equal frequencies.
- The higher the frequency of two values, the closer two values are.

Equation (3) reflects that different occurrence frequencies indicate distinct levels of attribute value significance. When the size of the data increases sharply, the log function can control the growth of similarity. To prevent the denominator from being zero, we add 1 to each term. Since $1 \leq |I(v_j^x)|, |I(v_j^y)| \leq m$, then $s_{Ia}^j(v_j^x, v_j^y) \in (0, 1]$. If $v_j^x \neq v_j^y$, $s_{Ia}^j(v_j^x, v_j^y)$ achieves the maximum value when $|I(v_j^x)| = |I(v_j^y)| = m/2$. For example, in Table 2, $|I(A_1)| = 3$ and $|I(A_2)| = 3$, $s_{Ia}^1(A_1, A_2) = 0.41$.

4.3 Inter-attribute similarity

The above intra-attribute similarity reflects the coupling relationships between the attribute values of one attribute a_j, however, it does not involve the couplings between other attributes a_k ($k \neq j$) and attribute a_j. Accordingly, we discuss the inter-attribute similarity, which involves the couplings between attributes and is much more complicated than intra-attribute couplings.

We note that the Modified Value Distance Matrix (MVDM) [38] measures the dissimilarity between categorical values w.r.t. class labels. It shows that attribute values are similar if they occur with similar relative frequency for all classifiers. Based on MVDM, Wang et. al. [10], [22] replaced the class labels with other attributes to enable unsupervised learning and proposed the *Inter-coupled Relative Similarity based on Power Set* (IRSP). They also proposed the *Inter-coupled Relative Similarity based on Join Set* (IRSJ), and the *Inter-coupled Relative Similarity based on Intersection Set* (IRSI), and proved that these measures are equivalent to each other in achieving the same accuracy in calculating value similarity [10], [22]. They prove that IRSI is the most efficient of the above measures; however, IRSI cannot retain the conditions of a metric similarity which are discussed in Section 3. Below, we propose a new inter-attribute similarity measure to satisfy metric properties.

Before calculating the inter-attribute similarity, we define the intersection set of co-occurrence conditional probability W_k.

Definition 3 (Intersection Set of Co-occurrence Conditional Probability of Attribute Values). The intersection set of co-occurrence conditional probability of values v_j^x, v_j^y of attribute a_j with co-occurrence values of attribute a_k ($j \neq k$) is:

$$W_k = V_k^{I(v_j^x)} \cap V_k^{I(v_j^y)}, \quad (4)$$

$V_k^{I(v_j^x)}$ is the set of values of attribute a_k for all objects in $I(v_j^x)$. W_k consists of those attribute values of attribute a_k which co-occur with both v_j^x and v_j^y.

The Jaccard similarity coefficient is widely used in clustering and classification. The Jaccard similarity coefficient is defined as:

$$J(\mathbf{f}, \mathbf{g}) = \frac{\sum_i \min(f_i, g_j)}{\sum_i \max(f_i, g_j)}, \quad (5)$$

where $\mathbf{f} = (f_1, f_2, ..., f_n)$ and $\mathbf{g} = (g_1, g_2, ..., g_n)$ are two vectors with all real numbers.

The Jaccard distance is a distance metric [46] as follows:

$$\delta_J(u_x, u_y) = 1 - J(x, y). \quad (6)$$

According to the Jaccard distance and Equation (1) discussed in the previous section, we define the inter-attribute similarity with the Jaccard similarity based on IRSI [10] and W_k as follows.

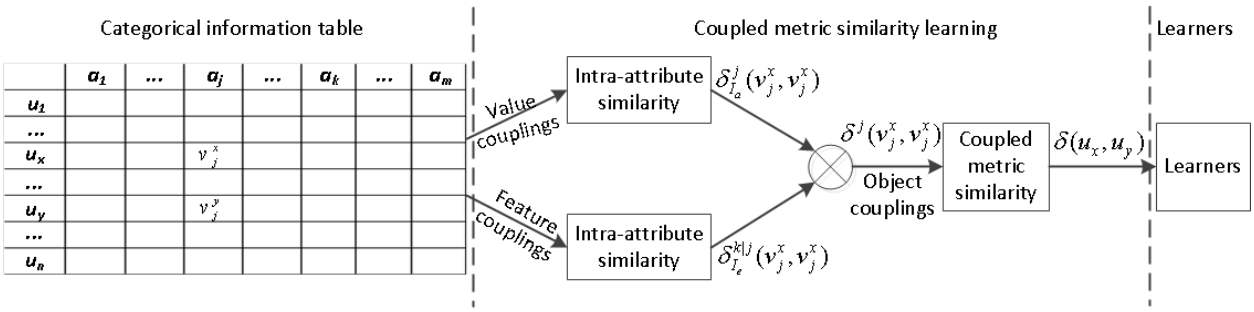

Fig. 1. The Framework of Coupling Metric Similarity Learning

Definition 4 (Inter-attribute Similarity of Attribute Values w.r.t. Another Attribute). The *inter-attribute similarity* between two attribute values v_j^x and v_j^y of attribute a_j with another attribute a_k is:

$$s_{Ie}^{k|j}(v_j^x, v_j^y) = \begin{cases} 1, \\ \text{if } v_j^x = v_j^y \\ \dfrac{\sum_{i=1}^{|W_k|} \max(p_x^i, p_y^i)}{2 \cdot \sum_{i=1}^{|W_k|} \max(p_x^i, p_y^i) - \sum_{i=1}^{|W_k|} \min(p_x^i, p_y^i) \frac{|W_k|}{|I|}}, \\ \text{otherwise} \end{cases} \tag{7}$$

where $p_x^i = p(w_k^i|v_j^x)$, $p_y^i = p(w_k^i|v_j^y)$, they are conditional probabilities of w_k^i with respect to v_j^x and v_j^y, p_x^i and p_y^i are calculated according to Equation (2). w_k^i is the i_{th} element in W_k which is calculated according to Equation (4). In particular, if W_k is empty, $s_{Ie}^j(v_j^x, v_j^y) = 0$.

In Table 2 for example, according to Equation (4), the similarity $s_{Ie}^{2|1}(A_1, A_2)$ depends on the values of attribute a_2. $I(A_1) = \{u_1, u_3, u_6\}$ and $I(A_2) = \{u_2, u_4, u_5, \}$, hence $V_2^{I(A_1)} = \{B_1, B_2\}$, $V_2^{I(A_2)} = \{B_1, B_2\}$ and $W_2 = \{B_1, B_2\}$. According to Equation (7), we calculate $\max(p(B_1|A_1), p(B_1|A_2))$, $\max(p(B_2|A_1), p(B_2|A_2))$, $\min(p(B_1|A_1), p(B_1|A_2))$, and $\min(p(B_2|A_1), p(B_2|A_2))$, and obtain $s_{Ie}^{2|1}(A_1, A_2) = 0.57$.

Following the above discussion, we further define the similarity between the value pair (v_j^x, v_j^y) of attribute a_j on top of the Jaccard similarity of other attributes a_k ($j \neq k$).

Definition 5 (Inter-attribute Similarity). The inter-attribute similarity between two attribute values v_j^x and v_j^y of attribute a_j is:

$$s_{Ie}^j(v_j^x, v_j^y) = \sum_{k=1, k\neq j}^{m} \gamma_{k|j} s_{Ie}^{k|j}(v_j^x, v_j^y), \tag{8}$$

where $\gamma_{k|j}$ represents the weight of each attribute a_k ($j \neq k$) to attribute a_j, $\sum_{k=1,k\neq j}^{m} \gamma_{k|j} = 1$, $\gamma_{k|j} \in [0, 1]$, and $s_{Ie}^{k|j}(v_j^x, v_j^y)$ is one of the inter-attribute similarity candidates with attribute a_k. $\gamma_{k|j}$ reflects the relation between attributes a_j and a_k.

Consequently, we have $s_{Ie}^{k|j}(v_j^x, v_j^y) \in [0, 1]$. Particularly, if W_k is not empty, $s_{Ie}^{k|j}(v_j^x, v_j^y) \in [0.5, 1]$. Since $s_{Ie}^{k|j}(v_j^x, v_j^y)$ is in $[0, 1]$, then $s_{Ie}^j(v_j^x, v_j^y) \in [0, 1]$.

In Table 2, for example $s_{Ie}^1(A_1, A_2) = 0.5 \cdot s_{Ie}^{1|2}(A_1, A_2) + 0.5 \cdot s_{Ie}^{1|3}(A_1, A_2) = 0.57$ if $\gamma_2 = \gamma_3 = 0.5$ by taking the equal weight. Furthermore, the coupled metric similarity (see Equation (10) in the following section) is obtained as $s^1(v_j^x, v_j^y) = 0.4904$ if $\alpha = 1$.

4.4 Coupled metric similarity

With the above defined intra-attribute similarity measure $s_{Ia}^j(v_j^x, v_j^y)$ and inter-attribute similarity measure $s_{Ie}^j(v_j^x, v_j^y)$, we now define the coupled metric similarity measure for attribute a_j.

An ideal similarity measure should be suitable for both IID and non-IID [13] data, hence we introduce a parameter α to satisfy this requirement. The parameter α can be adjusted to change the proportion of $s_{Ia}^j(v_j^x, v_j^y)$ against $s_{Ie}^j(v_j^x, v_j^y)$. We define the coupled similarity measure of v_j^x and v_j^y as follows.

Definition 6 (Coupled Metric Attribute Value Similarity). The *coupled metric attribute value similarity* (CMAVS) between attribute values v_j^x and v_j^y of attribute a_j is:

$$s^j(v_j^x, v_j^y) = \alpha s_{Ie}^j + (1 - \alpha) s_{Ia}^j, \tag{9}$$

where s_{Ia}^j and s_{le}^j are respectively the intra-attribute similarity and inter-attribute similarity of attribute values v_j^x and v_j^y. α is the control factor and $\alpha \in [0, 1)$ guarantees the positivity of CMAVS.

From the above definition, we can determine that α reflects the ratio of inter-attribute similarity in the overall similarity. A larger α indicates that more important couplings between attributes are captured in the similarity space and that the attribute a_j is more coupled with other attributes.

If all attributes are independent, $\alpha = 0$, correspondingly $s^j(v_j^x, v_j^y) = s_{Ia}^j$, indicating that only the couplings within an attribute are considered in measuring similarity. When α increases, $s^j(v_j^x, v_j^y)$ becomes closer to s_{Ie}^j. Therefore, by adjusting the parameter α, we can control $s^j(v_j^x, v_j^y)$ flexibly. Later in Section 6, we will show the effect of tuning parameter α on learning performance and demonstrate that it may be possible to find an empirically optimal α value for a given data set, while different data sets may share distinct α values.

We calculate the similarity between two objects u_x and u_y on top of CMAVS defined in Equation (9).

Definition 7 (Coupled Metric Similarity). The *coupled metric similarity* (CMS) between two objects u_x and u_y is $s(u_x, u_y)$:

$$s(u_x, u_y) = \sum_{j=1}^{m} \beta_j s^j(v_j^x, v_j^y), \tag{10}$$

where β_j represents the weight of the coupled metric attribute value similarity of an attribute a_j, $\sum_{j=1}^{m} \beta_j = 1$, $\beta_j \in [0, 1]$.

5 THEORETICAL ANALYSIS

This section proves that CMS is a valid similarity metric and analyses the theoretical properties and computational complexity of CMS.

5.1 CMS metric validity

Before we prove the validity of CMS, we first prove several theorems to lay the foundation.

Theorem 1. $s^j(v_j^x, v_j^y) = 1$, if and only if $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$ for every attribute a_j and when $\alpha \neq 0$.

Theorem 2. $s^j(v_j^x, v_j^y) = 1$, if and only if $s_{Ia}^j(v_j^x, v_j^y) = 1$ for every attribute a_j and when $\alpha = 0$.

Theorem 3. The coupled metric attribute value similarity s^j satisfies the triangle inequality if both intra-attribute similarity s_{Ia}^j and inter-attribute similarity s_{Ie}^j satisfy the triangle inequality for every attribute a_j.

Theorem 4. The intra-attribute similarity s_{Ia}^j satisfies the triangle inequality for any attribute a_j.

Theorem 5. The inter-attribute similarity s_{Ie}^j satisfies the triangle inequality for any attribute a_j.

These theorems are proved in the Appendix. Consequently, we prove that the validity of the proposed CMS, namely $s(u_x, u_y)$, satisfies the following metric properties.

1) **Positivity**: $0 < s(u_x, u_y) \leq 1$.
$s^j(v_j^x, v_j^y)$ consists of $s_{Ia}^j(v_j^x, v_j^y)$ and $s_{Ie}^j(v_j^x, v_j^y)$. According to Equation (3), $s_{Ia}^j(v_j^x, v_j^y)$ is in (0,1]. $s_{Ie}^j(v_j^x, v_j^y)$ is based on the linear product of $s_{Ie}^{k|j}(v_j^x, v_j^y)$, and $s_{Ie}^{k|j}(v_j^x, v_j^y)$ is the Jaccard similarity of vectors, accordingly $s_{Ie}^j(v_j^x, v_j^y) \in [0, 1]$. According to Equation (9), $s(u^x, u^y)$ is the weighted sum of all similarity measures for each attribute's $s_{Ia}^j(v_j^x, v_j^y)$ and $s_{Ie}^j(v_j^x, v_j^y)$. The similarity measure $s(u^x, u^y)$ therefore satisfies the positivity constraint.
2) **Reflexivity**: $s(u_x, u_y) = 1 \Leftrightarrow u_x = u_y$
We prove the necessity first.
If $u_x = u_y$ then it means $v_j^x = v_j^y$ for all attributes $\{a_j\}$. According to Equations (3) and (8), if $v_j^x = v_j^y$, then $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$, so $s^j(v_j^x, v_j^y) = 1$ and $s^j(u_x, u_y) = 1$. Hence, the similarity measure satisfies the necessity condition.
We further prove the sufficiency.
If $s(u_x, u_y) = 1$, we can conclude that $s^j(v_j^x, v_j^y) = 1$ for every attribute a_j according to Equation (10), because $s^j(v_j^x, v_j^y)$ is also in (0,1]. According to Theorem 1, $s^j(v_j^x, v_j^y) = 1$, if and only if $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$ and when $\alpha \neq 0$. According to Theorem 2, when $\alpha = 0$, $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$. From Equation (3), we can conclude that $s_{Ia}^j(v_j^x, v_j^y) = 1$ if and only if $v_j^x = v_j^y$. For all attributes a_j, $v_j^x = v_j^y$ means $u_x = u_y$. The similarity measure satisfies the sufficiency condition, therefore the proposed similarity measure $s(u_x, u_y) = 1$, if and only if $u_x = u_y$.
3) **Commutativity**: $s(u_x, u_y) = s(u_y, u_x)$.
All operations on our proposed similarity measure are addition, multiplication, maximum selection, and minimum selection. These operations are commutative. Hence the inequality holds implicitly.
4) **Triangle inequality**: $\frac{1}{s(u_x,u_y)} + \frac{1}{s(u_y,u_z)} \geq 1 + \frac{1}{s(u_x,u_z)}$
The resultant similarity $s(u_x, u_y)$ is the mean of all similarities $s^j(v_j^x, v_j^x)$ computed for every attribute. If $s^j(v_j^x, v_j^x)$ holds the triangle inequality, so does $s(u_x, u_y)$ (Its proof is similar to the proof of Theorem 3). If we can prove that each component of $s^j(v_j^x, v_j^x)$ (including $s_{Ia}^j(v_j^x, v_j^x)$ and $s_{Ie}^j(v_j^x, v_j^x)$) holds the triangle inequality, then $s^j(v_j^x, v_j^x)$ satisfies the triangle inequality according to Theorem 3.
The triangle inequality of $s_{Ia}^j(v_j^x, v_j^x)$ and $s_{Ie}^j(v_j^x, v_j^x)$ are easy to prove. According to Theorem 5, $s_{Ia}^j(v_j^x, v_j^x)$ satisfies the triangle inequality. $s_{Ie}^j(v_j^x, v_j^x)$ can be converted to the Jaccard distance according to Equation (1) since the Jaccard distance is a metric distance which holds the triangle inequality [47]. Hence, the triangle inequality is satisfied as well.

5.2 CMS theoretical property and computational complexity

The CMS measure $s(u_x, u_y)$ is an increasing function of s_{Ia}^j and s_{Ie}^j according to Equation (9). According to Definition 2, $s_{Ia}^j(v_j^x, v_j^x)$ reflects the intra-attribute similarity between two attribute values v_j^x and v_j^y. The higher the value $s_{Ia}^j(v_j^x, v_j^x)$ is, the closer the two attribute values are. Equation (7) shows that the inter-attribute similarity $s_{Ie}^{k|j}(v_j^x, v_j^y)$ is increasing with the size of co-occurrence set $|W_k|$. Hence, the larger s_{Ie}^j is, the more similar the two attribute values v_j^x and v_j^y are. In conclusion, we can obtain the increasing property of CMS, which means that the larger $s(u_x, u_y)$ is, the more similar two objects u_x and u_y are.

The CMS between two objects captures all intra-attribute and inter-attribute similarities of each attribute value pair in the corresponding information table. Accordingly, the computational complexity linearly depends on the number of attribute values. The most time-consuming element is the calculation of inter-attribute similarity which is quantified by the calculation of $s_{Ie}^{k|j}(v_j^x, v_j^y)$. Hash table is used as the data structure to store $p(v_k|v_j^x)$ of each pair. Suppose the maximal number of distinct values for each attribute is R and m is the number of attributes, the time complexity of calculating the hash table of $p(v_k|v_j^x)$ for all attributes

is $mR(R-1)$. The maximal value of $|W_k|$ is the number of objects n. According to Equation (7), the computational complexity of inter-attribute similarity is $nmR(R-1)$. Consequently, the upper bound of the time complexity of CMS for two objects is $O(nm^2R^2)$, and the upper bound of the total CMS complexity for n objects is $O(n^2m^2R^2)$.

6 Experiments and Evaluation

In this section, we evaluate the CMS performance in clustering categorical data. CMS is evaluated by comparison with other similarity measures for categorical data and clustering performance by incorporating CMS into popular clustering methods on 19 data sets.

6.1 Baseline clustering methods and measures

To evaluate the performance of CMS, the following five state-of-the-art similarity/distance measures are compared with CMS: ALGO_DISTANCE (ALGO for short) [9], Coupled Object Similarity (COS for short) [10], [22], Distance Metric of Hong (DM for short) [48], Matching similarity [7] (or Hamming Distance), and Occurrence Frequency (OF for short) [8].

We apply CMS to a typical similarity-based categorical clustering algorithm spectral clustering [49] and a distance-based algorithm k-modes [50] on categorical data. Similarly, we also incorporate ALGO, COS, DM, OF and Matching similarity measures into spectral clustering and the k-modes clustering algorithm. We compare their clustering performance to evaluate which similarity measure achieves better outcomes.

According to the distance-similarity or dissimilarity-similarity mapping function, i.e., Equation (1) in Section 3, we can derive the metric distance or dissimilarity measure from the coupled metric similarity as follows:

$$\delta(u_x, u_y) = \frac{1}{s(u_x, u_y)} - 1, \tag{11}$$

where $\delta(u_x, u_y)$ denotes the distance or dissimilarity between objects u_x and u_y, and $s(u_x, u_y)$ is the coupled metric similarity between u_x and u_y defined in Equation (10).

In Equation (10), we assign the weight vector $(\beta_k)_{1\times m}$ with values $\beta_k = 1/m$, and assign $\gamma_{k|j} = 1/(m-1)$ for every attribute in Equation (7) for simplicity. This simple setting is not optimal, but it can test whether the proposed CMS design is effective even in a non-optimal situation.

6.2 Data sets

Ninteen UCI data sets are used for the experiments[1]. The detailed characteristics of these 19 different data sets are described in terms of four data factors in Table 4. They are O - the number of objects, A - the number of attributes, V - the number of distinct values for all attributes, and C - the number of classes (we include the class information for evaluation only). *Abbr.* refers to the short form of a data set name. All numerical attributes in the data sets are removed to test the similarity for categorical data only.

1. https://archive.ics.uci.edu/ml/datasets.html

TABLE 4
The Data Characteristics of 19 UCI Data Sets

Data set	O	A	V	C	*Abbr.*
Soybeansmall	47	35	97	4	So
Zoo	101	16	36	7	Zo
DNAPromoter	106	57	228	2	Dp
Hayesroth	132	4	15	3	Ha
Lymphography	148	18	59	4	Ly
Hepatitis	155	13	36	2	He
Housevotes	232	16	32	2	Ho
Spect	267	22	44	2	Sp
Mofn3710	300	10	20	2	Mo
Soybeanlarge	307	35	132	19	Sol
Primarytumor	339	17	42	21	Pr
Dermatology	366	33	129	6	De
ThreeOf9	512	9	18	2	Tr
Wisconsin	683	9	89	2	Wi
Crx	690	9	45	2	Cr
Breastcancer	699	9	90	2	Br
Mammographic	830	4	20	2	Ma
Flare	1066	11	41	6	Fl
Titanic	2201	3	6	4	Ti

The UCI data is used here because it is public, relatively simple and easy to understand compared to more complex real-life data. The assumption made is as follows. If a non-IID-oriented measure can beat its baselines on UCI data, such a result should hold consistently on real-life data, which is usually non-IID, since UCI data is not likely to contain many sophisticated non-IID characteristics. Another reason for using the UCI data is the challenge of finding non-IID real-life data sets, since there is currently no tool that can verify whether or to what extent a data set is non-IID.

6.3 Evaluation methods

The performance of CMS and other similarity measures-based clustering is evaluated in terms of the internal criteria and external criteria. The internal criteria focus on the compactness and separability of clusters which are dependent on distance or similarity, which reflect the quality of resultant clusters to a certain extent. Although there are various internal clustering validation measures, most are related to the centers chosen by clusters [51]. Because it is difficult to decide the cluster centers for categorical data, we choose a center-independent internal criteria: Dunn's index [52], defined below.

$$D = \min_i \left\{ \min_j \left(\frac{\min_{u_x \in C_i, u_y \in C_j} \delta(u_x, u_y)}{\max_k \left\{ \max_{u_x, u_y \in C_k} \delta(u_x, u_y) \right\}} \right) \right\}, \tag{12}$$

where C_i is the ith cluster and $\delta(u_x, u_y)$ denotes the distance between u_x and u_y.

For CMS, we use the internal criteria to choose the α value in Equation (9). By greedy search of the Dunn's index results of CMS-enabled clustering, we can obtain the optimal α value. This is aligned with the purpose of CMS, that is, to maximize the similarity of similar objects and minimize the similarity of dissimilar objects. It is also worth noting that the α value chosen by the Dunn's index is not the

optimal value since it only reflects one aspect of similarity measurement.

For external criteria, we choose some commonly used criteria to compare the clustering results for different similarity measures. The external criteria estimate the difference between the cluster label of each object assigned by each clustering algorithm and the ground truth indicated by the data labels given in the source data. The criteria include *normalized mutual information* (*NMI*) and *F1-score*. The larger these criteria are, the better performance the clustering achieves; the corresponding similarity measure is accordingly more effective. The definitions of these three measures are given below.

- *Normalized Mutual Information*

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{c} n_{i,j}\log\left(\frac{n\cdot n_{i,j}}{n_i\cdot n_j}\right))}{\sqrt{\left(\sum_{i=1}^{k} n_i \log\frac{n_i}{n}\right)\left(\sum_{j=1}^{c} n_j \log\frac{n_j}{n}\right)}}, \quad (13)$$

  where c stands for the true number of classes, k is the number of clusters obtained by the algorithm, $n_{i,j}$ denotes the number of agreements between cluster i and class j, and n is the number of objects in the whole data set.

- *F1-score*

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}, \quad (14)$$

  where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

6.4 CMS α value w.r.t. capturing coupling relationships in data

The α value in Equation (9) reflects the extent of the coupling relationships embedded in a data set. We conduct experiments with different α values not only to choose a better α value but also to measure the coupling difference across data sets. The larger the value of α, the greater the proportion of inter-attribute similarity, hence the coupling relationships in a data set are stronger.

Tables 5 and 6 report the Dunn's index values of clustering results for different α values taken in spectral clustering and k-modes clustering. The results show that more than half of the data sets have coupling relationships between attributes.

TABLE 5
The Dunn's Index Value of CMS-enabled Spectral Clustering with Different α Values

Data sets	α value					
	0	0.2	0.4	0.6	0.8	1
So	0.461	**0.469**	0.432	0.387	0.447	0.442
Zo	0.106	**0.113**	0.111	0.092	0.088	0.088
Dp	**0.537**	0.535	0.535	0.535	0.535	0.536
Ha	0.212	0.230	0.237	0.241	0.243	**0.244**
Ly	0.149	0.188	0.191	0.191	0.191	**0.192**
He	0.077	**0.082**	0.081	0.081	0.081	0.081
Ho	**0.167**	0.161	0.160	0.151	0.151	0.151
Sp	**0.049**	0.045	0.044	0.044	0.045	0.045
Mo	**0.099**	0.098	0.098	0.098	0.098	0.098
Sol	0.047	0.049	0.045	0.046	0.049	**0.063**
Pr	0.072	0.068	0.071	**0.073**	0.072	0.071
De	**0.096**	0.089	0.088	0.088	0.093	0.095
Tr	0.121	0.121	**0.122**	0.121	0.121	0.121
Wi	0.136	**0.203**	0.108	0.109	0.111	0.111
Cr	0.091	**0.105**	0.104	0.103	0.103	0.103
Br	0.136	**0.203**	0.108	0.110	0.110	0.111
Ma	0.217	0.242	0.245	0.246	0.246	**0.247**
Fl	**0.103**	0.096	0.097	0.097	0.097	0.097
Ti	0.322	**0.337**	0.312	0.260	0.250	0.279

TABLE 6
The Dunn's Index Values of CMS-enabled K-modes Clustering with Different α Values

Data sets	α value					
	0	0.2	0.4	0.6	0.8	1
So	0.264	**1.000**	0.830	0.765	0.823	0.670
Zo	0.121	**1.000**	0.960	0.978	0.989	0.906
Dp	**0.541**	0.528	0.538	0.526	0.532	0.540
Ha	**0.213**	0.187	0.176	0.171	0.167	0.167
Ly	0.110	0.090	0.099	**0.102**	0.098	0.099
He	0.081	0.083	0.082	0.082	0.082	**0.082**
Ho	0.044	0.044	0.042	0.043	**0.044**	0.044
Sp	0.062	0.063	0.064	0.061	0.062	**0.065**
Mo	**0.110**	0.110	0.109	0.109	0.109	0.109
Sol	0.044	0.044	0.042	0.043	**0.044**	0.044
Pr	**0.090**	0.085	0.082	0.081	0.079	0.080
De	**0.083**	0.081	0.082	0.079	0.080	0.078
Tr	0.118	0.118	0.118	0.118	0.118	**0.118**
Wi	0.178	0.261	0.255	0.260	0.265	**0.269**
Cr	0.101	0.110	0.110	**0.111**	0.113	0.109
Br	0.173	0.237	0.264	0.258	0.261	**0.287**
Ma	0.224	0.240	**0.241**	0.239	0.238	0.237
Fl	**0.108**	0.108	0.107	0.108	0.106	0.108
Ti	**0.893**	0.620	0.591	0.538	0.491	0.491

6.5 Comparison of CMS and other similarity measures-enabled spectral clustering

Tables 7 and 8 report the results of spectral clustering by taking the similarity measures CMS, ALGO, COS, DM, OF and matching similarity in terms of performance measures *NMI* and *F1-score*. The overall performance is given in the bottom row w.r.t. the mean value. For each data set, the average performance is obtained by 100 tests of spectral clustering with distinct start points. In the clustering evaluation, the highest Dunn's index value, as shown in Table 5 is assigned to the α parameter in Equation (9). This reflects an acceptable balance between intra-attribute couplings and inter-attribute couplings in the respective data set.

The clustering results show that CMS-enabled spectral clustering outperforms the spectral clustering methods empowered by ALGO, COS, DM, OF and matching measures on seven out of 19 data sets in terms of both *NMI* in Table 7 and *F1-score* in Table 8. By contrast, COS, which also captures both intra-attribute and inter-attribute couplings, is superior only four times in terms of *NMI* and three times in terms of *F1-score*. CMS performs better than two best and most recent similarity baselines for categorical data: ALGO and COS by 3.4% and 6.8% respectively, and the worst performing measure DM by 10.4% in terms of the mean NMI value. In addition, CMS outperforms ALGO and COS by 2.4% and 4.7% respectively in terms of *F1-score*.

It is also interesting to note every other similarity measures performs better than at least one other in one or two data sets. This reflects the significant challenge of designing appropriate and generalized similarity measures for categorical data due to the difficulty of effectively capturing the intrinsic data characteristics in categorical data.

TABLE 7
NMI of CMS vs. ALGO, COS, DM, OF and Matching-enabled Spectral Clustering

Data set	CMS(α)	ALGO	COS	DM	OF	Matching
So	**0.958**(0.2)	0.953	0.946	0.957	0.941	0.952
Zo	0.675(0.2)	0.731	0.705	**0.748**	0.672	0.69
Dp	**0.281**(0)	0.209	0.154	0.227	0.103	0.25
Ha	**0.02**(1)	0.001	0.011	0.001	0.001	0.001
Ly	**0.209**(1)	0.159	0.164	0.138	0.119	0.152
He	0.182(0.2)	0.183	0.179	0.185	0.129	**0.209**
Ho	0.493(0)	0.522	**0.524**	0.522	0.493	0.493
Sp	**0.107**(0)	0.106	0.102	0.105	0.094	0.106
Mo	0.036(0)	**0.054**	0.054	0.017	0.017	0.017
Sol	0.734(1)	0.736	**0.768**	0.666	0.632	0.673
Pr	0.345(0.6)	0.348	0.337	**0.366**	0.351	0.344
De	0.798(0)	**0.814**	0.807	0.792	0.491	0.748
Tr	**0.035**(0.4)	0.002	0.002	0.034	0.03	0.026
Wi	0.775(0.2)	0.829	**0.831**	0.685	0.811	0.76
Cr	0.239(0.2)	0.035	0.043	0.024	**0.294**	0.201
Br	0.796(0.2)	0.818	**0.82**	0.664	0.801	0.749
Ma	**0.369**(1)	0.329	0.361	0.332	0.326	0.331
Fl	0.302(0)	0.318	0.269	0.192	**0.374**	0.28
Ti	0.129(0.2)	0.101	0.1	0.122	0.128	0.105
Mean	**0.394**	0.381	0.369	0.357	0.358	0.373

TABLE 8
F1-score of CMS vs. ALGO, COS, DM, OF and Matching-enabled Spectral Clustering

Data sets	CMS(α)	ALGO	COS	DM	OF	Matching
So	**0.926**(0.2)	0.911	0.893	0.915	0.898	0.925
Zo	0.507(0.2)	0.547	0.538	**0.588**	0.494	0.518
Dp	**0.783**(0)	0.753	0.726	0.753	0.675	0.771
Ha	**0.387**(1)	0.335	0.338	0.329	0.336	0.343
Ly	0.359(1)	0.366	**0.395**	0.286	0.319	0.34
He	0.661(0.2)	0.662	0.463	0.695	0.633	**0.704**
Ho	0.884(0)	0.888	**0.893**	0.888	0.884	0.884
Sp	**0.589**(0)	0.572	0.582	0.563	0.565	0.559
Mo	0.506(0)	**0.567**	0.567	0.509	0.509	0.511
Sol	0.563(0)	0.553	**0.609**	0.476	0.48	0.504
Pr	0.206(1)	0.209	0.196	**0.23**	0.213	0.205
De	**0.745**(0)	0.71	0.73	0.735	0.615	0.66
Tr	0.584(0.4)	0.522	0.519	**0.59**	0.582	0.578
Wi	0.935(0.2)	0.971	**0.973**	0.942	0.968	0.96
Cr	**0.794**(0.2)	0.551	0.493	0.527	0.791	0.753
Br	0.961(0.2)	0.969	**0.97**	0.937	0.966	0.957
Ma	**0.826**(1)	0.818	0.822	0.82	0.817	0.823
Fl	0.367(0)	0.392	0.352	0.32	**0.444**	0.359
Ti	0.366(0.2)	**0.375**	0.358	0.35	0.306	0.323
Mean	**0.629**	0.614	0.601	0.603	0.605	0.615

6.6 Comparison of CMS and other similarity measures-enabled k-modes clustering

Tables 9 and 10 report the *NMI* and *F1-score* results of k-modes clustering enabled by the distance measures CMS, ALGO, COS, DM, OF and Hamming distance. These similarity measures are transformed into distance measures according to Equation (11). The overall performance is given in the bottom row by the mean value. For each data set, the average performance is obtained via 100 tests of k-modes clustering with distinct start points. For CMS, the α parameter in Equation (9) is given the value with the highest Dunn's index as shown in Table 6.

In Table 9, the *NMI* results of CMS-enabled k-modes are superior to other measure-enabled k-modes on nine data sets, compared to five by ALGO, three by COS, and one by DM and OF respectively, but none by Hamming distance. Overall, CMS outperforms DM by 65%, OF by 45%, and COS by 4.7%, while it achieves the same mean performance as ALGO. The *F1-score* results of CMS-enabled k-modes outperform five other measures in seven out of 19 data sets as shown in Table 10, compared to seven by ALGO, two by COS, but none by OF and Hamming distance. With regard to the mean *F1-score*, CMS is superior to DM by 25%, and generally outperforms other measures.

TABLE 9
NMI of CMS vs. ALGO, COS, DM, OF and Matching-enabled K-modes Clustering

Data sets	CMS(α)	ALGO	COS	DM	OF	Hamming
So	0.865(0.2)	**0.887**	0.86	0.699	0.745	0.828
Zo	0.766(0.2)	0.788	**0.803**	0.757	0.736	0.759
Dp	**0.089(0)**	0.08	0.071	0.018	0.048	0.06
Ha	0.021(0)	0.014	**0.056**	0.012	0.037	0.015
Ly	0.159(0.6)	0.159	**0.164**	0.138	0.119	0.152
He	0.135(1)	**0.15**	0.102	0.128	0.099	0.131
Ho	**0.539(0.8)**	0.511	0.511	0.531	0.44	0.461
Sp	**0.097(1)**	0.092	0.087	0.092	0.089	0.088
Mo	**0.032(0)**	0.017	0.025	0.023	0.029	0.02
Sol	**0.688(0.8)**	0.683	0.674	0.286	0.602	0.631
Pr	0.359(0)	**0.374**	0.359	0.348	0.335	0.362
De	**0.748(0)**	0.734	0.743	0.152	0.272	0.578
Tr	**0.031(1)**	0.004	0.004	0.026	0.02	0.025
Wi	0.616(1)	**0.668**	0.606	0.028	0.102	0.482
Cr	0.182(0.6)	0.176	0.054	**0.229**	0.132	0.168
Br	**0.681(0.4)**	0.673	0.609	0.108	0.115	0.437
Ma	0.319(0.4)	**0.326**	0.277	0.271	0.269	0.305
Fl	0.373(0)	0.364	0.375	0.182	**0.409**	0.31
Ti	**0.118(0)**	0.112	0.117	0.113	0.106	0.109
Mean	**0.359**	0.358	0.342	0.218	0.248	0.312

6.7 Experimental summary

In summary, the above experiments show that CMS in most cases outperforms other state-of-the-art similarity and distance measures when they are incorporated into spectral clustering and k-modes clustering for clustering categorical data. The experimental results also show that, for most data sets, similarity/distance measures that involve coupling relationships, i.e., CMS, ALGO, DM and COS, almost always obtain better performance. This shows the importance of capturing the various coupling relationships embedded in complex categorical data [20]. By contrast, CMS significantly outperforms the five state-of-the-art categorical similarity measures in the overall results, indicating that CMS is better at capturing coupling relationships than other similarity measures.

The results also show that none of the similarity and distance measures can always win on all 19 data sets for unsupervised learning. This indicates the complexity and

TABLE 10
F1-score of CMS vs. ALGO, COS, DM, OF and Hamming-enabled K-modes Clustering

Data sets	CMS(α)	ALGO	COS	DM	OF	Hamming
So	0.807(0.2)	**0.854**	0.78	0.604	0.745	0.799
Zo	0.522(0.2)	0.554	**0.571**	0.52	0.514	0.534
Dp	**0.62**(0)	0.611	0.564	0.358	0.578	0.606
Ha	0.387(0)	0.374	**0.421**	0.375	0.398	0.359
Ly	**0.399**(0.6)	0.366	0.395	0.286	0.319	0.34
He	**0.645**(1)	0.641	0.62	0.516	0.615	0.636
Ho	0.865(0.8)	0.884	0.884	**0.896**	0.865	0.87
Sp	0.444(1)	0.445	0.403	**0.468**	0.463	0.42
Mo	**0.514**(0)	0.505	0.507	0.508	0.514	0.493
Sol	0.461(0.8)	0.489	**0.504**	0.081	0.449	0.452
Pr	0.227(0)	0.226	0.215	0.213	0.205	**0.228**
De	0.590(0)	**0.6**	0.577	0.196	0.317	0.517
Tr	**0.567**(1)	0.353	0.353	0.534	0.552	0.564
Wi	0.882(1)	**0.907**	0.887	0.434	0.536	0.815
Cr	0.653(0.6)	0.668	0.492	**0.751**	0.64	0.678
Br	0.834(0.4)	**0.928**	0.901	0.517	0.542	0.783
Ma	0.798(0.4)	**0.817**	0.769	0.759	0.759	0.793
Fl	**0.455**(0)	0.377	0.39	0.302	0.445	0.375
Ti	**0.326**(0)	0.314	0.32	0.313	0.324	0.318
Mean	**0.578**	0.574	0.556	0.454	0.515	0.557

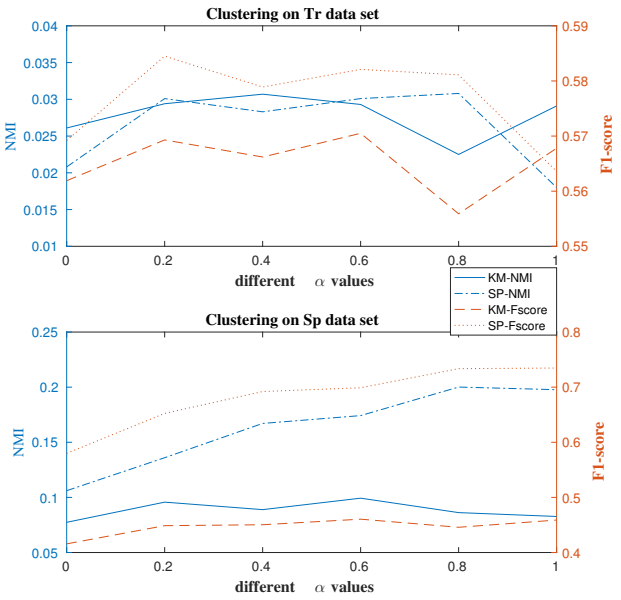

Fig. 2. Clustering results on Tr and Sp data sets w.r.t. different α values

significance of deeply understanding the intrinsic data characteristics of complex categorical data (which cannot be simply measured by matching/Hamming measure or the frequency of co-occurrence). In the following section, we discuss the underlying driving forces for CMS in capturing various hierarchical coupling relationships in categorical data.

7 Discussion on using CMS to capture coupling relationships

In this section, we empirically analyze why CMS achieves good performance. We explore the intrinsic working mechanisms of CMS, namely, by observing the impact of involving three levels of coupling relationships on clustering performance: intra-attribute similarity for capturing value couplings, inter-attribute similarity for capturing attribute couplings, and coupled similarity between objects which integrates both intra-attribute similarity and inter-attribute similarity. We particularly discuss how the intra-attribute similarity and inter-attribute similarity capture the intrinsic couplings within data.

7.1 Balance of intra-attribute and inter-attribute similarities

Coupled metric similarity integrates both intra-attribute similarity and inter-attribute similarity, as shown in Equation (9), in terms of their different contributions adjustable by parameter α. $\alpha = 0$ means we only consider the couplings within an attribute (i.e., intra-attribute similarity). $\alpha = 1$ indicates that we only consider the couplings between attributes (i.e., inter-attribute similarity). The effect of tuning parameter α on Dunn's index is shown in Table 5 and Table 6, although the outcomes only capture partial aspects of the data characteristics and may be sub-optimal. The α value corresponding to the highest Dunn's index value on a data set strikes a balance between the contributions made by intra-attribute similarity and inter-attribute similarity, which explains why the CMS incorporated by the corresponding α value obtains desirable clustering performance.

In our experiments, we also find that CMS with only intra-attribute similarity can obtain better clustering performance than some methods that consider both intra-attribute and inter-attribute similarities on data sets. As shown in Tables 7 and 8 for spectral clustering and Tables 9 and 10 for k-modes clustering, CMS performs consistently well on data sets Sp and Tr. Accordingly, we show the clustering performance of CMS-enabled spectral clustering and k-modes clustering for all the α values shown in Tables 5 and 6 to demonstrate the challenge of balancing intra-attribute and inter-attribute similarities in unsupervised learning.

Fig. 2 reflects the clustering performance on Tr and Sp data sets. It shows that different α values can lead to different clustering performance, hence, it is necessary to choose the optimal α value. The optimal α value on the Tr data set is 0.2 for both k-modes clustering and spectral clustering. In terms of spectral clustering, the optimal α value of data set Sp is 1. Meanwhile k-modes clustering is not sensitive to the change of α value for Sp when α value is larger than 0. The optimal α value of one data set is decided by not only data characteristics but also the clustering algorithm. We will study this issue in our future work.

7.2 Scrutinizing data characteristics

A effective similarity metric needs to capture the intrinsic data characteristics, which may be quantified in terms of data factors and indicators [53]. This section explores CMS in terms of capturing categorical data factors and hierarchical similarities. We illustrate this exploration by scrutinizing the characteristics of the Soybean-small data in Table 4.

The Soybean-small data set contains 47 objects, 35 attributes and 100 distinct attribute values. It is clearly a small

data set but it is relatively interesting due to its 'large' numbers of attributes and values compared to very 'small' object number in the UCI data. To illustrate its value and attribute coupling relationships, we select three attributes: *plant-stand*, *precip* and *temp*, and use *a1-a2, b1-b3, and c1-c3* to label the distinct values of these respective attributes.

At the attribute value level, we calculate the occurrence frequency of each value and the co-occurrence frequency of each value pair. Table 11 shows these statistics. The bold-faced values in the diagonal correspond to the occurrence frequencies of attribute values, and the other non-empty cells capture the co-occurrences of value pairs from different attributes.

Table 12 shows the intra-attribute similarity and inter-attribute similarity of attribute value pairs, as labelled as s^j_{Ia} for the intra-attribute similarity, s^j_{Ie} for the inter-attribute similarity, and s^j for the coupled object similarity, as defined in Equations (3), (8) and (9). The statistical information shown in Table 11 and the diverse similarities collected in Table 12 enable us disclose the intrinsic data characteristics in Soybean-small, and the power of CMS in terms of capturing such characteristics.

The intra-attribute similarity s^j_{Ia} of pair b1-b2, for example is 0.5880, which is larger than that of b1-b3. It consists of the relationship between frequencies of b1, b2 and b3 in Table 11, hence the intra-attribute similarity captures the frequency distributions and reflects the couplings between values within an attribute. The inter-attribute similarities s^j_{Ie} of most pairs (except pairs b1-b3 and c1-c2) in Table 12 are larger than their intra-attribute similarities. For pair c1-c2, its inter-attribute similarity is 0.2656, which is smaller than that of other pairs. In Table 11, the co-occurrence frequency of pairs c1-a2, c1-b2 and c1-b3 are 15, 15 and 2 respectively, while the co-occurrence frequencies of pairs c2-a2, c2-b2 and c2-b3 are all 0. This indicates that the co-occurring values of c1 are quite different from those of c2. The inter-attribute similarity s^j_{Ie} consists of the co-occurrence frequencies of attribute value pairs; it captures the latent couplings between different attributes. In this way, the similarity between attributes is transformed to the attribute-value similarity and then reflected in the similarity on the object level. The results in Table 12 also show the sensitivity of the integration of the intra- and inter-attribute similarities in calculating the value-to-attribute-to-object similarity.

Lastly, as shown in Section 7.1, CMS combines the intra-attribute similarity and inter-attribute similarity. The parameter α adjusts the contributions of intra-attribute similarities versus inter-attribute similarities. Since different data sets probably own diverse combinations of intra- and inter-attribute couplings, and different optimal α values accordingly exist. An optimal α value reflects the most appropriate distribution of intra- and inter-attribute similarities in a data set.

TABLE 11
The Frequencies and Co-occurrences of Attribute Values

Attribute Values	a1	a2	b1	b2	b3	c1	c2	c3
a1	**22**		10	13	0	2	6	14
a2		**25**	0	21	4	15	0	10
b1	10	0	**10**			0	6	4
b2	13	21		**33**		15	0	18
b3	0	4			**4**	2	0	2
c1	2	15	0	15	2	**17**		
c2	6	0	6	0	0		**6**	
c3	14	10	4	18	2			**24**

TABLE 12
The Value-to-Attribute-to-Object Similarities

	a1-a2	b1-b2	b1-b3	b2-b3	c1-c2	c1-c3	c2-c3
s^j_{Ia}	0.615	0.588	**0.491**	0.525	**0.538**	0.604	0.549
s^j_{Ie}	**0.682**	**0.703**	0.417	**0.827**	0.266	**0.734**	**0.626**
s^j	0.648	0.644	0.454	0.676	0.402	0.669	0.586

8 Conclusions and Future Work

Learning for non-IID data significantly challenges existing analytical and learning theories and metrics in effectively capturing the intrinsic heterogeneity and coupling relationships in non-IID data. Categorical data embedded with value-to-attribute-to-object hierarchical coupling relationships is particularly complex. Learning such data requires the appropriate representation and similarity metrics for capturing such hierarchical coupling relationships from attribute values to attributes and objects.

In this paper, we have proposed and evaluated a novel *coupled metric similarity* measure CMS for clustering hierarchical couplings in categorical data. Taking a data-driven approach that integrates intrinsic coupling relationships from low-level attributes and their values to objects, CMS grasps both attribute value frequency distribution and attribute dependency similarity in measuring attribute value similarity, attribute similarity and then object similarity.

Compared with the state-of-the-art similarity measures, including ALGO-distance, coupled object similarity, distance matrix, occurrence frequency-based measure, and matching-based measure, the incorporation of CMS and the above measures into two representative clustering methods, spectral clustering and k-modes, shows the great advantage of CMS against the baseline similarity measures in representing the above hierarchical couplings. CMS also incorporates a tuning mechanism for both distance-based and similarity-based analysis of either IID or non-IID data.

In addition to the experimental evaluation, we have investigated the driving factors of CMS-enabled performance improvement, as proved by the metric properties and explained by discussion on the underlying working mechanisms of CMS from statistical and data characteristic perspectives. None of the existing categorical similarity and dissimilarity measures provide such a theoretical foundation as CMS.

We are working on designing effective data structures and strategies for efficient enhancement and scalable clustering of large-scale non-IID categorical data using CMS. Another aspect we are working on is how to handle heterogeneous data with value-to-object couplings. The extension of CMS to numeric data clustering and classification is also on our future agenda.

APPENDIX
PROOF OF THE THEOREMS

Theorem 1. $s^j(v_j^x, v_j^y) = 1$, if and only if $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$ for every attribute a_j and when $\alpha \neq 0$.

Proof 1. We prove its necessity first. According to Equation (10), if $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$, then $s^j(v_j^x, v_j^y) = 1$.

We then prove its sufficiency by contradiction. Suppose $s^j(v_j^x, v_j^y) = 1$, then $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$ is false. Accordingly, the true cases may be one of following cases.

1) $s_{Ie}^j(v_j^x, v_j^y) = 1, s_{Ia}^j(v_j^x, v_j^y) \neq 1$:
so, $s^j(v_j^x, v_j^y) = 1$
$\Leftrightarrow \alpha + (1-\alpha)s_{Ia}^j = 1$
$\Leftrightarrow s_{Ia}^j = 1\ (\alpha \in [0, 1))$
This result contradicts the assumption that $s_{Ia}^j(v_j^x, v_j^y) \neq 1$.
2) $s_{Ie}^j(v_j^x, v_j^y) \neq 1, s_{Ia}^j(v_j^x, v_j^y) = 1$:
so, $s^j(v_j^x, v_j^y) = 1$
$\Leftrightarrow \alpha s_{Ie}^j + (1-\alpha) = 1$
$\Leftrightarrow s_{Ie}^j = 1\ (\alpha \in [0, 1))$
$\Leftrightarrow \alpha = 0$
This result contradicts the assumption that $s_{Ie}^j(v_j^x, v_j^y) \neq 1$.
3) $s_{Ie}^j(v_j^x, v_j^y) \neq 1, s_{Ia}^j(v_j^x, v_j^y) \neq 1$:
so, $s^j(v_j^x, v_j^y) = 1$
$\Leftrightarrow \alpha s_{Ie}^j + (1-\alpha)s_{Ia}^j = 1$
$\Leftrightarrow s_{Ia}^j = 1$ (if $\alpha = 0$)
This result contradicts the assumption that $s_{Ia}^j(v_j^x, v_j^y) \neq 1$.
or $\alpha \neq 0$ and $\alpha \in (0, 1)$
$\Leftrightarrow \alpha = \frac{1-s_{Ia}^j}{s_{Ie}^j - s_{Ia}^j}$
Since $s_{Ie}^j(v_j^x, v_j^y) \in (0, 1]$, $\frac{1-s_{Ia}^j}{s_{Ie}^j - s_{Ia}^j} \geq 1$ contradicts $\alpha < 1$.
Hence, we conclude that $s_{Ie}^j(v_j^x, v_j^y) = s_{Ia}^j(v_j^x, v_j^y) = 1$ if $s^j(v_j^x, v_j^y) = 1$.

Theorem 2. $s^j(v_j^x, v_j^y) = 1$, if and only if $s_{Ia}^j(v_j^x, v_j^y) = 1$ for every attribute a_j and when $\alpha = 0$.

Proof 2. According to Equation (9), $s^j(v_j^x, v_j^y) = (1-\alpha)s_{Ia}^j = s_{Ia}^j = 1$.

Theorem 3. The coupled metric attribute value similarity s^j satisfies the triangle inequality if both intra-attribute similarity s_{Ia}^j and inter-attribute similarity s_{Ie}^j satisfy the triangle inequality for every attribute a_j.

Proof 3. According to the conditions defined in Section 3, s_{Ia}^j satisfying the triangle inequality means that

$$\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} \geq 1 + \frac{1}{s_{Ia}^j(v_j^x, v_j^z)},$$

s_{Ie}^j satisfying the triangle inequality means that

$$\frac{1}{s_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)} \geq 1 + \frac{1}{s_{Ie}^j(v_j^x, v_j^z)}.$$

Hence, according to Equation (9)

$$\begin{aligned}
&\frac{1}{s^j(v_j^x, v_j^y)} + \frac{1}{s^j(v_j^y, v_j^z)} \\
=&\frac{1}{\alpha s_{Ie}^j(v_j^x, v_j^y) + (1-\alpha)s_{Ia}^j(v_j^x, v_j^y)} + \\
&\frac{1}{\alpha s_{Ie}^j(v_j^y, v_j^z) + (1-\alpha)s_{Ia}^j(v_j^y, v_j^z)} \\
=&\frac{\alpha(\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} - 1)}{\alpha(\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} + \frac{1}{s_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)}) - 1} + \\
&\frac{(1-\alpha)(\frac{1}{ss_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)} - 1)}{\alpha(\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} + \frac{1}{s_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)}) - 1} + \\
&\frac{(\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} - 1)(\frac{1}{s_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)} - 1)}{\alpha(\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} + \frac{1}{s_{Ie}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ie}^j(v_j^y, v_j^z)}) - 1} \\
\geq&\frac{\alpha\frac{1}{s_{Ia}^j(v_j^x, v_j^z)} + (1-\alpha)\frac{1}{s_{Ie}^j(v_j^x, v_j^z)} + \frac{1}{s_{Ie}^j(v_j^x, v_j^z)}\frac{1}{s_{Ie}^j(v_j^x, v_j^z)}}{\alpha\frac{1}{s_{Ia}^j(v_j^x, v_j^z)} + (1-\alpha)\frac{1}{s_{Ie}^j(v_j^x, v_j^z)}} \\
=&\frac{\alpha s_{Ie}^j(v_j^x, v_j^z) + (1-\alpha)s_{Ia}^j(v_j^x, v_j^z) + 1}{\alpha s_{Ie}^j(v_j^x, v_j^z) + (1-\alpha)s_{Ia}^j(v_j^x, v_j^z)p} \\
=&1 + \frac{1}{\alpha s_{Ie}^j(v_j^x, v_j^z) + (1-\alpha)s_{Ia}^j(v_j^x, v_j^z)} \\
=&1 + \frac{1}{s^j(v_j^x, v_j^z)}
\end{aligned}$$

Consequently, we conclude that the coupled metric attribute value similarity s^j satisfies the triangle inequality.

Theorem 4. The intra-attribute similarity s_{Ia}^j satisfies the triangle inequality for any attribute a_j.

Proof 4. We here prove that

$$\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} \geq 1 + \frac{1}{s_{Ia}^j(v_j^x, v_j^z)}$$

Considering the following cases:

1) $v_j^x = v_j^y$ or $v_j^y = v_j^z$, or $v_j^x = v_j^y = v_j^z$:
According to Equation (3) and $s_{Ia}^j \in (0, 1]$, the following holds:

$$\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} \geq 1 + \frac{1}{s_{Ia}^j(v_j^x, v_j^z)}$$

Hence, s_{Ia}^j satisfies the triangle inequality for this case.
2) $v_j^x \neq v_j^y$ and $v_j^y \neq v_j^z$:

$$\begin{aligned}
&\frac{1}{s_{Ia}^j(v_j^x, v_j^y)} + \frac{1}{s_{Ia}^j(v_j^y, v_j^z)} - \frac{1}{s_{Ia}^j(v_j^x, v_j^z)} - 1 \\
=&\frac{\log(xy) + \log x \cdot \log y}{\log x \cdot \log y} + \frac{\log(yz) + \log y \cdot \log z}{\log y \cdot \log z} - \\
&(\frac{\log(xz) + \log x \cdot \log z}{\log x \cdot \log z} + 1) \\
=&\frac{2}{\log y}
\end{aligned}$$

Since $|I(v_j^y)| \geq 1$, $y = |I(v_j^y)| + 1 \geq 2$, accordingly $\frac{2}{\log y} \geq 0$

Therefore, we have

$$\frac{1}{s^j_{Ia}(v^x_j, v^y_j)} + \frac{1}{s^j_{Ia}(v^y_j, v^z_j)} \geq 1 + \frac{1}{s^j_{Ia}(v^x_j, v^z_j)}$$

Consequently, we conclude that the intra-attribute similarity s^j_{Ia} satisfies the triangle inequality for any attribute a_j.

Theorem 5. The inter-attribute similarity s^j_{Ie} satisfies the triangle inequality for any attribute a_j.

Proof 5. According to Equation(8), if $s^{k|j}_{Ie}$ satisfies the triangle inequality, then s^j_{Ie} satisfies it as well.
Considering the following cases:

1) $v^x_j = v^y_j$ or $v^y_j = v^z_j$, or $v^x_j = v^y_j = v^z_j$:
According to Equation (7) and $s^j_{Ie} \in (0, 1]$, the following holds:

$$\frac{1}{s^j_{Ie}(v^x_j, v^y_j)} + \frac{1}{s^j_{Ie}(v^y_j, v^z_j)} \geq 1 + \frac{1}{s^j_{Ie}(v^x_j, v^z_j)}$$

2) $v^x_j \neq v^y_j$ and $v^y_j \neq v^z_j$:

$$s^{k|j}_{Ie}(v^x_j, v^y_j) =$$

$$\frac{\sum_1^{|W_k|} \max(p^1_x, p^1_y)}{2 \cdot \sum_1^{|W_k|} \max(p^1_x, p^1_y) - \sum_1^{|W_k|} \min(p^1_x, p^1_y)}$$

According to the distance-similarity mapping function (see Equation (6)), the distance is:

$$dist = 1 - \frac{\sum_1^{|W_k|} \min(p^1_x, p^1_y)}{\sum_1^{|W_k|} \max(p^1_x, p^1_y)}$$

Note that the above is the Jaccard distance. The Jaccard distance is a metric distance and satisfies the triangle inequality. Accordingly, we conclude that $s^{k|j}_{Ie}$ satisfies the triangle inequality.

Hence, s^j_{Ie} satisfies the triangle inequality.

ACKNOWLEDGMENTS

This work is partially supported by the China National High-tech R&D Program of China (863 Program) Grants (2012AA01A301, 2012AA010901, 2012AA010303 and 2015AA01A301), Program for New Century Excellent Talents in University by National Science Foundation China (61272142, 61402492, 61402486, 61379146 and 61272483), and by the Australian Research Council Discovery Grant (DP130102691).

REFERENCES

[1] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *The Journal of Machine Learning Research*, vol. 12, pp. 491–523, 2011.
[2] I. Alabdulmohsin, M. Cisse, X. Gao, and X. Zhang, "Large margin classification with indefinite similarities," *Machine Learning*, pp. 1–23, 2016.
[3] H. Gao, X.-W. Liu, Y.-X. Peng, and S.-L. Jian, "Sample-based extreme learning machine with missing data," *Mathematical Problems in Engineering*, 2015.
[4] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *The Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.
[5] B. Yang, K. Lu, Y.-h. Gao, X.-p. Wang, and K. Xu, "GPU acceleration of subgraph isomorphism search in large scale graph," *Journal of Central South University*, vol. 22, pp. 2238–2249, 2015.
[6] N. Pradhan, M. Gyanchandani, and R. Wadhvani, "A review on text similarity technique used in ir and its application," *International Journal of Computer Applications*, vol. 120, no. 9, 2015.
[7] E. Diday and H.-H. Bock, "Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data," 2000.
[8] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," *red*, vol. 30, no. 2, p. 3, 2008.
[9] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.
[10] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou, "Coupled nominal similarity in unsupervised learning," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 973–978.
[11] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 8, pp. 1378–1392, 2012.
[12] T. Xu, J. Sun, and J. Bi, "Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1345–1354.
[13] L. Cao, "Non-iidness learning in behavioral and social data," *The Computer Journal*, vol. 57, no. 9, pp. 1358–1370, 2014.
[14] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-iid samples," in *Proceedings of the 26th annual International Conference on Machine Learning*. ACM, 2009, pp. 1249–1256.
[15] W. Ping, Y. Xu, K. Ren, C.-H. Chi, and S. Furao, "Non-iid multi-instance dimensionality reduction by learning a maximum bag margin subspace." in *AAAI*, 2010.
[16] M. C. Ganiz, C. George, and W. M. Pottenger, "Higher order Naive Bayes: A novel non-iid approach to text classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1022–1034, 2011.
[17] Z.-C. Guo and L. Shi, "Classification with non-iid sampling," *Mathematical and Computer Modelling*, vol. 54, no. 5, pp. 1347–1364, 2011.
[18] L. Ralaivola, M. Szafranski, and G. Stempfel, "Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes," *The Journal of Machine Learning Research*, vol. 11, pp. 1927–1956, 2010.
[19] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," *Journal of Multivariate Analysis*, vol. 100, no. 1, pp. 175–194, 2009.
[20] L. Cao, "Coupling learning of complex interactions," *Information Processing and Management*, vol. 51, no. 2, pp. 167–186, 2015.
[21] H. Xu and S. Mannor, "Robustness and generalization," *Machine Learning*, vol. 86, no. 3, pp. 391–423, 2012.
[22] C. Wang, X. Dong, F. Zhou, L. Cao, and C.-H. Chi, "Coupled attribute similarity learning on categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 781–797, 2015.
[23] C. Liu and L. Cao, "A coupled k-nearest neighbor algorithm for multi-label classification," *PAKDD2015*, 2015.
[24] X. Cheng, D. Miao, C. Wang, and L. Cao, "Coupled term-term relation analysis for document clustering," pp. 1–8, 2013.
[25] X. Meng, L. Cao, and J. Shao, "Semantic approximate keyword query based on keyword and query coupling relationship analysis," *CIKM 2014*, pp. 529–538, 2014.
[26] F. Li, G. Xu, and L. Cao, "Coupled matrix factorization within non-iid context," *PAKDD2015*, 2015.
[27] L. Cao, H. Zhang, Y. Zhao, D. Luo, and C. Zhang, "Combined mining: Discovering informative knowledge in complex data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 699 – 712, 2011.
[28] L. Cao, "Combined mining: Analyzing object and pattern relations for discovering and constructing complex but actionable patterns," *WIREs Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 140–155, 2013.
[29] Y. Shi, H.-I. Suk, Y. Gao, and D. Shen, "Joint coupled-feature representation and coupled boosting for ad diagnosis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2721–2728.
[30] Y. Shi, W. Li, Y. Gao, L. Cao, and D. Shen, "Beyond iid: Learning to combine non-iid metrics for vision tasks," *AAAI 2017*, pp. 1–7, 2017.
[31] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Siam, 2007, vol. 20.
[32] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.
[33] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 503–507, 2007.
[34] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[35] A. Bellet, A. Habrard, and M. Sebban, "Good edit similarity learning by loss minimization," *Machine Learning*, vol. 89, no. 1-2, pp. 5–35, 2012.
[36] Gray, Katherine R and Aljabar, Paul and Heckemann, Rolf A and Hammers, Alexander and Rueckert, Daniel and Alzheimer's Disease Neuroimaging Initiative and others, "Random forest-based similarity measures for multi-modal classification of alzheimer's disease," *Neuroimage*, vol. 65, pp. 167–175, 2013.
[37] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "An affinity-based new local distance function and similarity measure for knn algorithm," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 356–363, 2012.
[38] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Machine learning*, vol. 10, no. 1, pp. 57–78, 1993.
[39] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *JAIR*, vol. 6, pp. 1–34, 1997.
[40] V. Chandola, S. Boriah, and V. Kumar, "Understanding categorical similarity measures for outlier detection," *Technology Report, University of Minnesota*, 2008.
[41] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, pp. 120–127, 2012.
[42] M. E. Houle, V. Oria, and U. Qasim, "Active caching for similarity queries based on shared-neighbor information," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 2010, pp. 669–678.
[43] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2013, p. 6.
[44] G. Das and H. Mannila, "Context-based similarity measures for categorical databases," in *Principles of Data Mining and Knowledge Discovery*. Springer, 2000, pp. 201–210.
[45] V. Bryant, *Metric spaces: iteration and application*. Cambridge University Press, 1985.
[46] A. H. Lipkus, "A proof of the triangle inequality for the tanimoto distance," *Journal of Mathematical Chemistry*, vol. 26, no. 1-3, pp. 263–265, 1999.
[47] M. Levandowsky and D. Winter, "Distance between sets," *Nature*, vol. 234, no. 5323, pp. 34–35, 1971.
[48] H. Jia, Y.-m. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 5, pp. 1065–1079, 2016.
[49] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
[50] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.
[51] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 911–916.
[52] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.
[53] L. Cao, X. Dong, and Z. Zheng, "e-nsp: Efficient negative sequential pattern mining," *Artificial Intelligence*, vol. 235, pp. 156–182, 2016.

Songlei Jian Songlei Jian is currently working toward the Ph.D. degree, jointly supervised at the National University of Defense Technology, China and the University of Technology Sydney, Australia. Her research interests include data analytics, machine learning and complex network analysis.

Longbing Cao Longbing Cao is a Professor at the University of Technology Sydney. He has a PhD in Pattern Recognition and Intelligent Systems and another in Computing Sciences. His research interests include data science, analytics and machine learning, and behavior informatics and their enterprise applications.

Kai Lu Kai Lu is a Professor and the Deputy Dean of College of Computer Science, National University of Defense Technology, China. His research interests include parallel and distributed system software, operating systems, and big data analytics.

Hang Gao Hang Gao is a PhD candidate in computer science at the National University of Defense Technology, China. His research interests include machine learning and data mining.