Coupled Attribute Similarity Learning on Categorical Data

Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao, *Senior Member, IEEE*, and Chi-Hung Chi

***Abstract*— Attribute independence has been taken as a major assumption in the limited research that has been conducted on similarity analysis for categorical data, especially unsupervised learning. However, in real-world data sources, attributes are more or less associated with each other in terms of certain coupling relationships. Accordingly, recent works on attribute dependency aggregation have introduced the co-occurrence of attribute values to explore attribute coupling, but they only present a local picture in analyzing categorical data similarity. This is inadequate for deep analysis, and the computational complexity grows exponentially when the data scale increases. This paper proposes an efficient data-driven similarity learning approach that generates a coupled attribute similarity measure for nominal objects with attribute couplings to capture a global picture of attribute similarity. It involves the frequency-based intra-coupled similarity within an attribute and the inter-coupled similarity upon value co-occurrences between attributes, as well as their integration on the object level. In particular, four measures are designed for the inter-coupled similarity to calculate the similarity between two categorical values by considering their relationships with other attributes in terms of power set, universal set, joint set, and intersection set. The theoretical analysis reveals the equivalent accuracy and superior efficiency of the measure based on the intersection set, particularly for large-scale data sets. Intensive experiments of data structure and clustering algorithms incorporating the coupled dissimilarity metric achieve a significant performance improvement on state-of-the-art measures and algorithms on 13 UCI data sets, which is confirmed by the statistical analysis. The experiment results show that the proposed coupled attribute similarity is generic, and can effectively and efficiently capture the intrinsic and global interactions within and between attributes for especially large-scale categorical data sets. In addition, two new coupled categorical clustering algorithms, i.e., CROCK and CLIMBO are proposed, and they both outperform the original ones in terms of clustering quality on UCI data sets and bibliographic data.**

Manuscript received April 6, 2013; revised April 23, 2014 and May 5, 2014; accepted May 11, 2014. Date of publication June 13, 2014; date of current version March 16, 2015. This work was supported in part by the National Privacy Principles, Tasmania, Australia, in part by the Australian Research Council Discovery under Grant DP1096218, in part by the Australian Research Council Linkage under Grant LP100200774, in part by the National Natural Science Foundation of China under Grant 71271125, in part by the Natural Science Foundation of China under Grant 61301183, and in part by the China Post-Doctoral Science Foundation under Grant 2013M540947.

C. Wang and C.-H. Chi are with the Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, TAS 7005, Australia (e-mail: canwang613@gmail.com; chihungchi@gmail.com).

X. Dong is with the School of Information, Qilu University of Technology, Ji'nan 250353, China (e-mail: dongxiangjun@gmail.com).

F. Zhou is with the Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: flying.zhou@163.com).

L. Cao is with the Advanced Analytics Institute, University of Technology at Sydney, Ultimo, NSW 2008, Australia (e-mail: longbing.cao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2325872

***Index Terms*— Clustering, coupled attribute similarity, coupled object analysis, similarity analysis, unsupervised learning.**

I. Introduction

SIMILARITY analysis has been a problem of great practical importance in several domains for decades, not least in recent work, including behavior analysis [1], document analysis [2], and image analysis [3]. A typical aspect of these applications is clustering, in which the similarity is usually defined in terms of one of the following levels: 1) between clusters; 2) between attributes; 3) between data objects; or 4) between attribute values. The similarity between clusters is often built on top of the similarity between data objects, e.g., centroid similarity. Further, the similarity between data objects is generally derived from the similarity between attribute values, e.g., Euclidean distance and simple matching similarity (SMS) [4]. The similarity between attribute values assesses the relationship between two data objects and even between two clusters. The more two objects or clusters resemble each other, the larger is the similarity [5]. The other similarity between attributes [6] can also be converted into the difference of similarities between pairwise attribute values [7]. Therefore, the similarity between attribute values plays a fundamental role in similarity analysis.

The similarity measures for attribute values are sensitive to the attribute types, which are classified as discrete and continuous. The discrete attribute is further typed as nominal (categorical) or binary [5]. The nominal data, a special case of the discrete type, has only a finite number of values, while the binary variable has exactly two values. In this paper, we regard the binary data as a special case of the nominal data.

Compared with the intensive study on the similarity between two numerical variables, such as Euclidean and Minkowski distance, and between two categorical values in supervised learning, e.g., heterogeneous distance functions [8] and modified value distance matrix (MVDM) [9], the similarity for nominal variables has received much less attention in unsupervised learning on unlabeled data. Only limited efforts [5] have been made, including SMS, which uses 0s and 1s to distinguish the similarity between distinct and identical categorical values, occurrence frequency (OF) [10] and information-theoretical similarity (Lin) [10], [11], to discuss the similarity between nominal values. The challenge is that these methods are too rough to precisely characterize the similarity between categorical attribute values, and they only deliver a local picture of the similarity and are not data-driven. In addition, none of them provides a comprehensive picture of similarity

TABLE I
INSTANCE OF THE MOVIE DATABASE

Movie	Director	Actor	Genre	Class
Godfather II	Scorsese	De Niro	Crime	l_1
Good Fellas	Coppola	De Niro	Crime	l_1
Vertigo	Hitchcock	Stewart	Thriller	l_2
N by NW	Hitchcock	Grant	Thriller	l_2
Bishop's Wife	Koster	Grant	Comedy	l_2
Harvey	Koster	Stewart	Comedy	l_2

between categorical attributes by combining relevant aspects. Below, we illustrate the problem with SMS and the challenge of analyzing the categorical data similarity.

As shown in Table I, six movie objects are divided into two classes with three nominal attributes: 1) director; 2) actor; and 3) genre. The SMS measure between directors Scorsese and Coppola is 0, but Scorsese and Coppola are very similar.[1] Another observation by following SMS is that the similarity between Koster and Hitchcock is equal to that between Koster and Coppola; however, the similarity of the former pair should be greater because both directors belong to the same class l_2.

The above examples show that it is much more complex to analyze the similarity between nominal variables than between continuous data. The SMS and its variants fail to capture a global picture of the genuine relationship for nominal data. With the exponential increase of categorical data, such as that derived from social networks, it is important to develop effective and efficient measures for capturing the similarity between nominal variables.

The similarity between categorical values is sensitive to the data characteristics. In general, two attribute values are expected to be similar if they present analogous frequency distributions within one attribute (e.g., OF and Lin) [10], [11]; this reflects the intra-coupled similarity within attributes. For example, two directors are very similar if they appear with almost the same frequency, such as Scorsese with Coppola and Koster with Hitchcock. However, the reality is that the former director pair is more similar than the latter. Ahmad and Dey [12] introduced the co-occurrence probability of categorical values from different attributes and compared this probability for two categorical values from the same attribute. This means that the similarity between directors relates to the dependency of director on other attributes, such as actor and genre over all the movie objects, namely, the inter-coupled similarity between attributes. They both capture local pictures of the similarity from different perspectives. No work has been reported on systematically considering both intra-coupled similarity and inter-coupled similarity. The incomplete description of the categorical value similarity leads to tentative and less effective learning performance. In addition, it is usually very costly to consider the similarity between values in relation to the dependency between attributes and the aggregation of such dependency [12], which is verified in Section VI.

In this paper, we explicitly discuss the data-driven intra-coupled similarity and inter-coupled similarity, as well as their global aggregation in unsupervised learning on nominal data. The key contributions are as follows.

1) We propose a coupled attribute similarity for objects (CASO) measure based on the coupled attribute similarity for values (CASV), by considering both the intra-coupled and inter-coupled attribute value similarities (IaASV and IeASV), which globally capture the attribute value frequency distribution and attribute dependency aggregation with high accuracy and relatively low complexity.
2) We compare the accuracy and efficiency of the four proposed measures for IeASV in terms of four relationships: power set; universal set; joint set; and intersection set; and obtain the most efficient candidate based on the intersection set (i.e., IRSI) from theoretical and experimental aspects.
3) A method is proposed to flexibly define the dissimilarity metrics with the proposed similarity building blocks according to specific requirements.
4) The proposed measures are compared with the state-of-the-art metrics on various benchmark data sets in terms of the internal and external clustering criteria. All the results are statistically significant.
5) We propose two new coupled categorical clustering algorithms: CROCK and CLIMBO.

This paper is organized as follows. In Section II, we briefly review the related work. Preliminary definitions are specified in Section III. Section IV proposes the framework of the coupled attribute similarity analysis. Section V defines the intra-coupled similarity, inter-coupled similarity, and their aggregation. The theoretical analysis is given in Section VI. We describe the CASO algorithm in Section VII. The effectiveness of CASO is empirically studied in Section VIII, two new categorical clustering methods (CROCK and CLIMBO) are introduced, and a flexible method to define dissimilarity metrics is also developed. Section IX discusses the coupled nominal similarity with open issues. Finally, we conclude this paper in Section X.

II. RELATED WORK

Some surveys, in particular [5] and [10], discuss the similarity between categorical attributes. The usual practice is to binarize the data and use binary similarity measures rather than directly considering nominal data. Cost and Salzberg [9] proposed MVDM based on labels, Wilson and Martinez [8] performed a detailed study of heterogeneous distance functions for instance based learning, and Figueiredo *et al.* [2] introduced word co-occurrence features for text classification. Unlike our focus, their similarities are only designed for supervised approaches.

A. Nominal Similarity in Unsupervised Learning

There are a number of existing data mining techniques for the unsupervised learning of nominal data [10], [12]. Well-known metrics include SMS [4] and its diverse variants, such as Jaccard coefficients [13], which are all intuitively based on the principle that the similarity measure is 1 with

[1] A conclusion drawn from a well-informed cinematic source.

identical values and 0 otherwise, which are not data driven. More recently, the frequency distribution of attribute values has been considered for similarity measures [10], such as OF and Lin. Similarity computation has been incorporated into the learning algorithm without explicitly defining general measures [14]. Neighborhood-based similarity [15], [16] was also explored to measure the proximity of objects using functions that operate on the intersection of two neighborhoods. They present the similarity between a pair of objects by considering only the relationships among data objects, which are built on the similarity between attribute values simply quantified by the variants of SMS. However, the couplings between attributes involve the similarity both between attribute values and between data objects. Such couplings are catered in our proposed similarity measure between attribute values, which is incorporated with the neighborhood-based similarity between data objects to more precisely describe the neighborhood of an object. It represents the neighborhood-based metric as a meta-similarity measure [10] in terms of both the couplings between attributes and between objects.

All the above methods are attribute-independent since similarity is calculated separately for two categorical values of individual attributes. However, an increasing number of researchers argue that the attribute value similarity is also dependent on the couplings of other attributes [1], [10]. The Pearson correlation coefficient [15] measures only the strength of linear dependence between two numerical variables. Das and Mannila [6] put forward the iterated contextual distances algorithm, believing that the attribute, object, and subrelation similarities are inter-dependent. They convert each object with binary attribute values to a continuous vector by a kernel smoothing function, and define the similarity between objects as the Manhattan distance between continuous vectors [6]. By contrast, we directly consider similarity for categorical values to maintain the least information loss. Andritsos *et al.* [17] introduced a context sensitive dissimilarity measure between attribute values based on the Jensen–Shannon divergence. Similarly, Ahmad and Dey [12] proposed an algorithm ADD to compute the dissimilarity between attribute values by considering the co-occurrence probability between each attribute value and the values of another attribute. Though the dissimilarity metric leads to high accuracy, the computation is usually very costly [12], which limits its application in large-scale problems. In addition, Ahmad and Dey's [12] approaches only focus on the interactions among different attributes, whereas our proposed measure also considers the couplings within each attribute globally.

B. Categorical Clustering

Clustering algorithms [16], including partition-based methods, such as k-means and hierarchy-based methods like divisive approaches [5], are more suitable for clustering data with numerical attributes than categorical data.

Clustering of categorical data (categorical clustering for short) is a difficult, yet important task. Many fields, from statistics to psychology, deal with categorical data. Despite this fact, categorical clustering has received limited attention with only a handful of relevant publications. Guha *et al.* [16] proposed a robust hierarchical clustering algorithm ROCK, which uses the link-based similarity measure to measure the similarity between two categorical data points and between two clusters. Gibson *et al.* [14] first constructed a hypergraph according to the database, and then cluster the hypergraph using a discrete dynamic system STIRR. Andritsos *et al.* [17] introduced a scalable hierarchical categorical clustering algorithm LIMBO that builds on the information bottleneck (IB) framework for quantifying the relevant information preserved when clustering. An incremental algorithm called COOLCAT [18] was proposed to cluster categorical attributes using entropy; however, it is based on the assumption of independence between the attributes. Clustering with sLOPE (CLOPE), presented in [19], uses a global criterion function instead of a local one defined by the pairwise similarity to cluster categorical data, especially transactional data. Rather than a measure of similarity, CLICKS [20] uses a graph-theoretic approach to find k disjoint sets of vertices in a graph constructed for a particular data set. The last three algorithms, i.e., COOLCAT, CLOPE, and CLICKS, have a different focus from our proposed coupled similarity.

Therefore, in the experiments, we compare the clustering quality of ROCK, STIRR, and LIMBO with the coupled versions of them, i.e., when our proposed coupled similarity measure replaces the original similarity measure between attribute values in ROCK and LIMBO.

TABLE II
EXAMPLE OF INFORMATION TABLE

U \ A	a_1	a_2	a_3
u_1	$\mathcal{A}_1$	$\mathcal{B}_1$	$\mathcal{C}_1$
u_2	$\mathcal{A}_2$	$\mathcal{B}_1$	$\mathcal{C}_1$
u_3	$\mathcal{A}_2$	$\mathcal{B}_2$	$\mathcal{C}_2$
u_4	$\mathcal{A}_3$	$\mathcal{B}_3$	$\mathcal{C}_2$
u_5	$\mathcal{A}_4$	$\mathcal{B}_3$	$\mathcal{C}_3$
u_6	$\mathcal{A}_4$	$\mathcal{B}_2$	$\mathcal{C}_3$

III. PRELIMINARY DEFINITIONS

A large number of data objects with the same attribute set can be organized by an information table $S =< U, A, V, f >$, where universe $U = \{u_1, \ldots, u_m\}$ is composed of a nonempty finite set of data objects; $A = \{a_1, \ldots, a_n\}$ is a finite set of attributes; $V = \cup_{j=1}^{n} V_j$ is a collection of attribute value sets, in which V_j is the set of attribute values from attribute $a_j (1 \leq j \leq n)$; and $f = \cup_{j=1}^{n} f_j$, $f_j : U \rightarrow V_j (1 \leq j \leq n)$ is an information function that assigns a particular value of attribute a_j to every object. For instance, Table II is an information table consisting of six objects $\{u_1, \ldots, u_6\}$ and three attributes $\{a_1, a_2, a_3\}$, the attribute value of object u_1 for attribute a_2 is $f_2(u_1) = \mathcal{B}_1$, and the set of all attribute values for a_2 is $V_2 = \{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3\}$.

Generally speaking, the similarity between two objects $u_x, u_y (\in U)$ can be built on top of the similarities between their attribute values $v_j^x, v_j^y (\in V_j)$ for all attributes $a_j \in A$. Here, v_j^x and v_j^y indicate the respective attribute values of objects u_x and u_y for the attribute a_j, for example, $v_2^1 = \mathcal{B}_1$

and $v_1^2 = \mathcal{A}_2$. By proposing a coupled attribute value similarity measure, we define a new object similarity for categorical data. The basic concepts below facilitate the formulation for a coupled attribute value similarity measure. They are exemplified by Table II. Below, an information table S is given, and |set| is the number of elements in a certain set.

Definition 3.1 (SIF): Two set information functions (SIFs) are defined as

$$F_j : 2^U \to 2^{V_j}, \quad F_j(U') = \{f_j(u_x)|u_x \in U'\} \tag{1}$$

$$G_j : 2^{V_j} \to 2^U, \quad G_j(V_j') = \{u_i|f_j(u_i) \in V_j'\} \tag{2}$$

where $1 \le j \le n$, $1 \le i \le m$, $U' \subseteq U$, and $V_j' \subseteq V_j$.

These SIFs describe the relationships between objects and attribute values from different levels. Function $F_j(U')$ assigns the associated value set of attribute a_j to the object set U'. Function $G_j(V_j')$ maps the value set V_j' of attribute a_j to the dependent object set. For example, based on the attribute a_2, $F_2(\{u_1, u_2, u_3\}) = \{\mathcal{B}_1, \mathcal{B}_2\}$ collects the attribute values of u_1, u_2 and u_3; and $G_2(\{\mathcal{B}_1, \mathcal{B}_2\}) = \{u_1, u_2, u_3, u_6\}$ returns the objects whose attribute values are $\mathcal{B}_1$ and $\mathcal{B}_2$. In Table I on the movie data, $G_1(\{\text{Hitchcock}\}) = \{\text{Vertigo, N by NW}\}$ while $F_2(\{\text{Vertigo, N by NW}\}) = \{\text{Stewart, Grant}\}$.

Note that in the two definitions below, the superscripts x and y of v_j are omitted, since any attribute value $v_j \in V_j$ used here is independent of the objects u_x and u_y. However, v_j^x and v_j^y are reused when defining the similarity in the following sections.

Definition 3.2 (IIF): The inter-information function (IIF) obtains a value subset of attribute a_k for the corresponding objects, which are derived from the value v_j of attribute a_j. It is defined as

$$\varphi_{j\to k} : V_j \to 2^{V_k}, \quad \varphi_{j\to k}(v_j) = F_k(G_j(\{v_j\})). \tag{3}$$

This IIF $\varphi_{j\to k}$ is the composition of F_k and G_j. The involved subscript $j \to k$ means that this mapping φ is performed from attribute a_j to attribute a_k. Intuitively, $\varphi_{j\to k}(v_j)$ computes the set of attribute values from attribute a_k that co-occurs with a particular attribute value v_j from attribute a_j. In other words, $\varphi_{j\to k}(v_j)$ returns an attribute value subset of V_k that shares the same objects as v_j. For example, $\varphi_{2\to 1}(\mathcal{B}_1) = \{\mathcal{A}_1, \mathcal{A}_2\}$ specifies the values $\mathcal{B}_1$ of attribute a_2 and $\{\mathcal{A}_1, \mathcal{A}_2\}$ of attribute a_1 are related by the corresponding objects: 1) u_1 and 2) u_2. Likewise, in Table I, $\varphi_{1\to 2}(\text{Hitchcock}) = \{\text{Stewart, Grant}\}$ due to the connected objects: Vertigo and N by NW.

Definition 3.3 (ICP): The value subset $V_k'(\subseteq V_k)$ of attribute a_k, and the value $v_j(\in V_j)$ of attribute a_j, then the information conditional probability (ICP) of V_k' with respect to v_j is $P_{k|j}(V_k'|v_j)$, defined as

$$P_{k|j}(V_k'|v_j) = \frac{|G_k(V_k') \bigcap G_j(\{v_j\})|}{|G_j(\{v_j\})|}. \tag{4}$$

Intuitively, when given all the objects with the value v_j of attribute a_j, ICP is the percentage of common objects whose values of attribute a_k fall in subset V_k' and whose values of attribute a_j are exactly v_j as well. For example, $P_{1|2}(\{\mathcal{A}_1\}|\mathcal{B}_1) = 0.5$. Hence, ICP quantifies the relative overlapping ratio of attribute values in terms of objects.

TABLE III
LIST OF MAIN NOTATIONS

Variable	Explanation
$\{u_1, \cdots, u_m\}$	The set of m objects U
$\{a_1, \cdots, a_n\}$	The set of n attributes A
$l(\in L)$	Any label in the label (class) set L
$V_j'(\subseteq V_j)$	The subset of value set V_j of attribute a_j
$R(= \max\|V_j\|)$	The maximal number of values of each attribute
$v_j^x, v_j^y(\in V_j)$	Specific values of attribute a_j for objects u_x, u_y
$v_k(\in V_k)$	Any value of attribute a_k

TABLE IV
LIST OF ABBREVIATIONS

Abbreviation	Full Name
IaASV (δ_j^{Ia})	Intra-coupled Attribute Similarity for Values
IRSP ($\delta_{j|k}^{P}$)	Inter-coupled Relative Similarity based on Power Set
IRSU ($\delta_{j|k}^{U}$)	Inter-coupled Relative Similarity based on Universal Set
IRSJ ($\delta_{j|k}^{J}$)	Inter-coupled Relative Similarity based on Join Set
IRSI ($\delta_{j|k}^{I}$)	Inter-coupled Relative Similarity based on Intersection Set
IeASV (δ_j^{Ie})	Inter-coupled Attribute Similarity for Values
CASV (δ_j^{A})	Coupled Attribute Similarity for Values
CASO ($CASO$)	Coupled Attribute Similarity for Objects
CADO ($CADO$)	Coupled Attribute Dissimilarity for Objects

Back to Table I, $P_{2|1}(\{\text{De Niro, Stewart}\}|\text{Hitchcock}) = 0.5$ since actor subset {De Niro, Stewart} and director Hitchcock co-occur in only one movie Vertigo given Hitchcock directs two movies: Vertigo and N by NW. Note that the use of subset V_k' and element v_j is to facilitate the definitions in Section V, and make them mathematically solid and consistent.

All these concepts and functions form the foundation for formalizing the coupled interactions within and between categorical attributes, as presented below. The main notations in this paper are listed in Table III. In addition, several important abbreviations are defined in Table IV to facilitate the reading of this paper.

IV. FRAMEWORK OF THE COUPLED ATTRIBUTE SIMILARITY ANALYSIS

In this section, a framework for coupled attribute similarity analysis is proposed from a global perspective of the intra-coupled interaction within an attribute, the inter-coupled interaction among multiple attributes, and the integration of both.

With respect to the intra-coupled interaction, the similarity between attribute values is considered by examining their occurrence frequencies within one attribute. For the inter-coupled interaction, the similarity between attribute values is captured by exposing their co-occurrence dependency on the values of other attributes. For example, the coupled value similarity between $\mathcal{B}_1$ and $\mathcal{B}_2$ (i.e., values of attribute a_2) concerns both the intra-coupled relationship specified by the repeated times of values $\mathcal{B}_1$ and $\mathcal{B}_2$: 2 and 2, and the inter-coupled interaction triggered by the other two attributes (a_1 and a_3). The coupled interaction is then derived by the integration of intra-coupling and inter-coupling. In this way, the couplings of attributes lead to more accurate similarity ($\in [0, 1]$) between attribute values, rather than a rude assignment of either 0 or 1.

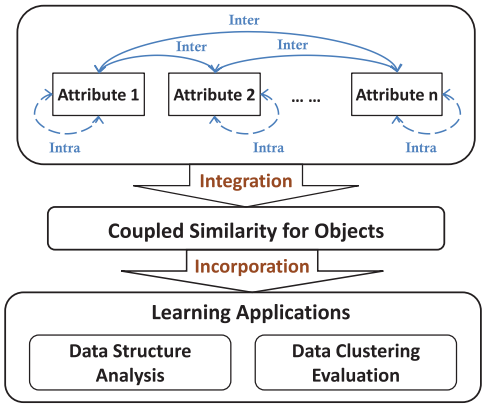

Fig. 1. Framework of coupled attribute similarity analysis, where ←----→ indicates intra-coupling and ⟷ refers to inter-coupling.

In the framework described in Fig. 1, the couplings of attributes are revealed via the similarity between attribute values v_j^x (e.g., Scorsese in Table I) and v_j^y (e.g., Coppola in Table I) of each attribute a_j (e.g., Director in Table I) by means of the intra-coupling and inter-coupling. Further, the coupled similarity for objects is built on top of the pairwise similarity between attribute values according to the integration of couplings. Finally, two learning tasks are explored for the data structure analysis and data clustering evaluation by incorporating the coupled interactions, revealing that the couplings of attributes are essential to applications in empirical studies.

Given an information table S with a set of m objects U and a set of n attributes A, we specify those interactions and couplings in the following sections.

V. Coupled Attribute Similarity

The attribute couplings are proposed in terms of both intra-coupled and inter-coupled similarities. Below, the intra-coupled and inter-coupled relationships, as well as the integrated coupling, are formalized and exemplified. Note that all the main notations are listed in Table III.

A. Intra-Coupled Interaction

According to [5], the discrepancy in attribute value occurrence times reflects the value similarity in terms of frequency distribution. It reveals that greater similarity is assigned to the attribute value pair which owns approximately equal frequencies. The higher these frequencies are, the closer the two values are. Different occurrence frequencies therefore indicate distinct levels of attribute value significance.

These principles are also consistent with the similarity theorem presented in [11], in which the commonality corresponds to the product of frequencies and the full description relates to the total sum of individual frequencies and their product. In addition, a comparative evaluation on similarity measures for categorical data has been done in [10], delivering OF and Lin as the two best similarity measures among 14 existing measures on 18 data sets. Both these measures assign higher weights to mismatches or matches on frequent values, and the maximum similarity is attained when the attribute values exhibit approximately equal frequencies [10].

Thus, when calculating attribute value similarity (e.g., the similarity between Scorsese and Coppola in Table I), we consider the relationship between the attribute value frequencies of an attribute, proposed as intra-coupled similarity to satisfy the above principles.

Definition 5.1 (IaASV): The intra-coupled attribute similarity for values (IaASV) between values v_j^x and v_j^y of attribute a_j is

$$\delta_j^{Ia}(v_j^x, v_j^y) = \frac{|G_j(\{v_j^x\})|\cdot|G_j(\{v_j^y\})|}{|G_j(\{v_j^x\})|+|G_j(v_j^y)|+|G_j(\{v_j^x\})|\cdot|G_j(\{v_j^y\})|}. \quad (5)$$

Since $1 \le |G_j(v_j^x)|, |G_j(v_j^y)| \le m$ and $2 \le |G_j(v_j^x)| + |G_j(v_j^y)| \le m$, then $\delta_j^{Ia} \in [1/3, m/(m+4)]$ is obtained according to Proof (a) in the Appendix. For example, in Table II, both $\mathcal{B}_1$ and $\mathcal{B}_2$ are observed twice, $\delta_2^{Ia}(\mathcal{B}_1, \mathcal{B}_2) = 0.5$. In Table I, we have $\delta_1^{Ia}(\text{Scorsese}, \text{Coppola}) = 1/3$ since both directors Scorsese and Coppola appear once, i.e., $|G_1(\{\text{Scorsese}\})| = |G_1(\{\text{Coppola}\})| = 1$.

Note that there is still an issue in the above definition: if two attribute values v_j^x and v_j^y have the same frequency, then we have $\delta_j^{Ia}(v_j^x, v_j^x) = \delta_j^{Ia}(v_j^x, v_j^y)$. This is somewhat intuitively problematic, but the inter-coupled similarity proposed in the next section remedies this issue because the inter-coupled similarities between v_j^x, v_j^x and between v_j^x, v_j^y are overwhelmingly distinct.

By taking the frequency of attribute values into consideration, IaASV characterizes the value similarity in terms of attribute value occurrence times.

B. Inter-Coupled Interaction

The IaASV considers the interaction between attribute values within an attribute a_j. It does not involve the couplings between attributes [e.g., $a_k (k \neq j)$ and a_j] when calculating the attribute value similarity. For this, we discuss the dependency aggregation, i.e., inter-coupled interaction.

In 1993, Cost and Salzberg [9] presented a powerful new method MVDM for measuring the dissimilarity between categorical values of a given attribute. The MVDM considers the overall similarity of classification of all objects on each possible value of each attribute. The dissimilarity $D_{j|L}$ between two attribute values v_j^x and v_j^y (e.g., Scorsese and Coppola in Table I) for a specific attribute a_j (e.g., Director in Table I) regarding labels L (e.g., Class in Table I) is

$$D_{j|L}(v_j^x, v_j^y) = \sum_{l \in L} |P_{l|j}(\{l\}|v_j^x) - P_{l|j}(\{l\}|v_j^y)| \quad (6)$$

where l ($\in L$, e.g., l_1 and l_2 in Table I) is a label in the information table. $P_{l|j}$ is the ICP defined in (4) by replacing the attribute a_k with the label l, the attribute value subset V_k' with the label subset $L' \subseteq L$ (here $L' = \{l\}$), in which $G_l(L')$ refers to the set of objects whose labels fall in L'. $D_{j|L}$ indicates that values are identified as being similar if they occur with the same relative frequency for all classes.

TABLE V
EXAMPLE OF COMPUTING SIMILARITY USING IRSP

V_1'	$\overline{V_1'}$	$P_{1\|2}(V_1'\|\mathcal{B}_1)$	$P_{1\|2}(\overline{V_1'}\|\mathcal{B}_2)$	$2-P_{1\|2}(V_1'\|\mathcal{B}_1)-P_{1\|2}(\overline{V_1'}\|\mathcal{B}_2)$
$\varnothing$	$\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,\mathcal{A}_4\}$	0	1	1
$\{\mathcal{A}_1\}$	$\{\mathcal{A}_2,\mathcal{A}_3,\mathcal{A}_4\}$	0.5	1	0.5
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,\mathcal{A}_4\}$	$\varnothing$	1	0	1

The idea behind is that we wish to establish that values are similar if they occur with the same relative frequency for all classifications.

According to the principle [21] that, for the categorical data distribution, the sum of $L1$ dissimilarities and twice the total variation dissimilarity are equivalent, we have

$$D_{j|L}(v_j^x, v_j^y) = 2 \cdot \max_{L' \subseteq L} |P_{l|j}(L'|v_j^x) - P_{l|j}(L'|v_j^y)|. \tag{7}$$

The detailed proof on the equivalence of (6) and (7) is specified by Proof (b) in the Appendix.

In the absence of labels, the above (7) is adapted to satisfy our target problem by replacing the class label information with other attribute knowledge to enable unsupervised learning. In Table I, for instance, we consider the attribute Actor rather than Class when calculating the similarity between Scorsese and Coppola for Director, since Class is invisible in unsupervised learning process. We regard this interaction between attributes as inter-coupled similarity in terms of the co-occurrence comparisons of ICP. The most intuitive variant of (7) is IRSP.

Definition 5.2 (IRSP): The inter-coupled relative similarity based on power set (IRSP) between values v_j^x and v_j^y of attribute a_j based on another attribute a_k is defined as $\delta_{j|k}^P(v_j^x, v_j^y, V_k)$ (below $\delta_{j|k}^P$ for short)

$$\delta_{j|k}^P = \min_{V_k' \subseteq V_k} \{2 - P_{k|j}(V_k'|v_j^x) - P_{k|j}(\overline{V_k'}|v_j^y)\} \tag{8}$$

where $\overline{V_k'} = V_k \backslash V_k'$ is the complementary set of a set V_k' under the complete value set V_k of attribute a_k.

The main difference between (8) and (7) includes: 1) the multiplier 2 in (7) is omitted; 2) labels are replaced with other values of a particular attribute a_k, i.e., V_k' and V_k are substituted for L' and L, respectively; 3) a complementary set $\overline{V_k'}$ rather than the original set V_k' is concerned for v_j^y in ICP, note that $P_{k|j}(\overline{V_k'}|v_j^y) = 1 - P_{k|j}(V_k'|v_j^y)$; and 4) dissimilarity is considered rather than similarity; the new dissimilarity measure

$$D'_{j|k}(v_j^x, v_j^y) = \max_{V_k' \subseteq V_k} |P_{k|j}(V_k'|v_j^x) + P_{k|j}(\overline{V_k'}|v_j^y) - 1| \tag{9}$$

is obtained by following the previous three steps, then we have $\delta_{j|k}^P = 1 - D'_{j|k}(v_j^x, v_j^y)$. The detailed conversion process is provided in Proof (c) in the Appendix. Two attribute values are closer to each other if they have more similar probabilities with other attribute value subsets in terms of co-occurrence object frequencies.

In Table II, by employing (8), we want to obtain $\delta_{2|1}^P(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}_{i=1}^4)$, i.e., the similarity between two attribute values $\mathcal{B}_1, \mathcal{B}_2$ of a_2 regarding attribute a_1 with its values $\{\mathcal{A}_i\}_{i=1}^4$. As shown in Table V, the set of all attribute values of a_1 is $V_1 = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$. The number of all power sets within V_1 is 2^4, i.e., the number of the combinations consisting of $V_1' \subseteq V_1$ and $\overline{V_1'} \subseteq V_1$ is 2^4. In detail, for the second row when we consider $V_1' = \{\mathcal{A}_1\}$ and $\overline{V_1'} = \{\mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$, we have $P_{1|2}(V_1'|\mathcal{B}_1) = |\{u_1\}|/|\{u_1, u_2\}| = 0.5$ and $P_{1|2}(\overline{V_1'}|\mathcal{B}_2) = |\{u_3, u_6\}|/|\{u_3, u_6\}| = 1$. Therefore, $2 - P_{1|2}(V_1'|\mathcal{B}_1) - P_{1|2}(\overline{V_1'}|\mathcal{B}_2) = 2 - 0.5 - 1 = 0.5$. The other elements in the power set of V_1 follow the same rule. Accordingly, the minimal value among them is 0.5, which indicates that the corresponding similarity $\delta_{2|1}^P$ is 0.5.

This process shows that the combinational explosion brought about by the power set needs to be considered when calculating attribute value similarity by IRSP. For a given set of attribute values, the power set considers all the subsets. However, the universal set concerns all the elements involved, which effectively reduces the number of items involved. The joint and intersection sets focus on parts of the elements, which further optimize the calculation time. We start with the power set-based IRSP, and will proceed to the universal set-based IRSU, the joint set-based IRSJ, and the intersection set-based IRSI to see whether the problem can be reduced in this way. We therefore try to define three more similarity metrics IRSU, IRSJ, and IRSI based on IRSP.

Definition 5.3 (IRSU, IRSJ, IRSI): The inter-coupled relative similarity based on universal set (IRSU), joint set (IRSJ), and intersection set (IRSI) between values v_j^x and v_j^y of attribute a_j based on another attribute a_k are defined as $\delta_{j|k}^U(v_j^x, v_j^y, V_k)$, $\delta_{j|k}^J(v_j^x, v_j^y, V_k)$, and $\delta_{j|k}^I(v_j^x, v_j^y, V_k)$ (below $\delta_{j|k}$, $\delta_{j|k}^J$, and $\delta_{j|k}^I$ for short), respectively

$$\delta_{j|k}^U = 2 - \sum_{v_k \in V_k} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\} \tag{10}$$

$$\delta_{j|k}^J = 2 - \sum_{v_k \in \cup} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\} \tag{11}$$

$$\delta_{j|k}^I = \sum_{v_k \in \cap} \min\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\} \tag{12}$$

where $v_k \in \cup$ and $v_k \in \cap$ denote $v_k \in \varphi_{j\to k}(v_j^x) \cup \varphi_{j\to k}(v_j^y)$ and $v_k \in \varphi_{j\to k}(v_j^x) \cap \varphi_{j\to k}(v_j^y)$, respectively.

In the above definition, universal set, joint set, and intersection set are proposed to quantify the inter-coupled similarity based on the power set. Let us explain the intuitive meaning and understanding behind those complex formulas. For example, in Table I, the IRSP similarity between directors Scorsese and Coppola is measured by examining their individual connections with all the actor subsets ($V_2' \subseteq V_2$), such as {De Niro}, {De Niro, Stewart}, {Stewart,

TABLE VI
COMPUTING SIMILARITY USING IRSU

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	max
$\mathcal{A}_1$	0.5	0	0.5
$\mathcal{A}_2$	0.5	0.5	0.5
$\mathcal{A}_3$	0	0	0
$\mathcal{A}_4$	0	0.5	0.5

TABLE VII
COMPUTING SIMILARITY USING IRSJ

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	max
$\mathcal{A}_1$	0.5	0	0.5
$\mathcal{A}_2$	0.5	0.5	0.5
$\mathcal{A}_4$	0	0.5	0.5

Grant}.[2] Alternatively, the IRSU similarity between directors Scorsese and Coppola is defined upon all the actors ($v_2 \in V_2$), including De Niro, Stewart, and Grant. The IRSJ similarity between directors Scorsese and Coppola is further built on the subset of {De Niro, Stewart, Grant} according to the joint rule ($v_2 \in \cup$), which produces all the possible values to share objects. The IRSI similarity between directors Scorsese and Coppola is composed by another subset of {De Niro, Stewart, Grant} based on the intersection rule ($v_2 \in \cap$), which returns the common values to share objects. In addition, the joint and intersection rules correspond to different selection schemes of actors that co-occur with directors Scorsese and/or Coppola in movies. The former collects the co-occurrence actors with either Scorsese or Coppola, while the latter gains the co-occurrence actors with both Scorsese and Coppola. As discussed later in Section VI, these four options are actually equivalent to one another, though present varying computational efficiency.

In detail, each value $v_k (\in V_k)$ of attribute a_k, rather than its value subset $V'_k \subseteq V_k$, is considered to reduce computational complexity. As shown in Table VI, we have $P_{1|2}(\{\mathcal{A}_1\}|\mathcal{B}_1) = |\{u_1\}|/|\{u_1, u_2\}| = 0.5$ and $P_{1|2}(\{\mathcal{A}_1\}|\mathcal{B}_2) = |\varnothing|/|\{u_3, u_6\}| = 0$ when v_k takes $\mathcal{A}_1$, thus the maximum is 0.5. Accordingly, the similarity $\delta^U_{2|1}$ based on IRSU is $\delta^U_{2|1}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}^4_{i=1}) = 2 - 0.5 - 0.5 - 0 - 0.5 = 0.5$. Since IRSU only concerns all the single attribute values rather than exploring the whole power set, it solves the combinational explosion issue to a great extent. In IRSU, ICP is merely calculated eight times compared with 32 times by IRSP, which leads to a substantial improvement in efficiency.

The IIF (3) is used to further reduce the time cost of ICP with two more similarity measures: IRSJ (11) and IRSI (12). With (11), the calculation of $\delta^J_{2|1}$ is further simplified since $\mathcal{A}_3 \notin \varphi_{2\to1}(\mathcal{B}_1) \cup \varphi_{2\to1}(\mathcal{B}_2)$, where $\varphi_{2\to1}(\mathcal{B}_1) = \{\mathcal{A}_1, \mathcal{A}_2\}$ and $\varphi_{2\to1}(\mathcal{B}_2) = \{\mathcal{A}_2, \mathcal{A}_4\}$. As shown in Table VII, we obtain $\delta^J_{2|1}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}^4_{i=1}) = 2 - 0.5 - 0.5 - 0.5 = 0.5$ by calculating ICP with $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_4\}$ rather than the whole value set $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$ of attribute a_1. This reveals the fact that it is enough to compute ICP with $w \in V_1$ that belongs to $\varphi_{2\to1}(\mathcal{B}_1) \cup \varphi_{2\to1}(\mathcal{B}_2)$ instead of all the elements in V_1. Thus, IRSJ further reduces the complexity compared with IRSU.

[2]This is just an illustration, the whole power set includes eight subsets.

TABLE VIII
COMPUTING SIMILARITY USING IRSI

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	min
$\mathcal{A}_2$	0.5	0.5	0.5

Based on IRSU, an alternative IRSI is concerned. With (12), the calculation of $\delta^I_{2|1}$ is once again simplified as in Table VIII since only $\mathcal{A}_2 \in \varphi_{2\to1}(\mathcal{B}_1) \cap \varphi_{2\to1}(\mathcal{B}_2)$, where $\varphi_{2\to1}(\mathcal{B}_1) = \{\mathcal{A}_1, \mathcal{A}_2\}$ and $\varphi_{2\to1}(\mathcal{B}_2) = \{\mathcal{A}_2, \mathcal{A}_4\}$. Then, we easily obtain $\delta^I_{2|1}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}^4_{i=1}) = 0.5$ as the final result, though ICP has been calculated only once. In this case, it is sufficient to compute ICP with $\mathcal{A}_2 \in V_1$, which only belongs to $\varphi_{2\to1}(\mathcal{B}_1) \cap \varphi_{2\to1}(\mathcal{B}_2)$ (i.e., $\{\mathcal{A}_1, \mathcal{A}_2\} \cap \{\mathcal{A}_2, \mathcal{A}_4\}$). In detail, we have $P_{1|2}(\{\mathcal{A}_2\}|\mathcal{B}_1) = |\{u_2\}|/|\{u_1, u_2\}| = 0.5$ and $P_{1|2}(\{\mathcal{A}_2\}|\mathcal{B}_2) = |\{u_3\}|/|\{u_3, u_6\}| = 0.5$, so the minimum is 0.5. It is trivial that the cardinality of intersection $\cap$ is always no larger than that of joint set $\cup$. Thus, IRSI is more efficient than IRSU due to the reduction of intra-coupled relative similarity complexity.

Intuitively, IRSI is the most efficient of all the proposed inter-coupled relative similarity measures: IRSP, IRSU, IRSJ, and IRSI. All four measures lead to the same similarity result, such as 0.5 in our example. These measures are mathematically equivalent to one another. This assumption is proved in Section VI.

Accordingly, the similarity between the value pair (v^x_j, v^y_j) of attribute a_j can be calculated on top of these four optional measures by aggregating all the relative similarity on attributes other than a_j. In Table I, the inter-coupled similarity between Scorsese and Coppola can be naturally constructed by summarizing both the co-occurrences from attributes other than Director, which are Actor and Genre.

Definition 5.4 (IeASV): The inter-coupled attribute similarity for values (IeASV) between attribute values v^x_j and v^y_j of attribute a_j is

$$\delta^{Ie}_j(v^x_j, v^y_j, \{V_k\}_{k\neq j}) = \sum_{k=1,k\neq j}^{n} \alpha_k \delta_{j|k}(v^x_j, v^y_j, V_k) \qquad (13)$$

where α_k is the weight parameter for attribute a_k, $\sum^n_{k=1,k\neq j} \alpha_k = 1$, $\alpha_k \in [0, 1]$, and $\delta_{j|k}(v^x_j, v^y_j, V_k)$ is one of the inter-coupled relative similarity candidates.

Therefore, $\delta^{Ie}_j \in [0, 1]$. Intuitively, δ^{Ie}_j calculates the inter-coupled similarity of values by aggregating all the connections between other attributes and the attribute a_j. Back to Table I, the inter-coupled similarity between Scorsese and Coppola (δ^{Ie}_1)[3] is determined by the respective co-occurrences of Actor $(\delta_{1|2})$ and Genre $(\delta_{1|3})$ with them. The parameter α_k specifies how strongly two attributes are dependent on each other. In this paper, we simply assign $\alpha_k = 1/(n-1)$, which indicates that every two attributes uniformly connect with each other. Thus, different similarity values rest with distinct extents of co-occurrence between attributes.

For example, in Table II, we have $\delta^{Ie}_2(\mathcal{B}_1, \mathcal{B}_2, \{V_1, V_3\}) = 0.5 \cdot \delta_{2|1}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}^4_{i=1}) + 0.5 \cdot \delta_{2|3}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{C}_i\}^3_{i=1}) = 0.25$

[3]The symbols within brackets are omitted if no ambiguity arises.

if α_1 and α_3 equal to 0.5. The calculation of the first component has been displayed above, while the second component can also be obtained by following the same approach. Alternatively, the parameter α_k which reflects the coupling weight of categorical attributes can also be defined to capture the average connection degree of attribute values inspired by [22] that proposes the support of attribute values. Later, we will explore and analyze this strategy in our future work.

C. Coupled Interaction

So far, we have built formal definitions for both IaASV and IeASV measures. The IaASV emphasizes the attribute value OF, while IeASV focuses on the co-occurrence comparison of ICP with four inter-coupled relative similarity options. Then, the CASV is naturally derived by simultaneously considering both measures.

Definition 5.5 (CASV): The CASV between attribute values v_j^x and v_j^y of attribute a_j is

$$\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) = \delta_j^{Ia}(v_j^x, v_j^y) \cdot \delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j}) \quad (14)$$

where $V_k (k \neq j)$ is a value set of attribute a_k different from a_j to enable the inter-coupled interaction. δ_j^{Ia} and δ_j^{Ie} are IaASV and IeASV, respectively, which will be detailed in the following sections.

As indicated in (14), CASV gets larger by increasing either IaASV or IeASV. Here, we choose the multiplication of these two components. The rationale is twofold: 1) IaASV is associated with how often the value occurs, while IeASV reflects the extent of the value difference brought by other attributes, hence intuitively, the multiplication of them indicates the total amount of attribute value difference and 2) the multiplication method is consistent with the adapted simple matching distance introduced in [5]. Alternatively, in our future work, we could consider other combination forms of IaASV and IeASV according to the data structure, such as $\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) = \beta \cdot \tilde{\delta}_j^{Ia}(v_j^x, v_j^y) + \gamma \cdot \delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j})$, where $\tilde{\delta}_j^{Ia} \in [0, 1]$ is the normalized intra-coupled similarity,[4] $0 \leq \beta, \gamma \leq 1$ $(\beta + \gamma = 1)$ are the corresponding weights. Thus, IaASV and IeASV can be controlled flexibly to display in which cases the former is more significant than the latter, and vice versa.

In addition, $\delta_j^A = \delta_j^{Ia} \cdot \delta_j^{Ie} \in [0, m/(m+4)]$ since we have $\delta_j^{Ia} \in [1/3, m/(m+4)](m \geq 2)$ as well as $\delta_j^{Ie} \in [0, 1]$. For example, in Table II, the CASV of attribute values $\mathcal{B}_1$ and $\mathcal{B}_2$ is $\delta_2^A(\mathcal{B}_1, \mathcal{B}_2, \{V_1, V_2, V_3\}) = \delta_2^{Ia}(\mathcal{B}_1, \mathcal{B}_2) \cdot \delta_2^{Ie}(\mathcal{B}_1, \mathcal{B}_2, \{V_1, V_3\}) = 0.5 \times 0.25 = 0.125$. For the Movie data, the coupled similarity between Scorsese and Coppola (δ_j^A) is obtained by multiplying the intra-coupled similarity (δ_1^{Ia}) characterized by frequency with the inter-coupled similarity (δ_1^{Ie}) quantified by co-occurrence. For other director pairs, we accordingly obtain that $\delta^A_{\text{Director}}(\text{Scorsese,Coppola}) = \delta^A_{\text{Director}}(\text{Coppola,Coppola}) = 0.33$, and $\delta^A_{\text{Director}}(\text{Koster, Coppol}a) = 0$ while $\delta^A_{\text{Director}}(\text{Koster, Hitchcock}) = 0.25$. They correspond to the fact that Scorsese and Coppola are very similar directors just as Coppola is to himself, and the similarity between Koster and Hitchcock is larger than that between Koster and Coppola, as clarified in Section I.

[4]The normalization can be made by $\tilde{\delta}_j^{Ia} = (\delta_j^{Ia} - \min)/(\max - \min)$, where min and max denote the minimum and maximum values of δ_j^{Ia} for each attribute, respectively.

TABLE IX
TIME COST OF ICP

Metric	Calculation Times of ICP	$\delta_{2\|1}(\mathcal{B}_1, \mathcal{B}_2)$
IRSP	$2 \cdot 2^{\|V_k\|}$	32
IRSU	$2 \cdot \|V_k\|$	8
IRSJ	$2 \cdot \|\varphi_{j\to k}(v_j^x) \bigcup \varphi_{j\to k}(v_j^y)\|$	6
IRSI	$2 \cdot \|\varphi_{j\to k}(v_j^x) \bigcap \varphi_{j\to k}(v_j^y)\|$	2

In the following theoretical analysis in Section VI, the computational accuracy and complexity of the four inter-coupled relative similarity options are analyzed.

VI. THEORETICAL ANALYSIS

This section compares the proposed four inter-coupled relative similarity measures (IRSP, IRSU, IRSJ, and IRSI) in terms of their computational accuracy and complexity.

A. Accuracy Equivalence

According to the set theory, these four measures are equivalent to one another in calculating value similarity; we therefore have the following theorem. This theorem is deduced by Proof (d) in the Appendix.

Theorem 6.1: The IRSP, IRSU, IRSJ, and IRSI are all equivalent to one another.

The above theorem indicates that IRSP, IRSU, IRSJ, and IRSI are equivalent to one another in terms of the information and knowledge they present. It also explains the similarity result in Section V-B. Thus, these measures can induce exactly the same computational accuracy in different learning tasks, including classification and clustering.

B. Computational Complexity Comparison

When calculating the similarity between every pair of attribute values for all attributes, the computational complexity linearly depends on the time cost of ICP, which is quantified by the calculation counts of ICP. This reflects the efficiency difference between distinct similarity measures. Table IX summarizes the time costs of the four inter-coupled relative similarity measures.

Let $|\text{ICP}_{j|k}^{(M)}|$ represent the time cost of ICP for $\delta_{j|k}^M(v_j^x, v_j^y)$ with the associated measure $M = \{P, U, J, I\}$, whose elements are IRSP, IRSU, IRSJ, and IRSI, respectively. From Table IX, $|\text{ICP}_{j|k}^{(P)}| \geq |\text{ICP}_{j|k}^{(U)}| \geq |\text{ICP}_{j|k}^{(J)}| \geq |\text{ICP}_{j|k}^{(I)}|$ holds constantly. It demonstrates the competitive efficiency of IRSI compared with the other three measures. In Table II, 32 calculation counts of ICP are required in IRSP, compared with only two calculation counts when using IRSI.

Suppose the maximal number of values for each attribute is $R(= \max_{j=1}^n |V_j|)$. In total, the number of value pairs for all the attributes is at most $n \cdot R(R-1)/2$, which is also the number of calculation steps. For each inter-coupled relative similarity, we calculate ICP for $|\text{ICP}_{j|k}^{(M)}|$ times. As we have n attributes,

TABLE X
COMPUTATIONAL COMPLEXITY FOR CASV

Metric	Calculation Steps	Flops per Step	Complexity
IRSP	$nR(R-1)/2$	$2(n-1)2^R$	$O(n^2R^22^R)$
IRSU	$nR(R-1)/2$	$2(n-1)R$	$O(n^2R^2R)$
IRSJ	$nR(R-1)/2$	$2(n-1)R_\cup$	$O(n^2R^2R)$
IRSI	$nR(R-1)/2$	$2(n-1)R_\cap$	$O(n^2R^2R)$

the total ICP time cost for CASV is $2\cdot|\text{ICP}_{j|k}^{(M)}|\cdot(n-1)$ flops per step. The computational complexity for calculating all four options of CASV is shown in Table X.

As indicated in Table X, all the measures have the same calculation steps, while their flops per step are sorted in descending order since $2^R > R \geq R_\cup \geq R_\cap$, in which $R_\cup$ and $R_\cap$ are the cardinality of the join and intersection sets of the corresponding IIFs, respectively. This evidences that the computational complexity essentially depends linearly on the time cost of ICP with given data. Specifically, IRSP has the largest complexity $O(n^2R^22^R)$, compared with the smaller equal ones $O(n^2R^3)$ presented by the other three measures (IRSU, IRSJ, and IRSI). Of the latter three candidates, though they have the same computational complexity, IRSI is the most efficient due to $R_\cap \leq R_\cup \leq R$. The dissimilarity ADD that Ahmad and Dey [12] used for mixed data clustering corresponds to the worst measure IRSP.

Considering both the accuracy analysis and complexity comparison, we conclude that IRSI is the best performing measure because it indicates the least complexity but maintains equal accuracy to present couplings. Thus, we only consider IRSI in the following algorithmic design and experimental comparisons.

VII. COUPLED SIMILARITY ALGORITHM

In previous sections, we have discussed the construction of CASV and its theoretical comparison among the inter-coupled relative similarity candidates. In this section, a coupled similarity between objects is built based on CASV. Below, we consider the sum of all these CASV measures, following the Manhattan dissimilarity [5].

Definition 7.1 (CASO): Given an information table S, the CASO between objects u_x and u_y is $\text{CASO}(u_x, u_y)$

$$CASO(u_x,u_y)=\sum_{j=1}^{n}\delta_j^A(v_j^x,v_j^y,\{V_k\}_{k=1}^n) \tag{15}$$

where δ_j^A is the CASV measure defined in (14), v_j^x and v_j^y are the attribute values of attribute a_j for objects u_x and u_y, respectively, and $1\leq x,y\leq m$, $1\leq j\leq n$.

For CASO, all the CASVs with each attribute are summed up for two objects. For example the similarity between u_2 and u_3 in Table II is $\text{CASO}(u_2,u_3)=\sum_{j=1}^{3}\delta_j^A(v_j^2,v_j^3,\{V_k\}_{k=1}^3)=0.5+0.125+0.125=0.75$.

The CASO has the properties of non-negativity since $\text{CASO}(u_x,u_y)\in[0,mn/(m+4)]$, in particular $\text{CAS}O(u_x,u_x)\in[n/3,mn/(m+4)]$, and symmetry, i.e., $\text{CASO}(u_x,u_y)=\text{CASO}(u_y,u_x)$, although it does not guarantee the property of triangle inequality. Therefore, CASO is a nonmetric similarity measure.

Algorithm 1 Coupled Attribute Similarity for Objects

Data: Data set $S_{m\times n}$ with m objects and n attributes, object $u_x, u_y (x,y\in[1,m])$, and weight $\alpha=(\alpha_k)_{1\times n}$.
Result: Coupled Similarity for objects $CASO(u_x,u_y)$.
begin
  //Compute pairwise similarity for any two values of the same attribute.
  for *attribute* $a_j,\ j=1:n$ **do**
    for *every value pair* $(v_j^x,v_j^y\in[1,|V_j|])$ **do**
      $U_1\longleftarrow\{i|v_j^i==v_j^x\}$, $U_2\longleftarrow\{i|v_j^i==v_j^y\}$;
      //Compute intra-coupled similarity for two values v_j^x and v_j^y.
      $\delta_j^{Ia}(v_j^x,v_j^y)=$
      $(|U_1||U_2|)/(|U_1|+|U_2|+|U_1||U_2|)$;
      //Compute coupled similarity for two attribute values v_j^x and v_j^y.
      $\delta_j^A(v_j^x,v_j^y,\{V_k\}_{k=1}^n)\longleftarrow$
      $\delta_j^{Ia}(v_j^x,v_j^y)\cdot IeASV(v_j^x,v_j^y,\{V_k\}_{k\neq j})$;
  //Compute coupled similarity between two objects u_x and u_y.
  $CASO(u_x,u_y)\longleftarrow sum(\delta_j^A(v_j^x,v_j^y,\{V_k\}_{k=1}^n))$;
  end
end
Function $IeASV(v_j^x,v_j^y,\{V_k\}_{k\neq j})$
begin
  //Compute inter-coupled similarity for two attribute values v_j^x and v_j^y.
  for *attribute* $(k=1:n)\wedge(k\neq j)$ **do**
    $\{v_k^z\}_{z\in U_3}\longleftarrow\{v_k^x\}_{x\in U_1}\cap\{v_k^y\}_{y\in U_2}$;
    for *intersection* $z=U_3(1):U_3(|U_3|)$ **do**
      $U_0\longleftarrow\{i|v_k^i==v_k^z\}$;
      $ICP_x\longleftarrow|U_0\cap U_1|/|U_1|$;
      $ICP_y\longleftarrow|U_0\cap U_2|/|U_2|$;
      $Min_{(x,y)}\longleftarrow min(ICP_x,ICP_y)$;
    //Compute IRSI for v_j^x and v_j^y.
    $\delta_{j|k}^I(v_j^x,v_j^y,V_k)=sum(Min_{(x,y)})$;
  $\delta_j^{Ie}(v_j^x,v_j^y,\{V_k\}_{k\neq j})=sum[\alpha(k)\times\delta_{j|k}^I(v_j^x,v_j^y,V_k)]$;
  return $\delta_j^{Ie}(v_j^x,v_j^y,\{V_k\}_{k\neq j})$;
end

We then design an algorithm CASO(), given in Algorithm 1, to compute the coupled object similarity with IRSI (i.e., the best inter-coupled relative similarity candidate). The whole process of this algorithm is summarized as follows: 1) Compute the IaASV for values v_j^x and v_j^y of attribute a_j (Line 1); 2) Compute the IeASV for attribute values v_j^x and v_j^y based on IRSI (Line 1–Line 1); 3) Compute the CASV for attribute values v_j^x and v_j^y (Line 1); and 4) Compute the CASO for objects u_x and u_y (Line 1).

Before the similarity calculation is performed, some data preprocessing is conducted to enable this algorithm. In detail, all the categories of each attribute need to be encoded as numberings, starting at one and increasing to the maximum,

which is the respective number of attribute values. To reduce unnecessary iterations in Line 1, pairwise CASV similarity for any two values of the same attribute, rather than the only two values involved of each attribute, is precalculated for reuse when computing the object similarity. Explicitly, this pseudocode also embodies the fact that the computational complexity for IRSI is indeed $O(n^2R^3)$. However, it might not be very attractive for extremely large data sets with attributes that take too many values. Thus, we are working on strategies of attribute reduction to effectively reduce the number of coupled attributes.

VIII. Experiments and Evaluation

In this section, extensive experiments are performed on several UCI and bibliographic data sets to show the effectiveness of our proposed coupled similarity measures. For simplicity, we assign the weight vector $\alpha = (\alpha_k)_{1\times n}$ with values $\alpha(k) = 1/(n-1)$ in Definition 5.4.

In this part of our experiments, we focus on comparing our novel coupled attribute dissimilarity for objects (CADO) induced from CASO with existing categorical dissimilarity measures. Four independent groups of experiments are conducted with extensive data sets based on machine learning applications. In the following, we evaluate the CADO, which is derived from (15):

$$CADO(u_x, u_y) = \sum_{j=1}^{n} h_1(\delta_j^{Ia}(v_j^x, v_j^y)) \cdot h_2(\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j})) \quad (16)$$

where $h_1(t)$ and $h_2(t)$ are decreasing functions. Based on intra-coupled and inter-coupled similarities, $h_1(t)$ and $h_2(t)$ can be flexibly chosen to build dissimilarity measures according to specific requirements. In terms of the capability of revealing the data relationship, the better the induced dissimilarity, the better is its similarity.

Here, we consider $h_1(t) = 1/t - 1$ and $h_2(t) = 1 - t$ to reflect the complementarity between similarity and dissimilarity measures, since they are both decreasing functions of t. The rationale behind these two functions is as follows. The first conversion corresponds to the improved simple matching dissimilarity (SMD) with frequency [5], if only 0 and 1 are assigned to δ_j^{Ie} (i.e., SMD [23]: dissimilarity 0 for identical values, and otherwise 1). The second transformation guarantees the consistency of CADO with the dissimilarity measure ADD [12], when a constant is fixed for δ_j^{Ia}. In addition, $h_1(t) = 1/t - 1$ is also consistent with the converted measures proposed in [11]; $h_2(t) = 1 - t$ follows the way of converting OF to OFD [10] as well, presented in the next section. Both these functions are designed to include existing classical measures as special cases of our proposed coupled similarity. The detailed specialization to the improved SMD and the ADD are explained in Section IX.

A. Data Structure Analysis

This section performs experiments to explicitly specify the internal structures for the labeled data. Clusterings are normally evaluated by assigning the best score to the algorithm that produces clusters with highest similarity within a cluster and lowest similarity between clusters based on a certain similarity measure. We work in a different way, in which similarity measures are evaluated with clustering criteria and given labels. In this way, a better cluster structure can be clarified with a better similarity measure in terms of the clustering internal descriptors, such as sum-square, Davies–Bouldin index (DBI) [24], and Dunn index (DI) [25].

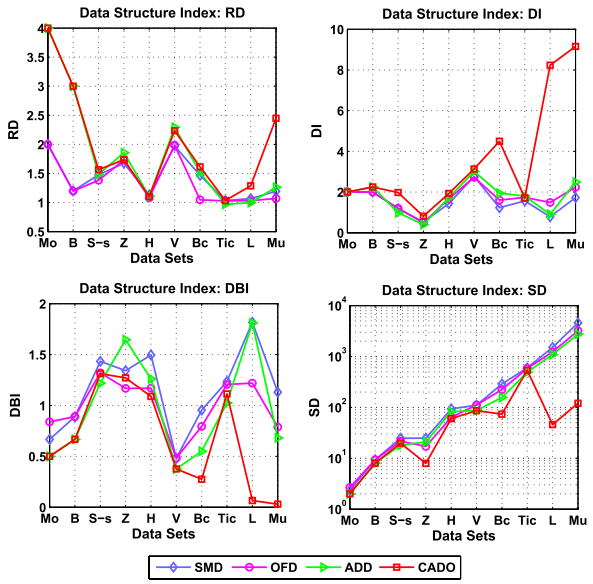

Fig. 2. Data structure index comparison.

To reflect the data cluster structure more clearly, the induced dissimilarity metrics are evaluated by four descriptors. 1) Relative dissimilarity (RD). 2) DBI. 3) DI. 4) Sum-dissimilarity (SD). In detail, RD is the ratio of average inter-cluster dissimilarity upon average intra-cluster dissimilarity for all cluster labels. SD is the sum of object dissimilarities within all the clusters. Since internal criteria seek clusters with high intra-cluster similarity and low inter-cluster similarity, dissimilarity metrics that produce clusters with high RD or DI and low DBI or SD are more desirable.

Four object dissimilarity metrics are considered: 1) SMD [5] (i.e., Hamming distance [23]); 2) OF dissimilarity (OFD) [10]; 3) ADD introduced in [12]; and 4) our proposed CADO. The SMD is a simple, well-known measure for categorical data, while OFD considers matching in terms of attribute value frequency distribution, both formalized as the sum of value dissimilarities for all the attributes. Further, value dissimilarities $D_j^{\text{SMD}} = D_j^{\text{OFD}} = 0$ if $v_j^x = x_j^y$, otherwise they equal 1 and $1 - \left[1 + log(m/|G_j(\{v_j^x\})|) \cdot log(m/|G_j(\{v_j^y\})|)\right]^{-1}$ for SMD and OFD, respectively. The dissimilarity measure ADD, derived from (15) with the worst inter-coupled relative similarity candidate IRSP, considers the sum of inter-coupled interactions between all the corresponding attribute values. These three measures only concern local pictures, while CADO is globally formalized based on (16).

The cluster structures produced by the above four dissimilarity metrics are then analyzed on 10 data sets in different scales. The results after dissimilarity normalization are shown

TABLE XI
CLUSTERING EVALUATION ON SIX DATA SETS

	Data Set ($\|U\|$)	*KM*				*SC*			
		SMD	*OFD*	*ADD*	*CADO*	*SMD*	*OFD*	*ADD*	*CADO*
AC	Shuttle (15)	0.653	0.672	0.733	**0.787**	0.716	0.733	0.762	**0.844**
	Balloon (20)	0.631	0.686	0.720	**0.760**	0.681	0.705	0.727	**0.757**
	Soybean-small (47)	0.734	0.779	0.796	**0.877**	0.749	0.800	0.761	**0.912**
	Zoo (101)	0.604	0.672	0.680	**0.792**	0.660	0.693	0.706	**0.800**
	Soybean-large (307)	0.465	0.477	0.498	**0.539**	0.489	0.507	0.527	**0.636**
	BreastCancer (699)	0.784	0.810	0.812	**0.908**	0.852	0.892	0.900	**0.944**
NMI	Shuttle (15)	0.220	0.230	0.301	**0.371**	0.267	0.249	0.330	**0.456**
	Balloon (20)	0.156	0.235	0.330	**0.381**	0.272	0.291	0.286	**0.364**
	Soybean-small (47)	0.732	0.782	0.845	**0.885**	0.764	0.797	0.791	**0.924**
	Zoo (101)	0.681	0.725	0.684	**0.761**	0.704	0.721	0.728	**0.787**
	Soybean-large (307)	0.604	0.617	0.622	**0.657**	0.625	0.635	0.688	**0.725**
	BreastCancer (699)	0.378	0.416	0.433	**0.595**	0.548	0.590	0.640	**0.705**

in Fig. 2, where the x-axis refers to the data sets Movie, Balloon, Soybean-small, Zoo, Hayesroth, Voting, Breast-cancer, Tic, Letter, and Mushroom, respectively. They are ordered according to the number of objects involved (i.e., m) to describe distinct data scales, ranging from 6 to 8124. As discussed previously, larger RD, larger DI, smaller DBI, and smaller SD indicate better characterization of the cluster differentiation capability, which corresponds to a better dissimilarity metric being induced. From Fig. 2, we observe that, with the exception of a few items, the corresponding RD and DI indexes on CADO are mostly the largest ones when compared with those on SMD, OFD, and ADD; while the associated DBI and SD index curves on CADO are mostly below the other three curves. The results show that our proposed CADO is better than SMD and OFD in terms of differentiating objects in distinct clusters. ADD also seems to be slightly better than SMD and OFD in most cases. The degrees of improvement of CADO upon SMD, OFD, and ADD mainly depend on data structure rather than on data scale $|U|(=m)$ alone.

B. Data Clustering Evaluation

To demonstrate the effectiveness of our proposed CADO and CASO in applications, we conduct two groups of experiments: 1) k-modes (KM) and spectral clustering (SC) and 2) ROCK and CROCK. The former compares two classical clustering methods based on four dissimilarity metrics on six data sets. The latter considers the clustering quality of the adapted method CROCK by integrating our proposed CASO with the categorical clustering algorithm ROCK [16].

1) KM and SC: One of the clustering approaches is the KM algorithm [5], designed to cluster categorical data sets. The main idea of KM is to specify the number of clusters k and then to select k initial modes, followed by allocating every object to the nearest mode. The other is a branch of graph-based clustering, i.e., SC [26], which makes use of Laplacian Eigenmaps on a dissimilarity matrix to perform dimensionality reduction for clustering before the k-means algorithm. Below, we aim to compare the performance of CADO (16) against SMD [5], OFD [10], and ADD [12] as used in data cluster analysis for further clustering evaluation.

We consider eight strategies for clustering on six UCI data sets: KM with SMD, KM with OFD, KM with ADD, KM with CADO, and SC with SMD, SC with OFDSC with ADD, SC with CADO. The clustering performance is evaluated by comparing the obtained cluster of each object with that provided by the data label in terms of accuracy (AC) and normalized mutual information (NMI) [27], which are essentially the external criteria compared with the internal criterion analysis in Section VIII-A. The AC $\in [0, 1]$ is a degree of closeness between the obtained clusters and its actual data labels, while NMI $\in [0, 1]$ is a quantity that measures the mutual dependence of two variables: clusters and labels. The larger AC or NMI is, the better the clustering is, and the better the corresponding dissimilarity metric is.

Table XI reports the results on six data sets with different $|U|$, ranging from 15 to 699 in the increasing order. The performance of the aforementioned eight schemes is evaluated on AC and NMI individually. Followed by Laplacian Eigenmaps, the subspace dimensions are determined by the number of labels in SC. For each data, the average performance is computed over 100 tests for KM and SC with distinct start points. Note that the highest measure score of each experimental setting is highlighted in boldface.

As Table XI indicates, the clustering methods with CADO, whether KM or SC, outperform those with SMD, OFD, and ADD on both AC and NMI. In addition, CADO is better than ADD for measuring clustering quality, ADD is in general superior to OFD, OFD performs better than SMD for most cases. These findings are consistent with the results uncovered in Section VIII-A. In addition, Ahmad and Dey [12] also evidenced that their proposed metric ADD outperforms SMD in terms of KM clustering. Thus, we only analyze the performance improvement of CADO upon ADD in details. The reason is that the effect of inter-coupled interaction of categorical attributes is generally stronger than that of intra-coupled interaction, and the consideration of a complete coupling relationship leads to the largest improvement on clustering accuracy since it discloses the implicit whole structure hidden in data.

For KM, the AC improving rate ranges from 5.56% (Balloon) to 16.50% (Zoo), while the NMI improving rate falls within 4.76% (Soybean-s, i.e., Soybean-small) and 37.38% (Breastcancer). With regard to SC, the former rate takes the minimal and maximal ratios as 4.21% (Balloon) and 20.84%

(Soybean-l, i.e., Soybean-large), respectively, however, the latter rate belongs to [5.45% (Soybean-l), 38.12% (Shuttle)]. The AC and NMI evaluate clustering quality from different aspects; generally, they take minimal and maximal ratios on distinct data sets. Statistical analysis, namely the t-test, has been done on AC and NMI, at a 95% significance level. The null hypothesis that CADO is better than ADD in terms of AC and NMI is accepted. Another significant observation is that SC mostly outperforms KM whenever it has the same dissimilarity metric; this is consistent with the finding in [26], indicating that SC very often outperforms k-means for numerical data.

2) ROCK and CROCK: The ROCK, proposed by Guha *et al*. [16], is a robust clustering algorithm for categorical attributes. A link-based similarity measure between two data points is defined based on the neighborhood relation of the two data points, rather than distance or similarity with other data points in the data set.

During the process of choosing neighbors for each data object, Guha *et al*. [16] simply considered the Jaccard coefficient to capture the closeness between each pair of data objects, followed by the determination of neighbors with a user-defined threshold parameter. Their algorithm mainly focuses on the coupled relationship among objects, without any concern for the coupled relationships among attributes and their values. Therefore, we propose to replace the Jaccard coefficient with our proposed coupled nominal similarity CASO and to construct a coupled ROCK (CROCK) algorithm by considering both coupled objects and coupled attribute values. Specifically, we regard two data objects u_x and u_y to be neighbors if $\text{CASO}(u_x, u_y)/n \geq \theta$, instead of $|u_x \cap u_y|/|u_x \cup u_y| \geq \theta$ presented in [16]. The other procedures and functions remain the same as [16].

Below, we experiment with five real-life data sets, i.e., Movie, Hayesroth, SPECT, Voting, and Mushroom, to compare the cluster quality between ROCK and CROCK in terms of three measures: 1) precision (Pr); 2) recall (Re); and 3) specificity (Sp) [2], [17]. As described in [17], the larger these indexes, the better the clustering. The number of runs for each experiment here is set to be 20 to obtain corresponding average results for the evaluation measures, due to the high computational complexity.

Table XII shows the results of both algorithms on quality measures. We choose parameters to obtain the best results, such as $\theta = 0.75$ for Voting. As this table indicates, the adapted CROCK with our proposed CASO outperforms the original ROCK on almost all the evaluation measures. Statistical testing also supports the results on Pr, Re, and Sp, that CROCK performs better than ROCK, at a 95% significance level. Thus, CROCK's quality is verified to be superior to that of ROCK due to the fact that the former considers both the couplings between attributes with their values (through co-occurrence) and between objects (by links).

TABLE XII
CROCK VERSUS ROCK ON UCI DATA SETS

Data Set ($\|U\|$)	*ROCK*			*CROCK*		
	Pr	Re	Sp	Pr	Re	Sp
Movie (6)	0.88	0.75	0.92	**1**	**1**	**1**
Hayesroth (132)	0.43	0.39	0.62	**0.45**	**0.44**	**0.67**
SPECT (267)	**0.73**	0.62	0.57	0.71	**0.76**	**0.62**
Voting (435)	0.88	0.88	0.89	**0.90**	**0.90**	**0.92**
Mushroom (8124)	0.78	0.65	**0.76**	**0.87**	**0.79**	0.74

TABLE XIII
CLUSTERING QUALITIES FOR FIRST AUTHOR

Algorithm	Parameters	Pr	Sp	AC
STIRR	$\mathscr{M} = 9,\ \mathscr{N} = 10$	0.5112	0.2360	0.5278
LIMBO	$\phi = 0.0,\ S = \infty$	0.6176	**0.4891**	0.5375
CLIMBO	$\phi = 0.0,\ S = \infty$	**0.8045**	0.4718	**0.7333**

C. Intra-Attribute Value Clustering

In this part, we present the results of CASO applications to the problem of intra-attribute value clustering. We use the bibliographic data taken from the publicly-accessible bibliographic databases with 720 research papers [14]. Some 190 papers focus on database research, and the remaining 530 papers are written on theoretical computer science and related fields. For each paper, we record the name of the first author, the name of the second author, the name of the conference/journal, and the year of publication. We are interested in clustering the first authors, as well as the conferences/journals.

As mentioned in Section II, STIRR applies an iterative method based on a linear dynamic system to assign and propagate weights on the categorical values [14] to conduct the intra-attribute value clustering, and LIMBO defines a distance between attribute values on the basis of the IB framework to quantify the degree of interchangeability of attribute values within a single attribute to group them [17]. So, we substitute our proposed CADO in (16) for the distance $\delta I(c_i, c_j)$ described by the information loss (i.e., Jensen–Shannon divergence) in [17], and then propose a coupled version of LIMBO, i.e., CLIMBO. The LIMBO reveals that two attribute values are similar if the contexts in which they appear are similar. It is an alternative way to explicate the inter-coupled interactions among different attributes; however, it lacks the consideration of the intra-coupled interactions within each attribute. Thus CADO can be extended to measure the coupled distance between clusters by replacing an object u_i with a cluster c_i, then CLIMBO is naturally induced.

Below, two experiments are conducted to compare these algorithms for the intra-attribute value clustering. The parameters are specified in the second column of Table XIII. For STIRR, $\mathscr{M}$ and $\mathscr{N}$ are the numbers of initial configurations and iterations, respectively. For LIMBO and CLIMBO, ϕ indicates the size bound, S refers to the accuracy bound, and the addition operator is used. Note that the experiments in this part only run 20 times to display the average results, since the algorithms itself is computational costly.

The first experiment is designed to cluster the first authors of the 720 academic papers, and the labels for evaluation are the preknown research fields: 1) database research (190 papers) and 2) theoretical computer science (530 papers). Note that all authors are identified by their last names so that, for instance,

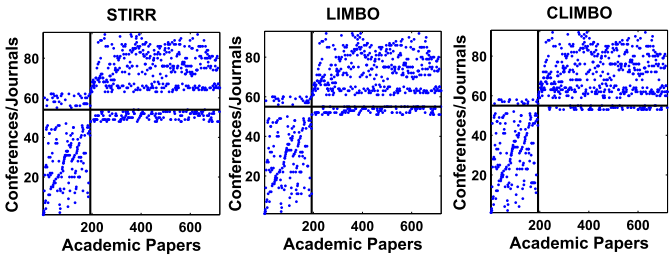

Fig. 3. Clusterings for conferences/journals.

an attribute value Wang actually represents several Wangs taken together. In addition, the second author is regarded as being the same as the first author if the research paper has only one author. These two aspects lead to the overall modest clustering quality. The STIRR, LIMBO, and CLIMBO are compared on the intra-attribute value clustering results of the attribute first author with regard to Pr, Sp [2], and AC [17]. Table XIII shows that CLIMBO is the best in terms of Pr and AC, and is comparable with LIMBO on Sp. All the results on Pr, Sp, and AC are supported by a statistical significant test at a 95% significance level.

We now turn to the problem of clustering the conferences/journals. Fig. 3 shows the clusters produced by STIRR, LIMBO, and CLIMBO. The x-axis represents the academic papers, while the y-axis denotes publishing venues. The thick horizontal line separates the clusters of conferences/journals, and the thick vertical line distinguishes between database research related papers (on the left) and theoretical computer science related papers (on the right). If an author has published a paper in a particular venue, this is represented by a point. From this figure, it is clear that CLIMBO yields the best partition, followed by LIMBO, and STIRR performs worst. However, even the clustering of CLIMBO is slightly mistaken by the conferences/journals between index 50 and 60, which is due to the influence of their co-authors.

The above two experiments therefore reveal that CLIMBO is better than LIMBO and STIRR on the clustering quality of intra-attribute values. In addition, LIMBO can also be clearly observed to outperform STIRR, which is consistent with the conclusion drawn in [17].

In summary: 1) intra-coupled relative similarity measures IRSP, IRSU, IRSJ, and IRSI all present the same learning accuracy, but IRSI is the most efficient, especially for large-scale data; 2) our proposed object dissimilarity metric CADO is better than others, i.e., the traditional SMD, frequency distribution only OFD, and dependency aggregation only ADD, for categorical data in terms of data structure analysis and clustering quality; and 3) the incorporation of CASO or CADO into existing categorical clustering algorithms, such as overlap-based methods (e.g., KM and ROCK), context-based methods (e.g., STIRR), and information-theoretic methods (e.g., LIMBO) can greatly lift their performance.

IX. Discussion

Below, we discuss the potential opportunities triggered by our proposed CASV, CASO, and CADO. The degenerative (first) aspect discusses the degeneration of CADO and CASV with special cases, while the extended (second) aspect focuses on the direct extension of CASO and CADO.

Degenerative Aspect: Many existing similarity measures for attribute values are special cases of our proposed CADO or CASV. On one hand, CADO could degenerate as an intra-attribute-independence measure if frequency functions $G_j(\{v_j^x\}), G_j(\{v_j^y\})$ take a nonzero constant value ξ. In this way, the dissimilarity measure ADD between v_j^x and v_j^y proposed in [12] is exactly $\xi/2 \cdot$ CADO, which considers the interactions between attributes, but lacks the couplings within each attribute. On the other hand, an inter-attribute-independence measure could be produced by considering $\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k=j})$ for IeASV, in which $\delta_{j|j}^{I}(v_j^x, v_j^y, V_j)$ replaces $\delta_{j|k}^{I}(v_j^x, v_j^y, V_k)$ $(k \neq j)$ for IRSI. Such an example is the improved SMD with frequency [5]. In addition, an intra-inter-attribute-independence measure could be obtained by specializing $G_j(\{v_j^x\}) = G_j(\{v_j^y\}) = \xi$ and $\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k=j})$ both, which corresponds to the classical similarity measure SMS and its variants, such as Jaccard coefficients [5]. So, our proposed measures have the capability of generalization to the existing similarity measures which assume independence and partial dependence among attributes.

Extended Aspect: The couplings or relationships between attribute values, attributes, objects, and even clusters should be considered to cater for the interactions among the data. We have already proposed a coupled discretization algorithm CD [28], which concerns both the information dependency and deterministic relationship to disclose the couplings of uncertainty and certainty. A coupled framework for clustering ensembles have been reported in [29] by considering both the relationships within each base clustering and the interactions between distinct base clusterings, in which CASO or CADO is applied. On the other hand, coupled attribute analysis [30] has also been carried out to quantify the relationships among continuous data. In addition, how to appropriately choose the weights α_k for IeASV defined in (13), rather than simply treating them as equal, is in great need of further exploration. Further, we are also working on a flexible way to control the respective importance of IaASV and IeASV by using corresponding weights β and γ, according to the specific data structure. Other data mining and machine learning tasks, e.g. fraud detection [1] and relational learning [31], can also be considered to involve coupled interactions.

X. Conclusion

We have proposed CASO, a novel data-driven coupled attribute similarity measure for objects incorporating both IaASV and inter-coupled attribute similarity for values in unsupervised learning on nominal data. The measure involves both attribute value frequency distribution (intra-coupling) and attribute dependency aggregation (inter-coupling) and the interaction of the two, which captures a global picture of the similarity and has been shown to improve learning accuracy in diverse similarity measures. Theoretical analysis have shown that the inter-coupled relative similarity measure

IRSI significantly outperforms the other options (IRSP, IRSU, and IRSJ) in terms of efficiency, while maintaining equal accuracy. In addition, our derived dissimilarity metric is more general and accurate in capturing the internal structures of the predefined clusters and clustering quality in accordance with intensive empirical results. Very substantial experiments on accuracy have been conducted on the data structure and clustering performance by incorporating the proposed similarity. This has clearly shown that the proposed coupled nominal similarity leads to more accurate learning performance on large scale categorical data sets, supported by statistical analysis. The reason is that our proposed measure is global as a result of effectively integrating different aspects of the similarity.

We are currently applying the CASO measure with IRSI to attribute discretization, clustering ensemble, and other data mining and machine learning tasks. We are working on the assignment of attribute weights, and the flexible engagement of IaASV and IeASV. We are designing the strategies of attribute reduction to fit the extremely large data. In addition, the proposed concepts inter-information function and information conditional probability have the potential to be used in other applications. Flexible dissimilarity measures can also be built on our fundamental similarity building blocks according to different requirements.

APPENDIX

Proof (a):

Theorem 1(a) (Definition 5.1): Intra-coupled attribute similarity for values (IaASV) between values v_j^x and v_j^y of attribute a_j is $\delta_j^{Ia}(v_j^x, v_j^y)$, we have $\delta_j^{Ia} \in [1/3, m/(m+4)]$.

Proof 1: According to Definition 5.1, we have that $1 \le |G_j(\{v_j^x\})|, |G_j(\{v_j^y\})| \le m$ holds, then

$$\begin{aligned}
&\delta_j^{Ia}(v_j^x, v_j^y) \\
&\quad = \frac{|G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|}{|G_j(\{v_j^x\})| + |G_j(v_j^y)| + |G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|} \\
&\quad = \frac{1}{|G_j(\{v_j^y\})|^{-1} + |G_j(\{v_j^x\})|^{-1} + 1} \\
&\quad \le \frac{1}{2\sqrt{|G_j(\{v_j^y\})|^{-1} \cdot |G_j(\{v_j^x\})|^{-1}} + 1}.
\end{aligned}$$

On one hand, $\delta_j^{Ia}(v_j^x, v_j^y)$ is a monotonously increasing function of variables $|G_j(\{v_j^x\})|$ and $|G_j(\{v_j^y\})|$, respectively. Therefore, $\delta_j^{Ia}(v_j^x, v_j^y)$ takes its minimum value $1/3$ when $|G_j(\{v_j^x\})| = |G_j(\{v_j^y\})| = 1$.

On the other hand, because of both $2 \le |G_j(\{v_j^x\})| + |G_j(\{v_j^y\})| \le m$ and the above function property, then $\delta_j^{Ia}(v_j^x, v_j^y)$ takes its maximum value $m/(m+4)$ when $|G_j(\{v_j^x\})| = |G_j(\{v_j^y\})| = m/2$.

Thus, considering both aspects above, we have

$$\delta_j^{Ia}(v_j^x, v_j^y) \in \left[\frac{1}{3}, \frac{m}{(m+4)}\right].$$

Proof (b):

Theorem 2(b) (Definition 5.2): Equations (6) and (7) are equal to each other: $D_{j|L}(v_j^x, v_j^y) = \sum_{l \in L} |P_{l|j}(\{l\}|v_j^x) - P_{l|j}(\{l\}|v_j^y)| = 2 \cdot \max_{L' \subseteq L} |P_{l|j}(L'|v_j^x) - P_{l|j}(L'|v_j^y)|$ holds. [**Note**] This theorem is deduced from a property in probability theory, which is the total variation distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a sigma-algebra $\mathcal{F}$ of the subsets of the sample space Ω is defined via $\delta(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$. For a finite alphabet, we can write $\delta(\mathbb{P}, \mathbb{Q}) = (1/2) \sum_{x \in \Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|$. If we regard $\mathbb{P} = P_{l|j}(\cdot|v_j^x)$ and $\mathbb{Q} = P_{l|j}(\cdot|v_j^y)$, $A = L'$ and $x = l$, then the above theorem holds accordingly.

Proof 2: Assume that $L = \{l_1, l_2, \ldots, l_n\}$ and $L' = \{l_1, l_2, \ldots, l_k\}$ $(k \le n)$, we have

$$\begin{aligned}
F(L') &= 2 \cdot \left| P_{l|j}(L'|v_j^x) - P_{l|j}(L'|v_j^y) \right| \\
&= \left| 2 \cdot \sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^x) - 2 \cdot \sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^y) \right|.
\end{aligned}$$

Since $\sum_{i=1}^{n} P_{l|j}(l_i|v_j^x) = \sum_{i=1}^{n} P_{l|j}(l_i|v_j^y) = 1$ holds, then

$$\begin{aligned}
F(L') &= \left| \left[\sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^x) + 1 - \sum_{i=k+1}^{n} P_{l|j}(\{l_i\}|v_j^x) \right] \right. \\
&\quad \left. - \left[\sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^y) + 1 - \sum_{i=k+1}^{n} P_{l|j}(\{l_i\}|v_j^y) \right] \right| \\
&= \left| \sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^x) - \sum_{i=1}^{k} P_{l|j}(\{l_i\}|v_j^y) \right. \\
&\quad \left. + \sum_{i=k+1}^{n} P_{l|j}(\{l_i\}|v_j^y) - \sum_{i=k+1}^{n} P_{l|j}(\{l_i\}|v_j^x) \right| \\
&= \left| \sum_{i=1}^{k} \left[P_{l|j}(\{l_i\}|v_j^x) - P_{l|j}(\{l_i\}|v_j^y) \right] \right. \\
&\quad \left. + \sum_{i=k+1}^{n} \left[P_{l|j}(\{l_i\}|v_j^y) - P_{l|j}(\{l_i\}|v_j^x) \right] \right| \\
&\le \sum_{i=1}^{k} \left| P_{l|j}(\{l_i\}|v_j^x) - P_{l|j}(\{l_i\}|v_j^y) \right| \\
&\quad + \sum_{i=k+1}^{n} \left| P_{l|j}(\{l_i\}|v_j^y) - P_{l|j}(\{l_i\}|v_j^x) \right| \\
&\le \sum_{i \in 1}^{n} \left| P_{l|j}(\{l_i\}|v_j^x) - P_{l|j}(\{l_i\}|v_j^y) \right| \\
&= \sum_{l \in L} \left| P_{l|j}(\{l\}|v_j^x) - P_{l|j}(\{l\}|v_j^y) \right|.
\end{aligned}$$

If there exists $k > 0$, such that

$$P_{l|j}(\{l_i\}|v_j^x) \ge P_{l|j}(\{l_i\}|v_j^y)$$

holds for $1 \le i \le k < n$ and

$$P_{l|j}(\{l_i\}|v_j^x) < P_{l|j}(\{l_i\}|v_j^y)$$

holds for $k+1 \le i \le n$, then $F(L')$ takes its maximal value: $\sum_{l \in L} |P_{l|j}(\{l\}|v_j^x) - P_{l|j}(\{l\}|v_j^y)|$.

If for all $1 \le i \le k < n$

$$P_{l|j}(\{l_i\}|v_j^x) < P_{l|j}(\{l_i\}|v_j^y)$$

holds, then we have

$$P_{l|j}\big(\{l_i\}|v_j^x\big) \geq P_{l|j}\big(\{l_i\}|v_j^y\big)$$

for $k+1 \leq i \leq n$. Thus, we alternatively consider

$$F(L'') = 2 \cdot \big|P_{l|j}\big(L''|v_j^y\big) - P_{l|j}\big(L''|v_j^x\big)\big|$$

where $L'' = L - L'$. In fact

$$\max_{L' \subseteq L} F(L') = \max_{L'' \subseteq L} F(L'')$$

holds. Similar to the above deduction

$$\begin{aligned}\max_{L' \subseteq L} F(L') &= \max_{L'' \subseteq L} F(L'') \\ &= \sum_{l \in L} \big|P_{l|j}\big(\{l\}|v_j^x\big) - P_{l|j}\big(\{l\}|v_j^y\big)\big|.\end{aligned}$$

The rest special case is that for $1 \leq i \leq n$

$$P_{l|j}\big(\{l_i\}|v_j^x\big) \geq P_{l|j}\big(\{l_i\}|v_j^y\big)$$

holds. This is in fact

$$P_{l|j}\big(\{l_i\}|v_j^x\big) = P_{l|j}\big(\{l_i\}|v_j^y\big)$$

for every possible i, then $F(L') = 0$ takes the maximal value as well (i.e., $\sum_{l \in L} \big|P_{l|j}\big(\{l\}|v_j^x\big) - P_{l|j}\big(\{l\}|v_j^y\big)\big|$).

Therefore, we have

$$\begin{aligned}D_{j|L}\big(v_j^x, v_j^y\big) &= \sum_{l \in L} \big|P_{l|j}\big(\{l\}|v_j^x\big) - P_{l|j}\big(\{l\}|v_j^y\big)\big| \\ &= 2 \cdot \max_{L' \subseteq L} \big|P_{l|j}\big(L'|v_j^x\big) - P_{l|j}\big(L'|v_j^y\big)\big|.\end{aligned}$$

Proof (c):

(Definition 5.2): The conversion is conducted from (7) to (8) via (9): $D_{j|L}\big(v_j^x, v_j^y\big) = 2 \cdot \max_{L' \subseteq L} |P_{l|j}(L'|v_j^x) - P_{l|j}(L'|v_j^y)|$ to $\delta^P_{j|k} = \min_{V'_k \subseteq V_k}\{2 - P_{k|j}(V'_k|v_j^x) - P_{k|j}(\overline{V'_k}|v_j^y)\}$.

Proof: The whole conversion procedural is divided into four steps.

1) The multiplier 2 in $D_{j|L}\big(v_j^x, v_j^y\big)$ is omitted

$$D^{(1)}_{j|L}\big(v_j^x, v_j^y\big) = \max_{L' \subseteq L} \big|P_{l|j}\big(L'|v_j^x\big) - P_{l|j}\big(L'|v_j^y\big)\big|.$$

2) Labels are replaced with other values of a particular attribute a_k

$$D^{(2)}_{j|k}\big(v_j^x, v_j^y\big) = \max_{V'_k \subseteq V_k} \big|P_{k|j}\big(V'_k|v_j^x\big) - P_{k|j}\big(V'_k|v_j^y\big)\big|.$$

3) A complementary set $\overline{V'_k}$ rather than the original one V'_k is concerned for v_j^y in *ICP*, based on $P_{k|j}\big(V'_k|v_j^y\big) = 1 - P_{k|j}\big(\overline{V'_k}|v_j^y\big)$

$$D^{(3)}_{j|k}\big(v_j^x, v_j^y\big) = \max_{V'_k \subseteq V_k} \big|P_{k|j}\big(V'_k|v_j^x\big) + P_{k|j}\big(\overline{V'_k}|v_j^y\big) - 1\big|$$

which is $D'_{j|k}\big(v_j^x, v_j^y\big)$ formalized in (9).

4) Dissimilarity is considered rather than similarity, we use $\delta^P_{j|k} = 1 - D'_{j|k}\big(v_j^x, v_j^y\big)$ for simplicity

$$\begin{aligned}D^{(4.1)}_{j|k}\big(v_j^x, v_j^y\big) &= 1 - D^{(3)}_{j|k}\big(v_j^x, v_j^y\big) \\ &= 1 - \max_{V'_k \subseteq V_k} \big|P_{k|j}\big(V'_k|v_j^x\big) + P_{k|j}\big(\overline{V'_k}|v_j^y\big) - 1\big|.\end{aligned}$$

If $P_{k|j}\big(V'_k|v_j^x\big) + P_{k|j}\big(\overline{V'_k}|v_j^y\big) - 1 \geq 0$, then we have

$$D^{(4.2)}_{j|k}\big(v_j^x, v_j^y\big) = \min_{V'_k \subseteq V_k} \big\{2 - P_{k|j}\big(V'_k|v_j^x\big) - P_{k|j}\big(\overline{V'_k}|v_j^y\big)\big\}$$

according to the fact that

$$1 - \max(|f(x)|) = \min(1 - f(x))$$

for all $f(x) \geq 0$ ($x \in \mathbb{R}$), where $f(x)$ is a function and $\mathbb{R}$ is the real number field.

If $P_{k|j}\big(V'_k|v_j^x\big) + P_{k|j}\big(\overline{V'_k}|v_j^y\big) - 1 < 0$, we alternatively use $V''_k = V_k - V'_k = \overline{V'_k}$. Then we have

$$D^{(4.1')}_{j|k}\big(v_j^x, v_j^y\big) = 1 - \max_{V''_k \subseteq V_k} \big|P_{k|j}\big(V''_k|v_j^x\big) + P_{k|j}\big(\overline{V''_k}|v_j^y\big) - 1\big|.$$

Since $P_{k|j}\big(V''_k|v_j^x\big) = 1 - P_{k|j}\big(V'_k|v_j^x\big)$ and $P_{k|j}\big(\overline{V''_k}|v_j^y\big) = P_{k|j}\big(V'_k|v_j^y\big) = 1 - P_{k|j}\big(\overline{V'_k}|v_j^y\big)$, we have

$$P_{k|j}\big(V''_k|v_j^x\big) + P_{k|j}\big(\overline{V''_k}|v_j^y\big) - 1 > 0.$$

Hence, we have

$$D^{(4.2')}_{j|k}\big(v_j^x, v_j^y\big) = \min_{V''_k \subseteq V_k} \big\{2 - P_{k|j}(V''_k|v_j^x) - P_{k|j}\big(\overline{V''_k}|v_j^y\big)\big\}$$

according to the fact that $1 - \max(|f(x)|) = \min(1 + f(x))$ for all $f(x) \geq 0$ ($x \in \mathbb{R}$), where $f(x)$ is a function and $\mathbb{R}$ is the real number field.

We can see that

$$D^{(4.1)}_{j|k}\big(v_j^x, v_j^y\big) = D^{(4.1')}_{j|k}\big(v_j^x, v_j^y\big).$$

Therefore, we have obtained that

$$\begin{aligned}D^{(4.1)}_{j|k}\big(v_j^x, v_j^y\big) &= D^{(4.1')}_{j|k}\big(v_j^x, v_j^y\big) \\ &= D^{(4.2)}_{j|k}\big(v_j^x, v_j^y\big) = D^{(4.2')}_{j|k}\big(v_j^x, v_j^y\big).\end{aligned}$$

By following the above four steps, we have successfully converted from (7) to (8) via (9): $D_{j|L}(v_j^x, v_j^y)$ to $D^{(4.2)}_{j|k}(v_j^x, v_j^y)$ or $D^{(4.2')}_{j|k}(v_j^x, v_j^y)$ via $D^{(3)}_{j|k}(v_j^x, v_j^y)$ or $D'_{j|k}(v_j^x, v_j^y)$. ■

Proof (d):

Theorem 3(d) (Theorem 6.1): IRSP, IRSU, IRSJ, and IRSI are all equivalent to one another.

Proof: Part (I) IRSP$\Longleftrightarrow$IRSU.

Let V_k^* be the value set of attribute a_k that makes

$$P_{k|j}\big(V'_k|v_j^x\big) + P_{k|j}\big(\overline{V'_k}|v_j^y\big)$$

maximal. Below, we show that for every $v_k \in V_k^*$

$$P_{k|j}\big(\{v_k\}|v_j^x\big) \geq P_{k|j}\big(\{v_k\}|v_j^y\big)$$

holds. If there exists v_k^z ($\in V_k^*$) satisfying

$$P_{k|j}\big(\{v_k^z\}|v_j^x\big) < P_{k|j}\big(\{v_k^z\}|v_j^y\big)$$

then set $V_k^{**} = V_k^* \backslash \{v_k^z\}$, $\overline{V_k^{**}} = \overline{V_k^*} \cup \{v_k^z\}$, it directly follows that:

$$P_{k|j}\big(V_k^{**}|v_j^x\big) + P_{k|j}\big(\overline{V_k^{**}}|v_j^y\big) > P_{k|j}\big(V_k^*|v_j^x\big) + P_{k|j}\big(\overline{V_k^*}|v_j^y\big).$$

This results in the contradiction between V_k^{**} and V_k^* because of the maximal assumption of V_k^*.

Similarly, for any $v_k \in \overline{V_k^*}$

$$P_{k|j}\left(\{v_k\}|v_j^x\right) \le P_{k|j}\left(\{v_k\}|v_j^y\right)$$

holds. Hence

$$\begin{aligned}
\delta_{j|k}^P\left(v_j^x, v_j^y\right) &= \min_{V_k' \subseteq V_k} \left\{2 - P_{k|j}\left(V_k'|v_j^x\right) - P_{k|j}\left(\overline{V_k'}|v_j^y\right)\right\} \\
&= 2 - \max_{V_k' \subseteq V_k} \left\{P_{k|j}\left(V_k'|v_j^x\right) + P_{k|j}\left(\overline{V_k'}|v_j^y\right)\right\} \\
&= 2 - \left[P_{k|j}\left(V_k^*|v_j^x\right) + P_{k|j}\left(\overline{V_k^*}|v_j^y\right)\right] \\
&= 2 - \left[\sum_{v_k \in V_k^*} P_{k|j}\left(\{v_k\}|v_j^x\right) + \sum_{v_k \in \overline{V_k^*}} P_{k|j}\left(\{v_k\}|v_j^y\right)\right] \\
&= 2 - \left[\sum_{v_k \in V_k^*} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right. \\
&\qquad \left. + \sum_{v_k \in \overline{V_k^*}} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right] \\
&= 2 - \sum_{v_k \in V_k} \max\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\} \\
&= \delta_{j|k}^U\left(v_j^x, v_j^y\right).
\end{aligned}$$

Part (II) IRSU$\Longleftrightarrow$IRSJ.

Note that in the following Parts (II) and (III), $v_k \in v_j^x \backslash v_j^y$ and $v_k \in v_j^y \backslash v_j^x$ are the abbreviated forms for $v_k \in \varphi_{j\to k}(v_j^x)\backslash\varphi_{j\to k}(v_j^y)$ and $v_k \in \varphi_{j\to k}(v_j^y)\backslash\varphi_{j\to k}(v_j^x)$, respectively.

Given $v_k \notin \varphi_{j\to k}\left(v_j^x\right) \cup \varphi_{j\to k}\left(v_j^y\right)$, that is

$$v_k \notin \varphi_{j\to k}\left(v_j^x\right) \quad \text{and} \quad v_k \notin \varphi_{j\to k}\left(v_j^y\right).$$

If $v_k \notin \varphi_{j\to k}\left(v_j^x\right)$, we then have

$$g_k^*(\{v_k\}) \bigcap g_j(v_j^x) = \varnothing$$

so, $P_{k|j}\left(\{v_k\}|v_j^x\right) = 0$. Similarly, $P_{k|j}\left(\{v_k\}|v_j^y\right) = 0$. Therefore

$$\begin{aligned}
\delta_{j|k}^U\left(v_j^x, v_j^y\right) &= 2 - \sum_{v_k \in V_k} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} \\
&= 2 - \left[\sum_{v_k \in \bigcup} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right. \\
&\qquad \left. + \sum_{v_k \notin \bigcup} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right] \\
&= 2 - \sum_{v_k \in \bigcup} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} \\
&= \delta_{j|k}^J\left(v_j^x, v_j^y\right).
\end{aligned}$$

Part (III) IRSJ$\Longleftrightarrow$IRSI.

If $v_k \in \varphi_{j\to k}\left(v_j^x\right)\backslash\varphi_{j\to k}\left(v_j^y\right)$, then $P_{k|j}\left(\{v_k\}|v_j^y\right) = 0$. Accordingly, we have

$$\max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} = P_{k|j}\left(\{v_k\}|v_j^x\right).$$

Similarly, if $v_k \in \varphi_{j\to k}\left(v_j^y\right)\backslash\varphi_{j\to k}\left(v_j^x\right)$, it indicates

$$\max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} = P_{k|j}\left(\{v_k\}|v_j^y\right).$$

Therefore, we have

$$\begin{aligned}
&\delta_{j|k}^J\left(v_j^x, v_j^y\right) \\
&= 2 - \sum_{v_k \in \bigcup} \max\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\} \\
&= 2 - \left[\sum_{v_k \in v_j^x \backslash v_j^y} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right. \\
&\qquad + \sum_{v_k \in v_j^y \backslash v_j^x} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} \\
&\qquad \left. + \sum_{v_k \in \bigcap} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right] \\
&= 2 - \left[1 - \sum_{v_k \in \bigcap} P_{k|j}\left(\{v_k\}|v_j^x\right) + 1 - \sum_{v_k \in \bigcap} P_{k|j}\left(\{v_k\}|v_j^y\right)\right. \\
&\qquad \left. + \sum_{v_k \in \bigcap} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\}\right] \\
&= \sum_{v_k \in \bigcap} \left[P_{k|j}\left(\{v_k\}|v_j^x\right) + P_{k|j}\left(\{v_k\}|v_j^y\right)\right] \\
&\qquad - \sum_{v_k \in \bigcap} \max\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} \\
&= \sum_{v_k \in \bigcap} \min\left\{P_{k|j}\left(\{v_k\}|v_j^x\right), P_{k|j}\left(\{v_k\}|v_j^y\right)\right\} = \delta_{j|k}^I\left(v_j^x, v_j^y\right).
\end{aligned}$$

■

ACKNOWLEDGMENT

The authors would like to thank Mr. Xin Cheng and Mr. Zhong She for their assistance in coding on the algorithms: ROCK, LIMBO, and STIRR. We would also like to thank the associate editor and all the anonymous reviewers for their invaluable comments on our paper.

REFERENCES

[1] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1378–1392, Aug. 2012.

[2] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira, Jr, "Word co-occurrence features for text classification," *Inform. Syst.*, vol. 36, no. 5, pp. 843–858, 2011.

[3] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from Flickr groups using fast kernel machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2177–2188, Nov. 2012.

[4] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.

[5] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA, USA: SIAM, 2007.

[6] G. Das and H. Mannila, "Context-based similarity measures for categorical databases," in *Proc. 4th Eur. Conf. Principles Pract. Knowl. Discovery Databases*, Lyon, France, Sep. 2000, pp. 201–210.

[7] T. Li, M. Ogihara, and S. Ma, "On combining multiple clusterings: An overview and a new perspective," *Appl. Intell.*, vol. 33, no. 2, pp. 207–219, 2009.

[8] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J. Artif. Intell. Res.*, vol. 6, no. 1, pp. 1–34, 1997.

[9] S. Cost and S. Salzberg, "A weighted nearest neighbor algorithm for learning with symbolic features," *Mach. Learn.*, vol. 10, no. 1, pp. 57–78, 1993.

[10] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *Proc. SIAM Int. Conf. Data Mining*, Atlanta, GA, USA, Apr. 2008, pp. 243–254.

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

[12] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.

[13] L. A. Ribeiro and T. Harder, "Generalizing prefix filtering to improve set similarity joins," *Inform. Syst.*, vol. 36, no. 1, pp. 62–78, 2011.

[14] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering categorical data: An approach based on dynamical systems," *VLDB J.*, vol. 8, no. 3, pp. 222–236, 2000.

[15] M. E. Houle, V. Oria, and U. Qasim, "Active caching for similarity queries based on shared-neighbor information," in *Proc. 19th Int. Conf. Inform. Knowl. Manage.*, Oct. 2010, pp. 669–678.

[16] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Inform. Syst.*, vol. 25, no. 5, pp. 345–366, 2000.

[17] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "LIMBO: Scalable clustering of categorical data," in *Proc. 9th Int. Conf. Extending Database Technol.*, Heraklion, Greece, Mar. 2004, pp. 123–146.

[18] D. Barbará, J. Couto, and Y. Li, "COOLCAT: An entropy-based algorithm for categorical clustering," in *Proc. 11th Int. Conf. Inform. Knowl. Manage.*, McLean, VA, USA, Nov. 2002, pp. 582–589.

[19] Y. Yang, X. Guan, and J. You, "CLOPE: A fast and effective clustering algorithm for transactional data," in *Proc. 8th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Edmonton, AB, Canada, Jul. 2002, pp. 682–687.

[20] M. J. Zaki, M. Peters, I. Assent, and T. Seidl, "Clicks: An effective algorithm for mining subspace clusters in categorical datasets," in *Proc. 11th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, Aug. 2005, pp. 736–742.

[21] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Statist. Rev.*, vol. 70, no. 3, pp. 419–435, 2002.

[22] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS—Clustering categorical data using summaries," in *Proc. 5th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, Aug. 1999, pp. 73–83.

[23] K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn, "Improved iris recognition through fusion of Hamming distance and fragile bit distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2465–2476, Dec. 2011.

[24] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[25] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974.

[26] U. Von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[27] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 1624–1637, Dec. 2005.

[28] C. Wang, M. Wang, Z. She, and L. Cao, "CD: A coupled discretization algorithm," in *Proc. 16th Pacific-Asia Conf. Knowl. Discovery Data Mining*, Kuala Lumpur, Malaysia, May 2012, pp. 407–418.

[29] C. Wang, Z. She, and L. Cao, "Coupled clustering ensemble: Incorporating coupling relationships both between base clusterings and objects," in *Proc. 29th Int. Conf. Data Eng.*, Brisbane, Australia, Apr. 2013, pp. 374–385.

[30] C. Wang, Z. She, and L. Cao, "Coupled attribute analysis on numerical data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, Aug. 2013, pp. 1736–1742.

[31] L. Getoor and B. E. Taskar, *Introduction to Statistical Relational Learning*. Cambridge, MA, USA: MIT Press, 2007.

Can Wang received the B.Sc. and M.Sc. degrees from the Faculty of Mathematics and Statistics, Wuhan University, Wuhan, China, in 2007 and 2009, respectively, and the Ph.D. degree in computing sciences from the Advanced Analytics Institute, University of Technology, Sydney, NSW, Australia, in 2013.

She is currently a Post-Doctoral Fellow with the Commonwealth Scientific and Industrial Research Organisation, Australia. Her current research interests include behavior analytics, data mining, machine learning, and knowledge representation.

Xiangjun Dong received the Ph.D. degree in computing sciences from the Beijing Institute of Technology, Beijing, China.

He is currently a Professor with the School of Information, Qilu University of Technology, Jinan, China. His current research interests include data mining, machine learning, data integration, and database techniques.

Fei Zhou received the B.Eng. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree from the Department of Electronics Engineering, Tsinghua University, Beijing, China, in 2013.

He has been a Post-Doctoral Fellow with the Shenzhen Graduate School, Tsinghua University, since 2013. His current research interests include image processing and pattern recognition in video surveillance, image superresolution, image interpolation, image quality assessment, and object tracking.

Longbing Cao (SM'06) received the Ph.D. degrees in pattern recognition and intelligent systems, and computing sciences.

He is currently a Professor with the University of Technology, Sydney, NSW, Australia, where he is the Founding Director of the Advanced Analytics Institute, and is the Data Mining Research Leader of the Australian Capital Markets Cooperative Research Center, Sydney. His current research interests include big data analytics, data mining, machine learning, behavior informatics, complex intelligent systems, agent mining, and their applications.

Chi-Hung Chi received the Ph.D. degree from Purdue University, West Lafayette, IN, USA.

He is currently a Science Leader with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. Before joining CSIRO in 2012, he was involved in the industry and universities for more than 20 years. He has authored more than 200 papers in international conferences and journals, and holds six U.S. patents. His current research interests include service engineering, cloud computing, big data analytics, social networking, and behavior informatics.