Coupled Attribute Similarity Learning on Categorical Data

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID:	TNNLS-2013-P-1599
Manuscript Type:	Paper
Date Submitted by the Author:	06-Apr-2013
Complete List of Authors:	Wang, Can; Advanced Analytics Institute, University of Technology, Sydney Cao, Longbing; Advanced Analytics Institute, University of Technology, Sydney
Keywords:	Coupled Behavior Analysis, Data Mining, Machine Learning, Unsupervised Learning

Coupled Attribute Similarity Learning on Categorical Data

Can Wang, *Student Member, IEEE,* Longbing Cao, *Senior Member, IEEE*

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, join 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. Substantial experiments on 20 UCI data sets verify the theoretical conclusions. In addition, 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 12 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.

Index Terms—Similarity analysis, coupled attribute similarity, coupled object analysis, unsupervised learning, clustering.

1 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: between clusters, between attributes, between data objects, or 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 [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 to 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 Simple Matching Similarity (*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 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 1, six movie objects are divided into two classes with three nominal attributes: director, actor and genre. The *SMS* measure between directors "*Scorsese*" and "*Coppola*" is 0, but "*Scorsese*" and "*Coppola*" are

- *C. Wang and L. Cao are with the Advanced Analytics Institute, University of Technology, Sydney, Australia. E-mail: see {canwang613, longbing.cao}@gmail.com.*

TABLE 1
An 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

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. *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 6.

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:

– 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.
– We compare the accuracy and efficiency of the four proposed measures for *IeASV* in terms of four relationships: power set, universal set, join set, and intersection set; and obtain the most efficient candidate based on the intersection set (i.e. *IRSI*) from theoretical and experimental aspects.
– A method is proposed to flexibly define the dissimilarity metrics with the proposed similarity building blocks according to specific requirements.
– 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.

The paper is organized as follows. In Section 2, we briefly review the related work. Preliminary definitions are specified in Section 3. Section 4 proposes the framework of the coupled attribute similarity analysis. Section 5 defines the intra-coupled similarity, inter-coupled similarity, and their aggregation. The theoretical analysis is given in Section 6. We describe the *CASO* algorithm in Section 7. The efficiency and effectiveness of *CASO* are empirically studied in Section 8 and a flexible method to define dissimilarity metrics is also developed. Section 9 discusses the coupled nominal similarity with open issues. Finally, we conclude this work in Section 10.

2 Related Work

Some surveys, in particular [5], [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.

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 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 by using functions that operate on the intersection of two

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

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 for 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 put forward the Iterated Contextual Distances algorithm, believing that the attribute, object and sub-relation similarities are inter-dependent [6]. 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.

3 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 $U = \{u_1, \cdots, u_m\}$ is composed of a nonempty finite set of data objects; $A = \{a_1, \cdots, a_n\}$ is a finite set of attributes; $V = \bigcup_{j=1}^{n} V_j$ 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 $f = \bigcup_{j=1}^{n} f_j$, $f_j : U \to V_j (1 \leq j \leq n)$ is an information function which assigns a particular value of attribute a_j to every object. For instance, Table 2 is an information table consisting of six objects $\{u_1, \cdots, u_6\}$ 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 $V_2 = \{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3\}$.

TABLE 2
An Example of Information Table

A U	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	$\mathcal{A}_4$	$\mathcal{B}_2$	$\mathcal{C}_3$

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 2. Below, an information table S is given, and $|\text{set}|$ is the number of elements in a certain set.

Definition 3.1 (SIF): Two **Set Information Functions (SIFs)** are defined as:

$$F_j : 2^U \to 2^{V_j}, \quad F_j(U') = \{f_j(u_x) | u_x \in U'\}, \tag{3.1}$$

$$G_j : 2^{V_j} \to 2^U, \quad G_j(V_j') = \{u_i | f_j(u_i) \in V_j'\}, \tag{3.2}$$

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

These *SIF*s 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 $\mathcal{B}_2$.

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:

$$\varphi_{j \to k} : V_j \to 2^{V_k}, \quad \varphi_{j \to k}(v_j) = F_k(G_j(\{v_j\})). \tag{3.3}$$

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. For example, $\varphi_{2 \to 1}(\mathcal{B}_1) = \{\mathcal{A}_1, \mathcal{A}_2\}$

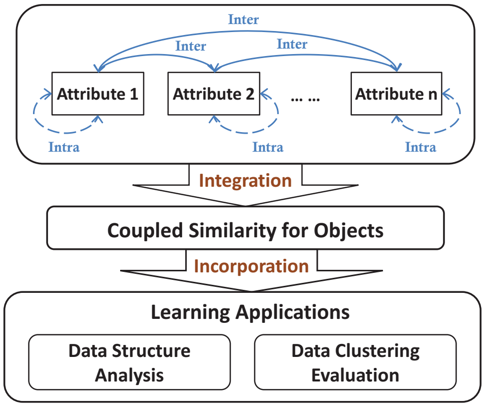

Fig. 1. A framework of coupled attribute similarity analysis, where $\dashleftarrow\dashrightarrow$ indicates intra-coupling and $\longleftrightarrow$ refers to inter-coupling.

TABLE 3
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

specifies that the attribute 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: u_1 and u_2.

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:

$$P_{k|j}(V_k'|v_j) = \frac{|G_k(V_k') \bigcap G_j(\{v_j\})|}{|G_j(\{v_j\})|}. \qquad (3.4)$$

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, $P_{1|2}(\{\mathcal{A}_1\}|\mathcal{B}_1) = 0.5$.

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 3.

4 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 B_1 and B_2 (i.e. values of attribute a_2) concerns both the intra-coupled relationship specified by the repeated times of values B_1 and B_2: 2 and 2, and the inter-coupled interaction triggered by the other two attributes (a_1 and a_3). Next, the coupled interaction is 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.

In the framework described in Fig. 1, the couplings of attributes are revealed via the similarity between attribute values v_j^x and v_j^y of each attribute a_j 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 learning 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.

5 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.

5.1 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, 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:

$$\delta_j^{Ia}(v_j^x, v_j^y) = \frac{|G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|}{|G_j(\{v_j^x\})| + |G_j(v_j^y)| + |G_j(\{v_j^x\})| \cdot |G_j(\{v_j^y\})|}. \tag{5.1}$$

Since $1 \leq |G_j(v_j^x)|, |G_j(v_j^y)| \leq m$ and $2 \leq |G_j(v_j^x)| + |G_j(v_j^y)| \leq m$, then $\delta_j^{Ia} \in [1/3, m/(m+4)]$ is obtained according to Proof (a) in the Appendix. For example, in Table 2, both $\mathcal{B}_1$ and $\mathcal{B}_2$ are observed twice, $\delta_2^{Ia}(\mathcal{B}_1, \mathcal{B}_2) = 0.5$.

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.

5.2 Inter-coupled Interaction

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. *MVDM* takes into account 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 for a specific attribute a_j regarding labels L is:

$$D_{j|L}(v_j^x, v_j^y) = \sum_{l \in L} |P_{l|j}(\{l\}|v_j^x) - P_{l|j}(\{l\}|v_j^y)|, \tag{5.2}$$

where $l (\in L)$ is a label in the information table S. $P_{l|j}$ is the *ICP* defined in (3.4) 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. According to the principle [18] that, for the categorical data distribution, the sum of L1 dissimilarities and twice the total variation dissimilarity are equivalent, we have:

$$D_{j|L}(v_j^x, v_j^y) = 2 \cdot \max_{L' \subseteq L} |P_{l|j}(L'|v_j^x) - P_{l|j}(L'|v_j^y)|. \tag{5.3}$$

The detailed proof on the equivalence of Equations (5.2) and (5.3) is specified by Proof (b) in the Appendix.

In the absence of labels, the above (5.3) is adapted to satisfy our target problem by replacing the class label information with other attribute knowledge to enable unsupervised learning. We regard this interaction between attributes as inter-coupled similarity in terms of the co-occurrence comparisons of *ICP*. The most intuitive variant of (5.3) 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):

$$\delta_{j|k}^P = \min_{V_k' \subseteq V_k} \{2 - P_{k|j}(V_k'|v_j^x) - P_{k|j}(\overline{V_k'}|v_j^y)\}, \tag{5.4}$$

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 (5.4) and (5.3) includes: 1) the multiplier 2 in (5.3) 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

$$D'_{j|k}(v_j^x, v_j^y) = \max_{V_k' \subseteq V_k} |P_{k|j}(V_k'|v_j^x) + P_{k|j}(\overline{V_k'}|v_j^y) - 1| \tag{5.5}$$

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 and relevant proof are provided in Proof (c) in the Appendix. In fact, 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 2, by employing (5.4), 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 attribute a_2 regarding attribute a_1. As shown in Table 4, the set of all attribute values of attribute 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. 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, the universal set concerns all the elements involved, and the join and intersection sets focus on parts of the elements. We start with the power set-based *IRSP*, and will proceed to the universal set-based *IRSU*, the join 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*, *IRSI* based on *IRSP*.

TABLE 4
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

TABLE 5
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 6
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

Definition 5.3 (IRSU, IRSJ, IRSI): The **Inter-coupled Relative Similarity based on Universal Set (IRSU), Join 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:

$$\delta_{j|k}^U = 2 - \sum_{v_k \in V_k} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (5.6)$$

$$\delta_{j|k}^J = 2 - \sum_{v_k \in \bigcup} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (5.7)$$

$$\delta_{j|k}^I = \sum_{v_k \in \bigcap} \min\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (5.8)$$

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

In the above, 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 5, the similarity $\delta_{2|1}^U$ based on *IRSU* is $\delta_{2|1}^U(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}_{i=1}^4) = 2 - 0.5 - 0.5 - 0 - 0.5 = 0.5$. 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 8 times compared with 32 times by *IRSP*, which leads to a substantial improvement in efficiency.

IIF (3.3) is used to further reduce the time cost of *ICP* with two more similarity measures: *IRSJ* (5.7) and *IRSI* (5.8). With (5.7), the calculation of $\delta_{2|1}^J$ is further simplified since $\mathcal{A}_3 \notin \varphi_{2\to 1}(\mathcal{B}_1) \bigcup \varphi_{2\to 1}(\mathcal{B}_2)$. As shown in Table 6, we obtain $\delta_{2|1}^J(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}_{i=1}^4) = 2 - 0.5 - 0.5 - 0.5 = 0.5$, which reveals the fact that it is enough to compute *ICP* with $w \in V_1$ that belongs to $\varphi_{2\to 1}(\mathcal{B}_1) \bigcup \varphi_{2\to 1}(\mathcal{B}_2)$ instead of all the elements in V_1. From this aspect, *IRSJ* further reduces the complexity compared to *IRSU*.

TABLE 7
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 (5.8), the calculation of $\delta_{2|1}^I$ is once again simplified as in Table 7 since only $A_2 \in \varphi_{2\to 1}(\mathcal{B}_1) \bigcap \varphi_{2\to 1}(\mathcal{B}_2)$. Then, we easily get $\delta_{2|1}^I(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}_{i=1}^4) = 0.5$. In this case, it is sufficient to compute *ICP* with $\mathcal{A}_2 \in V_1$ which only belongs to $\varphi_{2\to 1}(\mathcal{B}_1) \bigcap \varphi_{2\to 1}(\mathcal{B}_2)$. It is trivial that the cardinality of intersection $\bigcap$ is no larger than that of join set $\bigcup$. 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*, *IRSI*. In fact, 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 6.

Accordingly, the similarity between the value pair (v_j^x, v_j^y) 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.

Definition 5.4 (IeASV): The **Inter-coupled Attribute Similarity for Values (IeASV)** between attribute values v_j^x and v_j^y of attribute a_j is:

$$\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j}) = \sum_{k=1, k\neq j}^{n} \alpha_k \delta_{j|k}(v_j^x, v_j^y, V_k), \quad (5.9)$$

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

Therefore, $\delta_j^{Ie} \in [0, 1]$. For the parameter α_k, in this paper, we simply assign $\alpha_k = 1/(n-1)$. For example, in Table 2, we then have $\delta_2^{Ie}(\mathcal{B}_1, \mathcal{B}_2, \{V_1, V_3\}) = 0.5 \cdot \delta_{2|1}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{A}_i\}_{i=1}^4) + 0.5 \cdot \delta_{2|3}(\mathcal{B}_1, \mathcal{B}_2, \{\mathcal{C}_i\}_{i=1}^3) = 0.25$ if α_1 and α_3 equal to 0.5.

5.3 Coupled Interaction

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

Definition 5.5 (CASV): The **Coupled Attribute Similarity for Values (CASV)** between attribute values v_j^x and

v_j^y of attribute a_j is:

$$\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) = \delta_j^{Ia}(v_j^x, v_j^y) \cdot \delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k \neq j}), \tag{5.10}$$

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 Equation (5.10), *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; (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 $\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) = \beta \cdot \delta_j^{Ia}(v_j^x, v_j^y) + \gamma \cdot \delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k \neq j})$, where $0 \leq \beta, \gamma \leq 1$ $(\beta + \gamma = 1)$ 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.

Additionally, $\delta_j^A = \delta_j^{Ia} \cdot \delta_j^{Ie} \in [0, m/(m+4)]$ 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 2, 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 set, then $\delta_{Director}^A(Scorsese, Coppola) = \delta_{Director}^A(Coppola, Coppola) = 0.33$, and $\delta_{Director}^A\ (Koster, Coppola) = 0$ while $\delta_{Director}^A\ (Koster, Hitchcock) = 0.25$. 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 1.

In the following theoretical analysis in Section 6, the computational accuracy and complexity of the four inter-coupled relative similarity options are analyzed.

6 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.

1) 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: 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 5.2. Thus, these measures can induce exactly the same computational accuracy in different learning tasks including classification and clustering.

TABLE 8
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

2) 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 8 summarizes the time costs of the four inter-coupled relative similarity measures.

Let $|ICP_{j|k}^{(M)}|$ 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 8, $|ICP_{j|k}^{(P)}| \geq |ICP_{j|k}^{(U)}| \geq |ICP_{j|k}^{(J)}| \geq |ICP_{j|k}^{(I)}|$ holds constantly. It demonstrates the competitive efficiency of *IRSI* compared to the other three measures. In Table 2, 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 $|ICP_{j|k}^{(M)}|$ times. As we have n attributes, the total *ICP* time cost for *CASV* is $2 \cdot |ICP_{j|k}^{(M)}| \cdot (n-1)$ flops per step. The computational complexity for calculating all four options of *CASV* is shown in Table 9.

As indicated in Table 9, 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 *IIF*s, 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 to 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 $R_\cap \leq R_\cup \leq R$. In fact, 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.

TABLE 9
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	$nR(R-1)/2$	$2(n-1)R_\cap$	$O(n^2R^2R)$

7 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 **Coupled Attribute Similarity for Objects (CASO)** between objects u_x and u_y is $CASO(u_x, u_y)$:

$$CASO(u_x, u_y) = \sum_{j=1}^{n} \delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n), \qquad (7.1)$$

where δ_j^A is the *CASV* measure defined in (5.10), 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 *CASV*s with each attribute are summed up for two objects. For example the similarity between u_2 and u_3 in Table 2 is $CASO(u_2, u_3) = \sum_{j=1}^{3} \delta_j^A(v_j^2, v_j^3, \{V_k\}_{k=1}^3) = 0.5 + 0.125 + 0.125 = 0.75$.

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

We then design an algorithm *CASO_IRSI()*, given below, 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 5); (2) Compute the *IeASV* for attribute values v_j^x and v_j^y based on *IRSI* (Line 10 to Line 20); (3) Compute the *CASV* for attribute values v_j^x and v_j^y (Line 6); and (4) Compute the *CASO* for objects u_x and u_y (Line 7).

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 7, pairwise *CASV* similarity for any two values of the same attribute, rather than the only two values involved of each attribute, is pre-calculated for reuse when computing the object similarity. Explicitly, this pseudocode also embodies the fact that the computational complexity for *IRIS* 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.

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 $CASO(u_x, u_y)$.
1  begin
       // Compute pairwise similarity for any
          two values of the same attribute.
2      for attribute $a_j$, $j = 1 : n$ do
3          for every value pair $(v_j^x, v_j^y \in [1, |V_j|])$ do
4              $U_1 \longleftarrow \{i | v_j^i == v_j^x\}$, $U_2 \longleftarrow \{i | v_j^i == v_j^y\}$;
               // Compute intra-coupled similarity
                  for two values $v_j^x$ and $v_j^y$.
5              $\delta_j^{Ia}(v_j^x, v_j^y) = (|U_1||U_2|)/(|U_1| + |U_2| + |U_1||U_2|)$;
               // Compute coupled similarity for
                  two attribute values $v_j^x$ and $v_j^y$.
6              $\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) \longleftarrow$
               $\delta_j^{Ia}(v_j^x, v_j^y) \cdot IeASV(v_j^x, v_j^y, \{V_k\}_{k\neq j})$;
       // Compute coupled similarity between
          two objects $u_x$ and $u_y$.
7      $CASO(u_x, u_y) \longleftarrow sum(\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n))$;
8      end
9  Function $IeASV(v_j^x, v_j^y, \{V_k\}_{k\neq j})$
10 begin
       // Compute inter-coupled similarity for
          two attribute values $v_j^x$ and $v_j^y$.
11     for attribute $(k = 1 : n) \wedge (k \neq j)$ do
12         $\{v_k^z\}_{z\in U_3} \longleftarrow \{v_k^x\}_{x\in U_1} \bigcap \{v_k^y\}_{y\in U_2}$;
13         for intersection $z = U_3(1) : U_3(|U_3|)$ do
14             $U_0 \longleftarrow \{i | v_k^i == v_k^z\}$;
15             $ICP_x \longleftarrow |U_0 \bigcap U_1| / |U_1|$;
16             $ICP_y \longleftarrow |U_0 \bigcap U_2| / |U_2|$;
17             $Min_{(x,y)} \longleftarrow min(ICP_x, ICP_y)$;
           // Compute IRSI for $v_j^x$ and $v_j^y$.
18         $\delta_{j|k}^I(v_j^x, v_j^y, V_k) = sum(Min_{(x,y)})$;
19     $\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j}) = sum[\alpha(k) \times \delta_{j|k}^I(v_j^x, v_j^y, V_k)]$;
20     return $\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k\neq j})$;
```

8 Experiments and Evaluation

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

8.1 Coupled Similarity Comparison

To compare efficiency, experiments are conducted on the inter-coupled relative similarity measures: *IRSP*, *IRSU*, *IRSJ*, and *IRSI*. Experiments are first performed for efficiency comparison, followed by scalability analysis. Note that the time cost of *ICP* is quantified by the calculation counts of *ICP*.

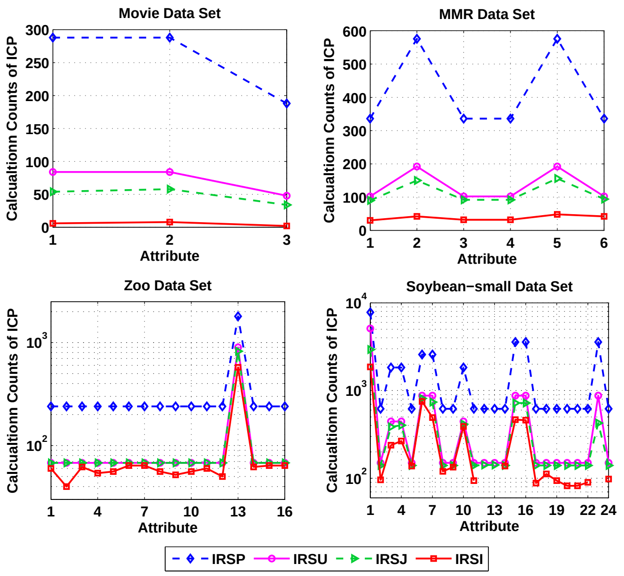

Fig. 2. Complexity on individual attributes.

8.1.1 Efficiency Comparison

The goal in this set of experiments is to show the obvious superiority of *IRSI* compared with the most time-consuming measure *IRSP*. As discussed in Section 6, the computational complexity linearly depends on the time cost of *ICP* with given data. Thus, we consider the comparison of complexity represented by the time cost of *ICP* from the following two aspects.

In terms of a single attribute, the time costs of *ICP* on *Movie* [12], *MMR*, *Soybean-small* and *Zoo* data sets are shown in Fig. 2. We only consider the attributes whose number of values is more than 1, thus, there are only 24 attributes for *Soybean-small* rather than 35. The horizontal axis refers to the ordinal number of nominal attributes, e.g., 1 indicates attribute a_1; while the vertical axis indicates the total time cost (i.e. calculation counts) of *ICP* for all value pairs of each attribute with four options: *IRSP*, *IRSU*, *IRSJ*, *IRSI*. The results show that for any individual attribute, *IRSI* always has the smallest time cost, followed by *IRSJ* and *IRSU*, while *IRSP* is far more time-consuming.

In more detail, we observe that the complexity of *IRSP* for each attribute is around three or four times the size of *IRSU* for these four data sets. Theoretically, this ratio $\xi(P/U)$ can be fixed within an interval based on the given data structure. Suppose we have an information table S with m objects and n attributes. For all the attributes, let $T(=\min_{k=1}^{n}|V_j|)$ and $R(=\max_{k=1}^{n}|V_j|)$ be their minimal and maximal number of values, respectively. Then, for any attribute a_j:

$$\xi_j(P/U) = \frac{|ICP_j^{(P)}|}{|ICP_j^{(U)}|} \in \left[\frac{2^T}{T}, \frac{2^R}{R}\right], \qquad (8.1)$$

where $|ICP_j^{(M)}|$ is the time cost of *ICP* for a_j. Proof (e) in the Appendix supports this statement. For *Zoo*, $T = 2$ and $R = 6$, and the corresponding multiples ξ_j, which range from 2.0 to 3.5, all fall in $[2, 10.7]$.

With respect to all attributes, all the time costs of *ICP* for all the attribute value pairs are considered. Table 10 reports the total time cost of *ICP* with four measures on 12 data sets in terms of relative proportion and direct frequency, where R and n denote the maximal number of attribute values and the number of attributes, respectively, and $|ICP^{(M)}|$ indicates the total time cost of *ICP* for all attributes. Let $\xi(U/P)$ and $\xi(I/J)$ denote the proportion $|ICP^{(U)}|/|ICP^{(P)}|$ and $|ICP^{(I)}|/|ICP^{(J)}|$, respectively. Then $\xi(U/P) \in [R/2^R, T/2^T]$ is deduced according to the proof of Equation (8.1). This property can be checked in Table 10, 27.1% ∈ [25%, 37.5%] for the data set *Hayesroth*.

These results also show that the efficiency advantage of *IRSU* over *IRSP* becomes more obvious when the maximal number of values R becomes larger, i.e., the proportion $\xi(U/P)$ reduces monotonously from 50% to 0.1% when R increases from 2 to 16. However, due to the fact that *IRSJ* and *IRSI* involve the relevant join set and intersection set respectively, the variation tendency of their relative efficiency ratio $\xi(I/J) \in [0, 1]$ mainly depends on the data structure rather than R and n alone. The probability of achieving a smaller ratio $\xi(I/J)$ increases as R grows, since we have more opportunity to obtain an intersection set smaller than a join set. This can be observed in Table 10 by the fact that there is a general decreasing tendency that nevertheless has several disorder ratios.

After fixing R, we consider the variation law for the efficiency of *IRSU* and *IRSI* with the increasing n. It is found that the *ICP* time costs of both measures become greater as n grows. For instance, the calculation frequency of *ICP* for *IRSI* increases from 78 to 4774 when n varies between 4 and 36 with $R = 3$. Similarly, the time costs of the other two options (*IRSU* and *IRSI*) also increase when either n or R goes up. The superiority of *IRSI* becomes more remarkable as the data grows more complicated and bigger compared to the other three metrics. Table 10 further evidences that *IRSI* is the most efficient measure in contrast to the worst measure, *IRSP*.

8.1.2 Scalability Analysis

As we have discussed in Section 6, the complexity for *IRSP* is $O(n^2R^22^R)$, while the other three have equal smaller complexity $O(n^2R^3)$. Here, scalability analysis is explored in terms of n and R separately.

From the perspective of the number of attributes n, the *Soybean-large* data set is considered with 307 objects and 35 attributes. Here, we fix R as 7, and focus on n ranging from 5 to 35 with a step length of 5. In terms of the total time cost of *ICP*, the computational complexity comparisons among four measures (*IRSP*, *IRSU*, *IRSJ* and *IRSI*) are depicted in Fig. 3 (a). The result indicates that the complexity of all these measures keeps increasing when n becomes larger. The acceleration of *IRSP* (from 3328 to 74128) is the greatest by contrast to the slightest acceleration of *IRSI* (from 632 to 15704).

TABLE 10
Complexity Comparison on All Attributes

Data Set	Corral	Voting	Led24	Lense	Tic	Chess	Movie	Hayesroth	Molecular	Solar	Mushroom	Letter
R	2	2	2	3	3	3	4	4	4	7	12	16
T	2	2	2	2	3	2	3	3	4	2	1	10
n	6	16	24	4	9	36	3	4	57	10	22	16
$\xi(U/P)$	50.0%	50.0%	50.0%	46.4%	37.5%	49.4%	27.8%	27.1%	25.0%	20.0%	1.7%	0.1%
$\xi(I/J)$	100%	100%	100%	100%	100%	88.7%	11.0%	100%	99.2%	82.3%	42.5%	48.4%
$\lvert ICP^{(U)}\rvert$	120	960	2208	78	1296	5390	212	468	153216	2544	76020	394294
$\lvert ICP^{(I)}\rvert$	120	960	2208	78	1296	4774	16	468	152022	1998	21736	140434

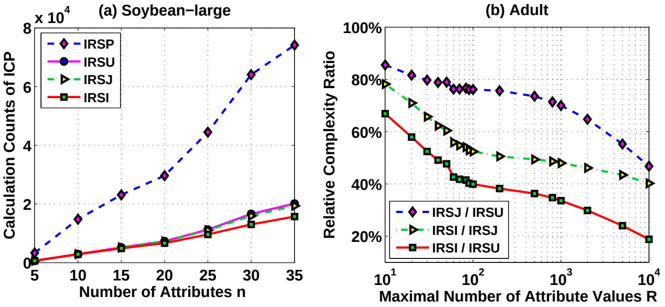

Fig. 3. Scalability on n and R respectively.

Apart from these two, the scalability curves are almost the same for *IRSU* and *IRSI*, though the complexity of *IRSU* is slightly higher than that of *IRSJ* with varied n. Therefore, *IRSI* is the most stable and efficient measure for calculating the intra-coupled relative similarity in terms of the scalability on n.

From the perspective of the maximal number of attribute values R, the variation of R is considered when n is fixed. Here, we take advantage of the *Adult* data set with 30718 objects and 13 attributes chosen. Specifically, the integer attribute "fnlwgt" is discretized into different intervals (from 10 to 10000) to form distinct R ranging from 16 to 10000, since one of the existing categorial attributes "education" already has 16 values. The outcomes are shown in Fig. 3(b), in which the horizontal axis refers to R, and the vertical axis indicates the relative complexity ratios in terms of $\xi(J/U)$, $\xi(I/J)$, and $\xi(I/U)$. From this figure, we observe all the ratios between 10% and 100%, which again verifies the complexity order for these four measures indicated in Section 6. Another issue is that all three curves decrease as R grows, which means the efficiency advantage of *IRSJ* over *IRSU* (from 85.5% to 46.8%), *IRSI* over *IRSJ* (from 78.2% to 40.2%), and *IRSI* over *IRSU* (from 66.9% to 18.8%) all become more and more obvious with the increase of R. The general downturn trend of these ratios comes from the fact that there is a higher probability of obtaining a join set smaller than the whole set, and an intersection set smaller than the join set, with larger R. The same conclusion also holds for the ratio $\xi(U/P)$, but this is due to the monotonously decreasing property of $\xi(U/P)$ on R, which has been proved in Proof (f) in the Appendix. We omit this ratio in Figure 3(b) since the denominator $|ICP^{(P)}|$ becomes exponentially large when R grows, e.g., it equals to 5.12×10^{83} when $R = 500$. Hence, *IRSI* is the least time-consuming intra-coupled similarity with regard to scalability on R.

In summary, all of the above experiment results clearly show that *IRSI* outperforms *IRSU*, *IRSJ* and *IRSI* on computational complexity, no matter how small or large, simple or complicated a data set is. In particular, with the increase in the number of either attributes or attribute values, *IRSI* demonstrates superior efficiency compared to the others. *IRSJ* and *IRSU* follow, with *IRSP* being the most time-consuming, especially for large-scale data.

8.2 Learning Applications

In this part of our experiments, we focus on two levels of algorithmic accuracy comparison:

1) Compare the proposed four intra-coupled measures: *IRSP*, *IRSU*, *IRSJ*, *IRSI*.
2) Compare our novel *Coupled Attribute Dissimilarity for Objects (CADO)* induced from *CASO* with existing categorial dissimilarity measures.

Three 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 (7.1):

$$CADO(u_x, u_y) \qquad (8.2)$$
$$= \sum_{j=1}^{n} h_1[\delta_j^{Ia}(v_j^x, v_j^y)] \cdot h_2[\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k \neq j})],$$

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 *SMD* with frequency [5], if only 0 and 1 are assigned to δ_j^{Ie} (i.e. *SMD* [19]: 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]; $h_2(t) = 1 - t$

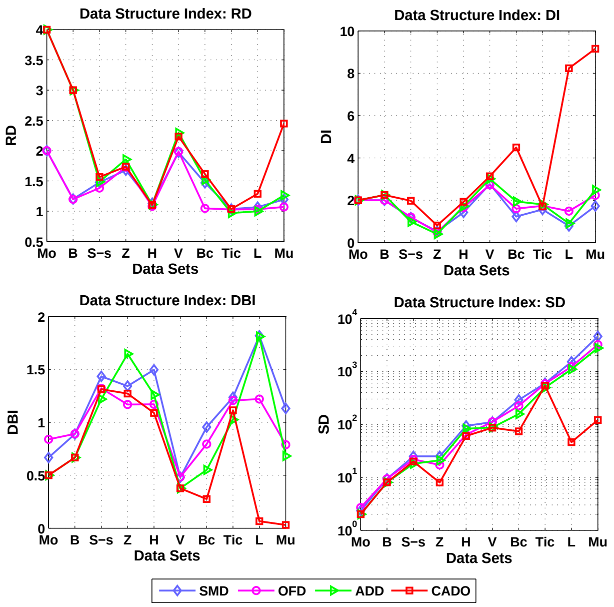

Fig. 4. Data structure index comparison.

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 9.

8.2.1 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) [20], and Dunn Index (DI) [21].

To reflect the data cluster structure more clearly, the induced dissimilarity metrics are evaluated by four descriptors: Relative Dissimilarity (RD), DBI, DI, and 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 here: Simple Matching Dissimilarity [5] (*SMD*, i.e. Hamming distance [19]), Occurrence Frequency Dissimilarity (*OFD*) [10], *ADD* proposed by Ahmad and Dey [12], and *CADO*. *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, attribute value dissimilarities $D_j^{SMD} = D_j^{OFD} = 0$ if $v_j^x = x_j^y$, otherwise they equal 1 and $1 - \left[1 + log\frac{m}{|G_j(\{v_j^x\})|} \cdot log\frac{m}{|G_j(\{v_j^y\})|}\right]^{-1}$ for *SMD* and *OFD*, respectively. The dissimilarity measure *ADD*, derived from (7.1) 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 the local picture, while our proposed *CADO* is globally formalized based on (8.2).

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 in Fig. 4, where the X axis refers to the data sets *Movie*, *Balloon*, *Soybean-small*, *Zoo*, *Hayesroth*, *Voting*, *Breastcancer*, *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. 4, 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.

In constructing *CADO*, all four candidates (*IRSP*, *IRSU*, *IRSJ*, and *IRSI*) are used. Just as we proved in Section 6, all the indexes are the same regardless of exactly what $\delta_{j|k}(x, y)$ refers to, which directly verifies that these four intra-coupled relative similarity measures present equal accuracy.

8.2.2 Clustering Evaluation

To demonstrate the effectiveness of our proposed *CADO* in clustering applications, we compare two classical clustering methods based on two dissimilarity metrics on six data sets. *CADO* is used with the outperforming measure *IRSI*.

One of the clustering approaches is the k-modes (*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. spectral clustering (*SC*) [22], which makes use of Laplacian Eigenmaps on a

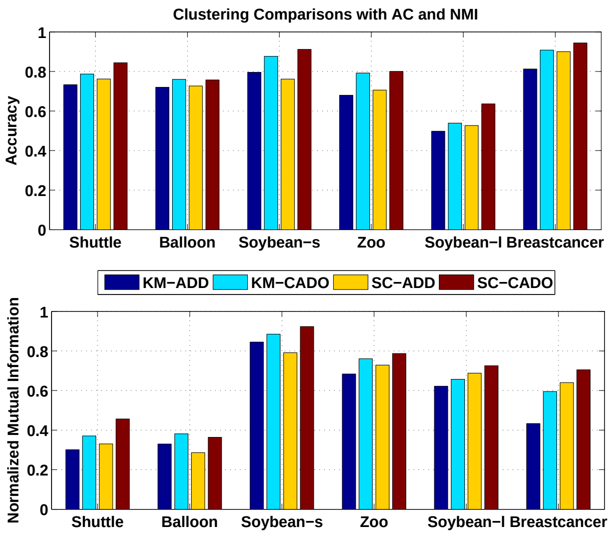

Fig. 5. Clustering evaluation on six data sets.

dissimilarity matrix to perform dimensionality reduction for clustering prior to the k-means algorithm. In respect of attribute dependency aggregation, however, Ahmad and Dey [12] evidenced that their proposed metric *ADD* outperforms *SMD* in terms of *KM* clustering. Thus, we aim to compare the performance of *CADO* (8.2) against *ADD* [12] for further clustering evaluation.

We conduct four groups of experiments on six UCI data sets: *KM* with *ADD*, *KM* with *CADO*, *SC* with *ADD*, and *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) [23], which are essentially the external criteria compared with the internal criterion analysis in Section 8.2.1. 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.

Fig. 5 reports the results on six data sets with different $|U|$, ranging from 15 to 699 in the increasing order. The performance of AC and NMI is individually evaluated for *KM-ADD*, *KM-CADO*, *SC-ADD*, and *SC-CADO*. Followed by Laplacian Eigenmaps, the subspace dimensions are determined by the number of labels in *SC*. For each data set, the average performance is computed over 100 tests for *KM* and *SC* with distinct start points.

As can clearly be seen from Fig. 5, the clustering methods with *CADO*, whether *KM* or *SC*, outperform those with *ADD* on both AC and NMI. That is to say, the dissimilarity metric *CADO* is better than *ADD* for measuring clustering quality. Specifically 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*)$]$. 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 [22], indicating that *SC* very often outperforms k-means for numerical data.

In summary, we have the following findings: 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; 3) the incorporation of *CASO* or *CADO* into existing categorical clustering algorithms such as overlap-based methods (e.g. *k-modes* can greatly lift their performance.

9 Discussions

Below, we discuss the potential opportunities triggered by our proposed *CASV*, *CASO* and *CADO*. The degenerative aspect discusses the degeneration of *CADO* and *CASV* with special cases, while the extended 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 by Ahmad and Dey [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]. Moreover, 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]. Therefore, our proposed measures have the capability of generalization on 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 may naturally induce various coupled tasks in data mining and machine learning, such as data discretization and clustering ensemble. We have already proposed a coupled discretization algorithm *CD* [24], which concerns both the information attribute dependency and deterministic attribute relationship to disclose the couplings of uncertainty and certainty degree. A coupled framework for clustering ensembles have been reported in [25] by considering both the relationships within each base clustering and the interactions between distinct base clusterings, in which *CASO* or *CADO* is applied. In addition, how to appropriately choose the weights α_k for *IeASV* defined in Equation (5.9), 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. The other data mining and machine learning tasks, e.g. fraud detection [1] and relational learning [26], can also be considered to involve coupled interactions.

10 Conclusion and Future Work

We have proposed *CASO*, a novel data-driven coupled attribute similarity measure for objects incorporating both intra-coupled attribute similarity for values 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 and substantial experiments have shown that the inter-coupled relative similarity measure *IRSI* significantly outperforms the other options (*IRSP*, *IRSU* and *IRSJ*) in terms of efficiency, in particular on a large-scale data set having a huge number of attribute values, while maintaining equal accuracy. Moreover, 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 and efficiency have been conducted on single attributes and on all attributes, as well as a scalability test on the number of attributes and the maximal number of attribute values, and 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, efficient and scalable 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 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 extremely large data. We are also considering extending the notion of "coupling" for the similarity of numerical data. Moreover, the proposed concepts *Inter-information Function* and *Information Conditional Probability* have the potential to be used in other applications. One of the clustering criteria, Minimal-Sum-Square, can also be adapted to involve the couplings of categorical data and thus can be improved. Flexible dissimilarity measures can also be built on our fundamental similarity building blocks according to a range of requirements.

References

[1] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE TKDE*, vol. PrePrint, 2011.

[2] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. André Gonçalves, and W. Meira Jr, "Word co-occurrence features for text classification," *Information Systems*, 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 TPAMI*, vol. PrePrints, 2012.

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

[5] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: ASA-SIAM Series on Statistics and Applied Probability, 2007.

[6] G. Das and H. Mannila, "Context-based similarity measures for categorical databases," in *PKDD 2000*, 2000, pp. 201–210.

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

[8] D. Wilson and T. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1–34, 1997.

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

[10] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," in *SDM 2008*, 2008, pp. 243–254.

[11] D. Lin, "An information-theoretic definition of similarity," in *ICML 1998*, 1998, pp. 296–304.

[12] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data and Knowledge Engineering*, vol. 63, pp. 503–527, 2007.

[13] L. A. Ribeiro and T. Harder, "Generalizing prefix filtering to improve set similarity joins," *Information Systems*, 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," *The VLDB Journal*, vol. 8, no. 3, pp. 222–236, 2000.

[15] M. Houle, V. Oria, and U. Qasim, "Active caching for similarity queries based on shared-neighbor information," in *CIKM 2010*, 2010, pp. 669–678.

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

[17] P. Andritsos, P. Tsaparas, R. Miller, and K. Sevcik, "LIMBO: Scalable clustering of categorical data," in *EDBT 2004*, 2004, pp. 123–146.

[18] A. Gibbs and F. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[19] K. Hollingsworth, K. Bowyer, and P. Flynn, "Improved Iris recognition through fusion of Hamming distance and fragile bit distance," *IEEE TPAMI*, vol. 33, no. 12, pp. 2465–2476, 2011.

[20] D. Davies and D. Bouldin, “A cluster separation measure,” *IEEE TPAMI*, vol. 1, no. 2, pp. 224–227, 1979.

[21] J. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Cybernetics and Systems*, vol. 4, no. 1, pp. 95–104, 1974.

[22] U. Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 1–32, 2007.

[23] D. Cai, X. He, and J. Han, “Document clustering using locality preserving indexing,” *IEEE TKDE*, vol. 17, no. 12, pp. 1624–1637, 2005.

[24] C. Wang, M. Wang, Z. She, and L. Cao, “CD: A coupled discretization algorithm,” in *PAKDD 2012*, 2012, pp. 407–418.

[25] C. Wang, Z. She, and L. Cao, “Coupled clustering ensemble: Incorporating coupling relationships both between base clusterings and objects,” in *ICDE 2013*, 2013.

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

APPENDIX

Proof (a)

*Theorem 10.1 (**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 $1 \leq |G_j(\{v_j^x\})|, |G_j(\{v_j^y\})| \leq m$ holds, then

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

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 $2 \leq |G_j(\{v_j^x\})| + |G_j(\{v_j^y\})| \leq m$ 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

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

Proof (b)

*Theorem 10.2 (**b**): [Definition **5.2**]* Two Equations (5.2) and (5.3) 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 $\delta(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \sum_{x \in \Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|$.” 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 $L = \{l_1, l_2, \cdots, l_n\}$ and $L' = \{l_1, l_2, \cdots, l_k\}$ ($k \leq n$), we have

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

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:

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

If there exists $k > 0$, such that

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

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

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

holds for $k+1 \leq i \leq n$, 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 $1 \leq i \leq k < n$,

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

holds, then we have

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

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

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

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

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

holds. Similar to the above deduction,

$$\begin{aligned}\max_{L' \subseteq L} F(L') &= \max_{L'' \subseteq L} F(L'') \\ &= \sum_{l \in L} |P_{l|j}(\{l\}|v^x_j) - P_{l|j}(\{l\}|v^y_j)|.\end{aligned}$$

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

$$P_{l|j}(\{l_i\}|v^x_j) \geq P_{l|j}(\{l_i\}|v^y_j)$$

holds. This is in fact

$$P_{l|j}(\{l_i\}|v^x_j) = P_{l|j}(\{l_i\}|v^y_j)$$

for every possible i, then $F(L') = 0$ takes the maximal value as well (i.e. $\sum_{l \in L} |P_{l|j}(\{l\}|v^x_j) - P_{l|j}(\{l\}|v^y_j)|$). Therefore, we have

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

Proof (c)

[Definition **5.2***]* The conversion is conducted from equations (5.3) to (5.4) via (5.5): "$D_{j|L}(v^x_j, v^y_j) = 2 \cdot \max_{L' \subseteq L} |P_{l|j}(L'|v^x_j) - P_{l|j}(L'|v^y_j)|$" to "$\delta^P_{j|k} = \min_{V'_k \subseteq V_k} \{2 - P_{k|j}(V'_k|v^x_j) - P_{k|j}(\overline{V'_k}|v^y_j)\}$".

Proof 3: The whole conversion procedural is divided into four steps.

(1) The multiplier 2 in $D_{j|L}(v^x_j, v^y_j)$ is omitted:

$$D^{(1)}_{j|L}(v^x_j, v^y_j) = \max_{L' \subseteq L} |P_{l|j}(L'|v^x_j) - P_{l|j}(L'|v^y_j)|.$$

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

$$D^{(2)}_{j|k}(v^x_j, v^y_j) = \max_{V'_k \subseteq V_k} |P_{k|j}(V'_k|v^x_j) - P_{k|j}(V'_k|v^y_j)|.$$

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

$$D^{(3)}_{j|k}(v^x_j, v^y_j) = \max_{V'_k \subseteq V_k} |P_{k|j}(V'_k|v^x_j) + P_{k|j}(\overline{V'_k}|v^y_j) - 1|,$$

which is $D'_{j|k}(v^x_j, v^y_j)$ formalized in equation (5.5).

(4) Dissimilarity is considered rather than similarity, we use $\delta^P_{j|k} = 1 - D'_{j|k}(v^x_j, v^y_j)$ for simplicity:

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

If $P_{k|j}(V'_k|v^x_j) + P_{k|j}(\overline{V'_k}|v^y_j) - 1 \geq 0$, then we have

$$D^{(4.2)}_{j|k}(v^x_j, v^y_j) = \min_{V'_k \subseteq V_k} \{2 - P_{k|j}(V'_k|v^x_j) - P_{k|j}(\overline{V'_k}|v^y_j)\}$$

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 $P_{k|j}(V'_k|v^x_j) + P_{k|j}(\overline{V'_k}|v^y_j) - 1 < 0$, we alternatively use $V''_k = V_k - V'_k = \overline{V'_k}$. Then we have

$$D^{(4.1')}_{j|k}(v^x_j, v^y_j) = 1 - \max_{V''_k \subseteq V_k} |P_{k|j}(V''_k|v^x_j) + P_{k|j}(\overline{V''_k}|v^y_j) - 1|$$

Since $P_{k|j}(V''_k|v^x_j) = 1 - P_{k|j}(V'_k|v^x_j)$ and $P_{k|j}(\overline{V''_k}|v^y_j) = P_{k|j}(V'_k|v^y_j) = 1 - P_{k|j}(\overline{V'_k}|v^y_j)$, we have

$$P_{k|j}(V''_k|v^x_j) + P_{k|j}(\overline{V''_k}|v^y_j) - 1 > 0.$$

Hence, we have

$$D^{(4.2')}_{j|k}(v^x_j, v^y_j) = \min_{V''_k \subseteq V_k} \{2 - P_{k|j}(V''_k|v^x_j) - P_{k|j}(\overline{V''_k}|v^y_j)\}$$

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.

In fact, we can see that

$$D^{(4.1)}_{j|k}(v^x_j, v^y_j) = D^{(4.1')}_{j|k}(v^x_j, v^y_j).$$

Therefore, we have obtained that

$$\begin{aligned}D^{(4.1)}_{j|k}(v^x_j, v^y_j) &= D^{(4.1')}_{j|k}(v^x_j, v^y_j) \\ &= D^{(4.2)}_{j|k}(v^x_j, v^y_j) = D^{(4.2')}_{j|k}(v^x_j, v^y_j).\end{aligned}$$

By following the above four steps, we have successfully converted from (5.3) to (5.4) via (5.5): $D_{j|L}(v^x_j, v^y_j)$ to $D^{(4.2)}_{j|k}(v^x_j, v^y_j)$ or $D^{(4.2')}_{j|k}(v^x_j, v^y_j)$ via $D^{(3)}_{j|k}(v^x_j, v^y_j)$ or $D'_{j|k}(v^x_j, v^y_j)$.

Proof (d)

*Theorem 10.3 (**d**): [Theorem **6.1**]* IRSP, IRSU, IRSJ and IRSI are all equivalent to one another.

Proof 4: Part (I) *IRSP⟺IRSU*
Let V^*_k be the value set of attribute a_k that makes

$$P_{k|j}(V'_k|v^x_j) + P_{k|j}(\overline{V'_k}|v^y_j)$$

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

$$P_{k|j}(\{v_k\}|v^x_j) \geq P_{k|j}(\{v_k\}|v^y_j)$$

holds. In fact, if there exists v^z_k ($\in V^*_k$) satisfying

$$P_{k|j}(\{v^z_k\}|v^x_j) < P_{k|j}(\{v^z_k\}|v^y_j),$$

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

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

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

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

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

holds. Hence,

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

Part (II) *IRSU*$\Longleftrightarrow$*IRSJ*

Note that in the following Part (II) and Part (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 $v_k \notin \varphi_{j\to k}(v_j^x)\bigcup\varphi_{j\to k}(v_j^y)$, that is

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

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

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

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

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

Part (III) *IRSJ*$\Longleftrightarrow$*IRSI*

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

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

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

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

Therefore, we have

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

Thus, *IRSP*, *IRSU*, *IRSJ*, and *IRSI* are all equivalent to one another.

Proof (e)

*Theorem 10.4 (**e**): [Experiment **8.1.1**]* For any attribute a_j, the proportion $\xi_j(P/U) \in [\frac{2^T}{T}, \frac{2^R}{R}]$. For all attributes, the proportion $\xi(P/U) \in [\frac{2^T}{T}, \frac{2^R}{R}]$.

Proof 5: According to Definitions 5.2 and 5.3, and Table 9, we know that

$$\xi_{j|k}(P/U) = \frac{|ICP_{a_{j|k}}^{(P)}|}{|ICP_{a_{j|k}}^{(U)}|} = \frac{2^{|V_k|}}{|V_k|},$$

where $|ICP_{a_{j|k}}^{(P)}|$ and $|ICP_{a_{j|k}}^{(U)}|$ represent the time costs of *ICP* for $\delta_{j|k}^P(v_j^x, v_j^y)$ and $\delta_{j|k}^U(v_j^x, v_j^y)$, respectively. Since $T = \min_{k=1}^n |V_j|$ and $R = \max_{k=1}^n |V_j|$, then $T \leq |V_k| \leq R$ for any set of attribute values V_k. We know $|V_k|$ is a positive integer, so based on Lemma 1 below, the statement

$$\xi_{j|k}(P/U) \in [\frac{2^T}{T}, \frac{2^R}{R}]$$

holds. In addition, we have

$$\begin{aligned}\xi_j(P/U) &= \frac{|ICP_j^{(P)}|}{|ICP_j^{(U)}|} = \frac{\sum_{k\neq j}|ICP^{(P)}_{a_{j|k}}|}{\sum_{k\neq j}|ICP^{(U)}_{a_{j|k}}|},\\ \xi(P/U) &= \frac{|ICP^{(P)}|}{|ICP^{(U)}|} = \frac{\sum_{1\le j\le n}|ICP_j^{(P)}|}{\sum_{1\le j\le n}|ICP_j^{(U)}|}.\end{aligned}$$

Based on Lemma 2 below, we then obtain that

$$\xi_j(P/U) \in \left[\frac{2^T}{T}, \frac{2^R}{R}\right] \text{ and } \xi(P/U) \in \left[\frac{2^T}{T}, \frac{2^R}{R}\right].$$

Lemma 1: If x is a positive integer, then function $q(x) = 2^x/x$ is a monotonically increasing function.

Proof 6: To verify the monotonically increasing property of function $q(x) = 2^x/x$, we only need to look at the derivative of $q(x)$ since $q(x)$ is a continuous function of x, that is

$$q'(x) = \frac{2^x\cdot\ln 2\cdot x - 2^x}{x^2} = \frac{2^x\cdot(\ln 2\cdot x - 1)}{x^2}.$$

If $q'(x) > 0$, then we can guarantee that $q(x)$ is a strictly monotonically increasing function. Here, $q'(x) > 0$ is equivalent to $x > 1/\ln 2$. As x is a positive integer, then $q(x) = 2^x/x$ is a strictly monotonically increasing function when $x \ge 2 > 1/\ln 2$. We also have $q(1) = 2$ when $x = 1$, and $q(2) = 2$ when $x = 2$, so $q(1) \le q(2)$. Thus, $q(x) = 2^x/x$ is a monotonically increasing function when x is a positive integer.

Lemma 2: If $x_1, \cdots, x_n$ are positive integers, where $T = \min_{1\le i\le n} x_i$ and $R = \max_{1\le i\le n} x_i$, then

$$\frac{2^T}{T} \le \frac{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}{x_1 + x_2 + \cdots + x_n} \le \frac{2^R}{R}.$$

Proof 7: Without loss of generality, we assume $1 \le x_1 \le x_2 \le \cdots \le x_n$, then $T = x_1$, $R = x_n$, and the question is to prove

$$\frac{2^{x_1}}{x_1} \le \frac{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}{x_1 + x_2 + \cdots + x_n} \le \frac{2^{x_n}}{x_n}$$

According to Lemma 1, we have for $i = 1, \cdots, n$,

$$\frac{2^{x_1}}{x_1} \le \frac{2^{x_i}}{x_i} \Longleftrightarrow 2^{x_1}\cdot x_i \le 2^{x_i}\cdot x_1.$$

Then, we can naturally obtain that

$$\sum_{i=1}^{n}(2^{x_1}\cdot x_i) \le \sum_{i=1}^{n}(2^{x_i}\cdot x_1),$$

which is equivalent to

$$2^{x_1}\cdot(x_1 + x_2 + \cdots + x_n) \le x_1\cdot(2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}).$$

Therefore, we have

$$\frac{2^{x_1}}{x_1} \le \frac{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}{x_1 + x_2 + \cdots + x_n}.$$

Similarly, we can prove that

$$\frac{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}{x_1 + x_2 + \cdots + x_n} \le \frac{2^{x_n}}{x_n}$$

Thus, the inequality

$$\frac{2^T}{T} \le \frac{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}{x_1 + x_2 + \cdots + x_n} \le \frac{2^R}{R}$$

holds for $T = \min_{1\le i\le n} x_i$ and $R = \max_{1\le i\le n} x_i$.

Proof (f)

*Theorem 10.5 (f): [Experiment **8.1.2**]* Multi-variant function $\xi(U/P)$ is a monotonously decreasing function on the maximal number of values R.

Proof 8: The multi-variant function $\xi(U/P)$ is

$$\begin{aligned}\xi(U/P) &= \frac{|ICP^{(U)}|}{|ICP^{(P)}|} = \frac{\sum_{1\le j\le n}|ICP_j^{(U)}|}{\sum_{1\le j\le n}|ICP_j^{(P)}|}\\ &= \frac{\sum_{1\le j\le n}\sum_{k\neq j}|ICP^{(U)}_{a_{j|k}}|}{\sum_{1\le j\le n}\sum_{k\neq j}|ICP^{(P)}_{a_{j|k}}|}.\end{aligned}$$

Since $|ICP^{(U)}_{a_{j|k}}| = |V_k|$ and $|ICP^{(P)}_{a_{j|k}}| = 2^{|V_k|}$, then the question is to prove that the function

$$F(x_1, x_2, \cdots, x_n) = \frac{x_1 + x_2 + \cdots + x_n}{2^{x_1} + 2^{x_2} + \cdots + 2^{x_n}}$$

is a monotonously decreasing function on x_n, if we assume integers $1 \le x_1 \le x_2 \le \cdots \le x_n$. To verify this, we only need to look at the partial derivative of F on x_n since F is a continuous function of x_n. If $\partial F/\partial x_n \le 0$, then we can say that F is a monotonously decreasing function on x_n. Suppose $M = \sum_{i=1}^{n-1} x_i$ and $N = \sum_{i=1}^{n-1} 2^{x_i}$, then we have

$$\begin{aligned}\frac{\partial F}{\partial x_n} &= \frac{N + 2^{x_n} - (M + x_n)\cdot 2^{x_n}\cdot\ln 2}{(N + 2^{x_n})^2}\\ &= \frac{(N - M\cdot 2^{x_n}\cdot\ln 2) + 2^{x_n}\cdot(1 - x_n\cdot\ln 2)}{(N + 2^{x_n})^2}.\end{aligned}$$

To judge whether $\partial F/\partial x_n \le 0$, we discuss the following four cases:

1) $2 \le x_{n-1} \le x_n$: In this case, according to Lemma 2, we have

$$\begin{aligned}\frac{N}{M} \le \frac{2^{x_{n-1}}}{x_{n-1}} \le 2^{x_n}\cdot\ln 2 &\iff N \le M\cdot 2^{x_n}\cdot\ln 2,\\ \frac{1}{ln2} < 2 \le x_n &\iff 1 < x_n\cdot\ln 2.\end{aligned}$$

Thus, $\partial F/\partial x_n < 0$ holds.

2) $x_{n-1} = 1$ and $2 \le x_n$: In this case, according to Lemma 2, we have

$$\begin{aligned}\frac{N}{M} \le \frac{2^1}{1} < 2^2\cdot\ln 2 \le 2^{x_n}\cdot\ln 2 &\iff N < M\cdot 2^{x_n}\cdot\ln 2,\\ \frac{1}{ln2} < 2 \le x_n &\iff 1 < x_n\cdot\ln 2.\end{aligned}$$

Thus, $\partial F/\partial x_n < 0$ holds.

3) $x_{n-1} = x_n = 1$: In this case, we have

$$F(x_1, \cdots, x_{n-1}, x_n) = F(1, \cdots, 1, 1) = \frac{n}{2n} = \frac{1}{2}.$$

For $x_{n-1} = 1, x_n = 2$, then

$$F(x_1, \cdots, x_{n-1}, x_n) = F(1, \cdots, 1, 2) = \frac{n-1+2}{2(n-1)+4} = \frac{1}{2}.$$

Thus, $F(x_1, \cdots, x_{n-1}, 2) \leq F(x_1, \cdots, x_{n-1}, 1)$.

4) $2 \leq x_{n-1}$ and $x_n = 1$: This case is impossible since we assume $x_{n-1} \leq x_n$.

Therefore, we discover that for both $x_{n-1} = 1$ and $2 \leq x_{n-1}$, F is a monotonously decreasing function on x_n. That is to say, multi-variant function $\xi(U/P)$ is a monotonously decreasing function on the maximal number of values R.

A Brief Description of the Difference between this Journal Version and the Conference Version

Can Wang, Longbing Cao[*]

Dear Editors and Reviewers:

This paper “Coupled Attribute Similarity Learning on Categorical Data” is a revised and enhanced version of the paper “Coupled Nominal Similarity in Unsupervised Learning” previously accepted by the 20th ACM Conference on Information and Knowledge Management (CIKM 2011) held in Glasgow, UK. To improve the originality and the overall quality of the previous conference version, we substantially expand the contents and address the main concerns of the reviewers from CIKM 2011 as follows. Note that all the pages and sections correspond to the new journal version, and the previous conference published version is attached along with this difference letter.

1. Substantial Extension

   The previous conference version has been substantially improved in terms of five aspects: We make a **Clear Differentiation** of our work with the state of the art; We specify an **Explicit Justification** for all the definitions and the connections between them; We provide a **Solid Theoretical Foundation** (i.e., six mathematical proofs) for all the statements in proposed theorems and experimental discoveries; We complement **Plenty of Supporting Experiments** (i.e, another three groups of experiments in addition to the only two parts of experiments in the previous conference version) to verify our proposed conclusions from a variety of perspectives including the statistical analysis; We offer a **Complete Presentation** of this work by additionally including a clear pseudocode of our main algorithm to help other peers and researchers to implement our method easily as well as discussions to analyze our method, and propose open issues with future work.

   The detailed improved and enhanced points of this journal version based on the previous conference version are listed in the following. Even com-

[*] The authors are with the Advanced Analytics Institute, University of Technology, Sydney, Australia. E-mail: see {canwang613, longbing.cao}@gmail.com.

pared to the initial submission version of CIKM 2011, most of the items below are newly addressed.

- We rewrite the Sections “Abstract” (on Page 1), “Introduction” (on Pages 1 and 2), and “Conclusion and Future Work” (on Page 13) to strengthen the motivation and significance of the preliminary version.
- We significantly revise the Section “Related Work” with the aim at providing a thorough review of the relevant research, better and more clearly justifying the differentiation of the current approaches with our work in terms of nominal similarity in unsupervised learning (on Pages 2 and 3).
- We unify all the notations throughout this paper in an easier and more understandable way in Section 3 (on Pages 3 and 4).
- We add a new section to exhibit the whole picture of our proposed *Coupled Attribute Similarity Analysis Framework* in Section 4 (on Page 4).
- We clearly explain the rationale of Definition 5.1 on the *Intra-coupled Attribute Similarity for Values (IaASV)*, motivate why two attribute values should be considered similar if their frequencies are similar, and clarify a small issue of this definition (on Pages 4 and 5).
- We study the derivation process of *IRSP* measure with details in the Appendix (the paragraphs around Definition 5.2: from Equations (5.2) to (5.3), from Equations (5.3) to (5.4) via (5.5)); and we give detailed examples for *IRSP*, *IRSU*, *IRSJ*, and *IRSI* measures in Section 5.2 (on Pages 5 and 6).
- We provide supporting arguments for the rationale of Definition 5.5 on the *Coupled Attribute Similarity for Values (CASV)* from twofold perspectives to explain why using multiplication to couple the measures, together with some discussion for its extension (the first paragraph after Definition 5.5 on Page 7).
- We add a concrete example (Table 8) to analyze the computational complexity in Section 6 (on Page 7).
- We depict the algorithm to compute the *Coupled Attribute Similarity for Objects (CASO)* in the form of pseudocode as Algorithm 1 and analyze it in Section 7 with details (on Page 8).
- We provide a detailed justification on the choices of $h_1(t)$ and $h_2(t)$ for the *Coupled Attribute Dissimilarity for Objects (CADO)* in Equation (8.2) (on Pages 10 and 11).
- We expand the experiment part to newly include Section 8.1.1 about Efficiency Comparisons in terms of a single attribute and all attributes (on Page 9 and in the Appendix), and Section 8.2.1 on Data Structure Analysis (on Page 11). The experiments in Section 8.2.2 are enhanced with statistical analysis (on Pages 11 and 12).

- We add the Section "Discussions" to further reveal the potential and future opportunities in terms of degenerative aspect and extended aspect (on Pages 12 and 13).
- We expand the Section "References" according to the up-to-date research progress emerging on this topic after the completion of our preliminary version and the newly involved measures and algorithms (on Pages 13 and 14).
- We attach the Appendix as supplementary material, including the detailed Proofs: the proof of a statement on Definition 5.1 (on Page 5), the derivation process of *IRSP* measure in Definition 5.2 with two proofs (on Page 5), the proof of Theorem 6.1 (on Page 7), the proof of a statement in Section 8.1.1 (on Page 9), and the proof of a statement in Section 8.1.2 (on Page 10).

2. Addressing CIKM Reviewers' Comments

   We revise and enhance the previous conference version according to the comments from CIKM 2011 reviewers.

   Those comments mainly focus on the lack of comparison to other existing measures. Thus, we design several additional experiments such as Efficiency Comparisons in Section 8.1.1, Data Structure Analysis in Section 8.2.1 (adding current similarity measure *OFD* and *ADD*).

   To improve the technical quality, we give six detailed proofs for the relevant theorems and statements in the Appendix, among them *Theorem (d) 6.1* is the most important; and we explain the rationale of Definition 5.5, as requested by the reviewers.

We are very grateful to the reviewers of CIKM 2011 for their helpful comments and suggestions, which allow us to improve the quality of our work. We are confident that a sufficient amount of new material (roughly 60%) has been added to warrant this journal version.

Coupled Nominal Similarity in Unsupervised Learning

Can Wang[*], Longbing Cao,
Jinjiu Li, Wei Wei, Yuming Ou
Centre for Quantum Computation and
Intelligent Systems
Advanced Analytics Institute
University of Technology, Sydney, Australia
Can.Wang@student.uts.edu.au

Mingchun Wang[†]
School of Science
Tianjin University of Technology and Education
Tianjin, China
mchwang123@163.com

ABSTRACT

The similarity between nominal objects is not straightforward, especially in unsupervised learning. This paper proposes coupled similarity metrics for nominal objects, which consider not only intra-coupled similarity within an attribute (i.e., value frequency distribution) but also inter-coupled similarity between attributes (i.e. feature dependency aggregation). Four metrics are designed to calculate the inter-coupled similarity between two categorical values by considering their relationships with other attributes. The theoretical analysis reveals their equivalent accuracy and superior efficiency based on intersection against others, in particular for large-scale data. Substantial experiments on extensive UCI data sets verify the theoretical conclusions. In addition, experiments of clustering based on the derived dissimilarity metrics show a significant performance improvement.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications–*data mining*

General Terms: Algorithms, Measurement, Performance

Keywords: Similarity measure, Complexity, Accuracy

1. INTRODUCTION

Similarity analysis has been a problem of great practical importance in several domains, including data mining, for decades [8]. By defining certain similarity measures between attribute values, it gauges the strength of the relationship between two data objects: the more two objects resemble each other, the larger the similarity is [7].

When objects are described by numerical features, their similarity measures geometric analogies which reflect the relationship of data values. For instance, the values $10m$ and $12m$ are more similar than $10m$ and $2m$. A variety of similarity metrics have been developed for numerical data, such as Euclidean and Minkowski distances [7]. By contrast, the similarity analysis between records described by nominal variables has received much less attention. Heterogeneous Distances [10] and Modified Value Distance Matrix ($MVDM$) [5], for example, depict the similarity between categorical values in supervised learning. For unlabeled data, only a few works [7], including Simple Matching Similarity (*SMS*, which only uses 0s and 1s to distinguish similarities between distinct and identical categorical values) and Occurrence Frequency [2], discuss the similarity between nominal values. We illustrate the problem with these works and the challenge of analyzing similarity for categorical data below.

Table 1: An Instance of the Movie Database

Movie	Director	Actor	Genre	Class
Godfather II	Scorsese	De Niro	Crime	G_1
Good Fellas	Coppola	De Niro	Crime	G_1
Vertigo	Hitchcock	Stewart	Thriller	G_2
N by NW	Hitchcock	Grant	Thriller	G_2
Bishop's Wife	Koster	Grant	Comedy	G_2
Harvey	Koster	Stewart	Comedy	G_2

Taking the Movie data (Table 1) as an example, six movie objects are divided into two classes with three nominal features: director, actor and genre. The *SMS* measure between directors "*Scorsese*" and "*Coppola*" is 0, but "*Scorsese*" and "*Coppola*" are very similar directors[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 since it belongs to the same class G_2.

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

Thus, we discuss the similarity for categorical values by considering data characteristics. Two attribute values are similar if they present analogous frequency distributions for one attribute [2]; this reflects the intra-coupled similarity within a feature. 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

[*] The first author of this paper for correspondence.

[†] The third author of this paper.

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

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

similar than the latter. To improve the accuracy of intra-coupled similarity, it is believed that the object co-occurrence probabilities of attribute values induced on other features are comparable [1]. To this end, the similarity between directors should also cater for the dependencies on other features such as “actor” and “genre” over all the movie objects, namely, the inter-coupled similarity between attributes. The coupling relationships between values and between attributes contribute to a more comprehensive understanding of object similarity [4]. No work that systematically considers both intra-coupled and inter-coupled similarities has been reported in the literature. This fact leads to the incomplete description of categorical value similarities, and apart from this, the similarity analysis on dependency aggregation is usually very costly.

In this paper, we propose a Coupled Object Similarity (COS) measure by considering both Intra-coupled and Inter-coupled Attribute Value Similarities ($IaAVS$ and $IeAVS$), which capture the attribute value frequency distribution and feature dependency aggregation with a high learning accuracy and relatively low complexity, respectively. We compare accuracies and efficiencies among the four proposed metrics for $IeAVS$, and come up with an optimal one from both theoretical and experimental aspects; we then evaluate our proposed measure with an existing metric on a variety of benchmark categorical data sets in terms of clustering qualities; and we develop a method to define dissimilarity metrics flexibly with our fundamental similarity building blocks according to specific requirements..

The paper is organized as follows. In Section 2, we briefly review the related work. Preliminary definitions are specified in Section 3. Section 4 proposes the coupled similarities, and the theoretical analysis is given in Section 5. We demonstrate the efficiency and effectiveness of COS in Section 6 with experiments. Finally, we end this paper in Section 7.

2. RELATED WORK

There are some surveys [2, 7] that discuss the similarity between categorical attributes. Cost and Salzberg [5] proposed $MVDM$ based on labels, while Wilson and Martinez [10] studied heterogeneous distances for instance based learning. Unlike our focus here, the measures in their study are only designed for supervised approaches.

For unsupervised learning, there exist some data mining techniques for nominal data [1, 2]. The most famous are the SMS measure and its diverse variants such as Jaccard coefficients [7], which are all intuitively based on the principle that the similarity measure is 1 with identical values and is otherwise 0. More recently, attribute value frequency distribution has been considered for similarity measures [2]; neighborhood-based similarities [8] are explored to describe the object neighborhood by using an overlap measure. They are different from our proposed method, which directly reveals the similarity between a pair of objects.

Recently, increasing numbers of researchers have argued that the attribute value similarities are also dependent on their coupling relations [2, 4]. Das and Mannila presented the Iterated Contextual Distances algorithm, believing that the feature and object similarities are inter-dependent [6]. Ahmad and Dey [1] proposed computing the dissimilarity by considering the co-occurrence. While the dissimilarity metric of the latter leads to high accuracy, the computation is usually very costly, which limits its application in large-scale problems.

Table 2: An Example of Information Table

A U	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

3. PROBLEM STATEMENT

A large number of data objects with the same features can be organized by an information table $S =< U, A, V, f >$, where $U = \{u_1, \cdots, u_m\}$ is composed of a nonempty finite set of data objects; $A = \{a_1, \cdots, a_n\}$ is a finite set of features; $V = \bigcup_{j=1}^{n} V_j$ is a set of all attribute values, in which V_j is the set of attribute values of feature $a_j (1 \leq j \leq m)$; and $f = \wedge_{j=1}^{n} f_j$ $(f_j : U \to V_j)$ is an information function which assigns a particular value of each feature to every object. For instance, Table 2 consists of six objects and three features, with $f_2(u_1) = B_1$ and $V_2 = \{B_1, B_2, B_3\}$.

Generally speaking, the similarity between two objects $u_{i_1}, u_{i_2} \in U$ is built on top of the similarities within their values $x, y \in V_j$ for all the features a_j. The basic concepts below are defined to facilitate the formulation for attribute value similarities, where $|H|$ is the number of elements in H.

Definition 3.1. *Given an information table S, three* ***Set Information Functions (SIFs)*** *are defined as $f_j^* : 2^U \to 2^{V_j}$, $g_j : V_j \to 2^U$, and $g_j^* : 2^{V_j} \to 2^U$. Specifically:*

$$f_j^*(\{u_{k_1}, \cdots, u_{k_t}\}) = \{f_j(u_{k_1}), \cdots, f_j(u_{k_t})\}, \tag{3.1}$$

$$g_j(x) = \{u_i | f_j(u_i) = x, 1 \leq j \leq n, 1 \leq i \leq m\}, \tag{3.2}$$

$$g_j^*(W) = \{u_i | f_j(u_i) \in W, 1 \leq j \leq n, 1 \leq i \leq m\}, \tag{3.3}$$

where $u_i, u_{k_1}, \cdots, u_{k_t} \in U$, and $W \subseteq V_j$.

These SIFs describe the relationships between objects and attribute values from different levels. For example, $f_2^*(\{u_1, u_2, u_3\}) = \{B_1, B_2\}$, $g_2(B_1) = \{u_1, u_2\}$ for value B_1, while $g_2^*(\{B_1, B_2\}) = \{u_1, u_2, u_3, u_6\}$ if given $W = \{B_1, B_2\}$.

Definition 3.2. *Given an information table S, its* ***Inter-information Function (IIF)*** *$\varphi_{j \to k} : V_j \to 2^{V_k}$ is defined:*

$$\varphi_{j \to k}(x) = f_k^*(g_j(x)). \tag{3.4}$$

This IIF $\varphi_{j \to k}$ is the composition of f_k^* and g_j. It obtains the kth attribute value subset for the corresponding objects, which are derived from the jth attribute value x. For example, $\varphi_{2 \to 1}(B_1) = \{A_1, A_2\}$.

Definition 3.3. *Given an information table S, the kth attribute value subset $W \subseteq V_k$, and the jth attribute value $x \in V_j$, the* ***Information Conditional Probability (ICP)*** *of W with respect to x is $P_{k|j}(W|x)$:*

$$P_{k|j}(W|x) = \frac{|g_k^*(W) \bigcap g_j(x)|}{|g_j(x)|}. \tag{3.5}$$

Intuitively, when given all the objects with the jth attribute value x, *ICP* is the percentage of the common objects whose kth attribute values fall in subset W and jth attribute value is exactly x as well. For example, $P_{1|2}(\{A_1\}|B_1) = 0.5$.

All these concepts and functions are composed to formalize the so-called coupled interactions between categorical attribute values, as presented below.

4. COUPLED SIMILARITIES

In this section, ***Coupled Attribute Value Similarity (CAVS)*** is proposed in terms of both intra-coupled and inter-coupled value similarities. When we consider the similarity between attribute values, "intra-coupled" indicates the involvement of attribute value occurrence frequencies within one feature, while the "inter-coupled" means the interaction of other features with this attribute. For example, the coupled value similarity between B_1 and B_2 concerns both the intra-coupled relationship specified by the repeated times of values B_1 and B_2: 2 and 2, and the inter-coupled interaction triggered by the other two features (a_1 and a_3).

Suppose we have the ***Intra-coupled Attribute Value Similarity (IaAVS)*** measure $\delta^{Ia}_j(x, y)$ and ***Inter-coupled Attribute Value Similarity (IeAVS)*** measure $\delta^{Ie}_j(x, y)$ for feature a_j and $x, y \in V_j$, then *CAVS* $\delta^A_j(x, y)$ is naturally derived by simultaneously considering both of them.

Definition 4.1. *Given an information table S, the* ***Coupled Attribute Value Similarity (CAVS)*** *between attribute values x and y of feature a_j is:*

$$\delta^A_j(x, y) = \delta^{Ia}_j(x, y) \cdot \delta^{Ie}_j(x, y) \tag{4.1}$$

where δ^{Ia}_j and δ^{Ie}_j are IaAVS and IeAVS, respectively.

4.1 Intra-coupled Interaction

According to [7], it is a fact that the discrepancy of attribute value occurrence times reflects the value similarity in terms of frequency distribution. Thus, when calculating attribute value similarity, we consider the relationship between attribute value frequencies on one feature, proposed as intra-coupled similarity in the following.

Definition 4.2. *Given an information table S, the* ***Intra-coupled Attribute Value Similarity (IaAVS)*** *between attribute values x and y of feature a_j is:*

$$\delta^{Ia}_j(x, y) = \frac{|g_j(x)| \cdot |g_j(y)|}{|g_j(x)| + |g_j(y)| + |g_j(x)| \cdot |g_j(y)|}. \tag{4.2}$$

In this way, different occurrence frequencies indicate distinct levels of attribute value significance. Gan et al. [7] reveal that greater similarity is assigned to the attribute value pair which owns approximately equal frequencies. The higher these frequencies are, the closer such two values are. Thus, function (4.2) is designed to satisfy these two principles. Besides, since $1 \leq |g_j(x)|, |g_j(y)| \leq m$, then $\delta^{Ia}_j \in [1/3, m/(m + 2)]$. For example, in Table 2, both values B_1 and B_2 are observed twice, so $\delta^{Ia}_2(B_1, B_2) = 0.5$.

Hence, by taking into account the frequencies of categories, an effective measure (*IaAVS*) has been captured to characterize the value similarity in terms of occurrence times.

4.2 Inter-coupled Interaction

In terms of *IaAVS*, we have considered the intra-coupled similarity, i.e., the interaction of attribute values within one feature a_j. This does not, however, involve the couplings between other features $a_k (k \neq j)$ and feature a_j when calculating attribute value similarity. Accordingly, we discuss this dependency aggregation, i.e., inter-coupled interaction.

In 1993, Cost and Salzberg [5] proposed a powerful method, *MVDM*, for measuring the dissimilarity between categorical values. *MVDM* considers the overall similarities of classification of all objects on each possible value of each feature. The idea is that attribute values are identified as being similar if they occur with the same relative frequency for all classifications. In the absence of labels, the above measure is adapted to satisfy our target problem by replacing the class label with some other feature to enable unsupervised learning. We regard this interaction between features as inter-coupled similarity in terms of the co-occurrence comparisons of *ICP*. The most intuitive variant is *IRSP*:

Definition 4.3. *Given an information table S, the* ***Inter-coupled Relative Similarity based on Power Set (IRSP)*** *between attribute values x and y of feature a_j based on another feature a_k is:*

$$\delta^P_{j|k}(x, y) = \min_{W \subseteq V_k} \{2 - P_{k|j}(W|x) - P_{k|j}(\overline{W}|y)\}, \tag{4.3}$$

where $\overline{W} = V_k \backslash W$ is the complementary set of a set W under the complete set V_k.

In fact, 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 2, by employing (4.3), we want to get $\delta^P_{2|1}(B_1, B_2)$, i.e. the similarity between two attribute values B_1, B_2 of feature a_2 regarding feature a_1. Since the set of all attribute values of feature a_1 is $V_1 = \{A_1, A_2, A_3, A_4\}$, the number of all power sets within V_1 is 2^4, i.e., the number of the combinations consisting of $W \subseteq V_1$ and $\overline{W} \subseteq V_1$ is 2^4. The minimal value among them is 0.5, which indicates that similarity $\delta^P_{2|1}(B_1, B_2) = 0.5$.

This process shows the combinational explosion brought about by the power set needs to be considered when calculating attribute value similarity by *IRSP*. We therefore try to define three more similarities based on *IRSP* as follows.

Definition 4.4. *Given an information table S, the* ***Inter-coupled Relative Similarity based on Universal Set (IRSU), Join Set (IRSJ), and Intersection Set (IRSI)*** *between attribute values x and y of feature a_j based on another feature a_k are the following formulae respectively:*

$$\delta^U_{j|k}(x, y) = 2 - \sum_{w \in V_k} \max\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}, \tag{4.4}$$

$$\delta^J_{j|k}(x, y) = 2 - \sum_{w \in \bigcup} \max\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}, \tag{4.5}$$

$$\delta^I_{j|k}(x, y) = \sum_{w \in \bigcap} \min\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}, \tag{4.6}$$

where $w \in \bigcup$ and $w \in \bigcap$ denote $w \in \varphi_{j\to k}(x) \bigcup \varphi_{j\to k}(y)$ and $w \in \varphi_{j\to k}(x) \bigcap \varphi_{j\to k}(y)$, respectively.

Each kth attribute value $w \in V_k$, rather than its value subset $W \subseteq V_k$, is considered to reduce computational complexity. In this way, *IRSU* is applied to compute similarity

$\delta^U_{2|1}(B_1, B_2)$, and we get $\delta^U_{2|1}(B_1, B_2) = 0.5$. Since *IRSU* only concerns all the single attribute values rather than exploring the whole power set, it has solved the combinational explosion issue to a great extent. In *IRSU*, *ICP* is merely calculated 8 times compared with 32 times by *IRSP*, which leads to a substantial improvement in efficiency. Then with (4.5), the calculation of $\delta^J_{2|1}(B_1, B_2)$ is further simplified since $A_3 \notin \varphi_{2\to1}(B_1) \bigcup \varphi_{2\to1}(B_2)$. Thus, we obtain $\delta^J_{2|1}(B_1, B_2) = 0.5$, which reveals the fact that it is enough to compute *ICP* with $w \in V_1$ that belongs to $\varphi_{2\to1}(B_1) \bigcup \varphi_{2\to1}$ (B_2) instead of all the elements in V_1. From this perspective, *IRSJ* reduces the complexity further when compared with *IRSU*. Based on *IRSU*, an alternative *IRSI* is considered. For example, with (4.6), the calculation of $\delta^I_{2|1}(B_1, B_2)$ is once again simplified since only $A_2 \in \varphi_{2\to1}(B_1) \bigcap \varphi_{2\to1}(B_2)$. Then, we easily get $\delta^I_{2|1}(B_1, B_2) = 0.5$. In this case, it is sufficient to compute *ICP* with $w \in V_1$ which only belongs to $\varphi_{2\to1}(B_1) \bigcap \varphi_{2\to1}(B_2)$. It is trivial that the cardinality of intersection $\bigcap$ is no larger than that of join set $\bigcup$. Thus, *IRSI* is further more efficient than *IRSU* due to the reduction of intra-coupled relative similarity complexity.

Intuitively speaking, it is a fact that *IRSI* is the most efficient of all the proposed inter-coupled relative similarity measures: *IRSP*, *IRSU*, *IRSJ*, *IRSI*. In addition, all four measures lead to the same similarity result, such as 0.5.

According to the above discussion, we can naturally define the similarity between the jth attribute value pair (x, y) on top of these four optional measures by aggregating all the relative similarities on features other than attribute a_j.

Definition 4.5. *Given an information table S, the **Inter-coupled Attribute Value Similarity (IeAVS)** between attribute values x and y of feature a_j is:*

$$\delta^{Ie}_j(x, y) = \sum_{k=1, k\neq j}^{n} \alpha_k \delta_{j|k}(x, y), \qquad (4.7)$$

where α_k is the weight parameter for feature a_k, $\sum_{k=1}^{n} \alpha_k = 1$, $\alpha_k \in [0, 1]$, and $\delta_{j|k}(x, y)$ is one of the inter-coupled relative similarity candidates.

Accordingly, we have $\delta^{Ie}_j \in [0, 1]$, then $\delta^A_j = \delta^{Ia}_j \cdot \delta^{Ie}_j \in [0, m/(m+2)]$ since $\delta^{Ia}_j \in [1/3, m/(m+2)]$. In Table 2, for example, $\delta^{Ie}_2(B_1, B_2) = 0.5{\cdot}\delta_{2|1}(B_1, B_2) + 0.5{\cdot}\delta_{2|3}(B_1, B_2) = (0.5 + 0)/2 = 0.25$ if $\alpha_1 = \alpha_3 = 0.5$ is taken with equal weight. Furthermore, coupled attribute value similarity (4.1) is obtained as $\delta^A_2(B_1, B_2) = \delta^{Ia}_2(B_1, B_2) \cdot \delta^{Ie}_2(B_1, B_2) = 0.5 \times 0.25 = 0.125$. For the Movie data set in Section 1, then $\delta^A_{Director}(Scorsese, Coppola) = \delta^A_{Director}(Coppola, Coppola) = 0.33$, and $\delta^A_{Director}$ $(Koster, Coppola) = 0$ while $\delta^A_{Director}$ $(Koster, Hitchcock) = 0.25$. 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 1.

After specifying *IaAVS* and *IeAVS*, a coupled similarity between objects is built based on *CAVS*. Then, we consider the sum of all these *CAVS*s analogous to the construction of Manhattan dissimilarity [7]. Formally, we have:

Definition 4.6. *Given an information table S, the **Coupled Object Similarity (COS)** between objects u_{i_1} and u_{i_2}:*

$$COS(u_{i_1}, u_{i_2}) = \sum_{j=1}^{n} \delta^A_j(x_{i_1 j}, x_{i_2 j}), \qquad (4.8)$$

where δ^A_j is the CAVS *measure defined in (4.1), $x_{i_1 j}$ and $x_{i_2 j}$ are the attribute values of feature a_j for objects u_{i_1} and u_{i_2} respectively, and $1 \le i_1, i_2 \le m$, $1 \le j \le n$.*

Table 3: Computational Complexity for *CAVS*

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	$nR(R-1)/2$	$2(n-1)P$	$O(n^2R^2R)$
IRSI	$nR(R-1)/2$	$2(n-1)Q$	$O(n^2R^2R)$

For *COS*, all the *CAVS*s with each feature are summed up for two objects. For example (Table 2), $COS(u_2, u_3) = \sum_{j=1}^{3} \delta_j(x_{2j}, x_{3j}) = 0.5 + 0.125 + 0.125 = 0.75$.

5. THEORETICAL ANALYSIS

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

1) Computational Accuracy Equivalence

From the aspect of set theory, these four measures are equivalent to one another in calculating value similarity.

Theorem 5.1. *IRSP, IRSU, IRSJ and IRSI are all equivalent to one another.*[2]

The above theorem also explains the similarity result in Section 4.2. Thus, these measures induce exactly the same computational accuracy in machine learning tasks.

2) Computational Complexity Comparison

Suppose we have an information table S with m objects and n features, the maximal number of attribute values for all the features is R. In total, the number of attribute value pairs for all the features 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 $|ICP^{(M)}_{j|k}|$ times by a measure ***IRSM***. As we have n attributes, the total *ICP* time costs for *CAVS* is $2|ICP^{(M)}_{j|k}| \cdot (n-1)$ flops per step. Since we have four options for $\boldsymbol{M}$, the computational complexities for calculating all the *CAVS*s are shown in Table 3.

As indicated in Table 3, all the measures have the same calculation steps, while their flops per step are sorted in descending order since $2^R > R \ge P \ge Q$, in which P and Q are the join and intersection sets of the corresponding *IIF*s, respectively. This evidences that the computational complexity essentially depends on the time costs of *ICP* linearly with given data. Specifically, *IRSP* has the largest complexity $O(n^2R^22^R)$, compared to 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 $Q \le P \le R$. In fact, the dissimilarity that Ahmad and Dey [1] have used for mixed data clustering corresponds to the worst measure *IRSP* discussed here.

Considering both the accuracy analysis and complexity comparison, we conclude that *IRSI* is the best performing because it indicates the least complexity but still maintains an equal accuracy to present coupling.

[2] All detailed proofs of Theorem 5.1 are available on request.

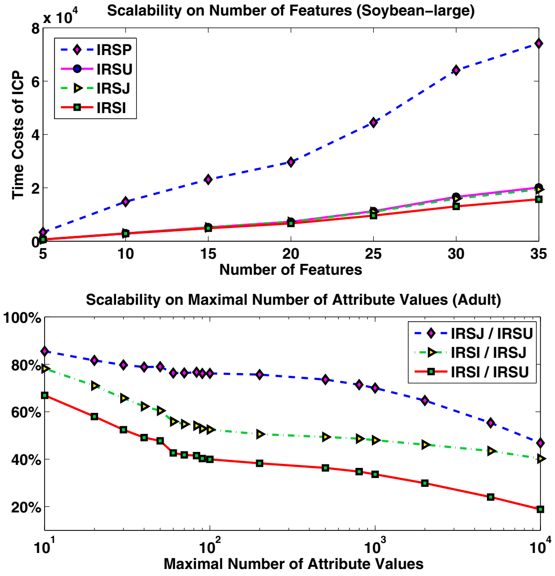

Figure 1: Scalability on $|A|$ and R respectively.

6. EXPERIMENT AND EVALUATION

In this section, several experiments are performed on extensive UCI data sets to show the effectiveness and efficiency of our proposed coupled similarities. The experiments are divided into two categories: coupled similarity comparison and COS application. For simplicity, we just assign the weight vector $\alpha = (\alpha_k)_{1\times n}$ with values $\alpha(k) = 1/n$ in (4.7).

6.1 Coupled Similarity Comparison

To compare efficiencies, we conduct extensive experiments on the inter-coupled relative similarity metrics: *IRSP*, *IRSU*, *IRSJ*, and *IRSI*. The goal in this set of experiments is to show the obvious superiority of *IRSI*, compared with the most time-consuming measure *IRSP*. As discussed in Section 5, the computational complexity linearly depends on the time costs of *ICP* with given data. Thus, we consider a comparison of complexities represented by the time costs of *ICP*. Also explained in Section 5, the complexity for *IRSP* is $O(n^2R^22^R)$, while the other three have equal smaller complexity $O(n^2R^3)$. Here, scalability analysis is explored in terms of these two factors separately: the number of features $|A|$ and the maximal number of attribute values R.

From the perspective of $|A|$, Soybean-large data set is considered with 307 objects and 35 features. Here, we fix R to be 7, and focus on $|A|$ ranging from 5 to 35 with step 5. In terms of the total time costs of *ICP*, the computational complexity comparisons among four measures (*IRSP*, *IRSU*, *IRSJ*, and *IRSI*) are depicted in Figure 1($|A|$). The result indicates that the complexities of all these measures keep increasing when $|A|$ becomes larger. The acceleration of *IRSP* (from 3328 to 74128) is the greatest compared with the slightest acceleration of *IRSI* (from 632 to 15704). Apart from these two, the scalability curves are almost the same for *IRSU* and *IRSI*, though the complexity of *IRSU* is slightly higher than that of *IRSJ* with varied $|A|$. Therefore, *IRSI* is the most stable and efficient measure to calculate the intra-coupled relative similarity in terms of $|A|$.

From the perspective of R, the variation of R is considered when $|A|$ is confirmed. Here, we take advantage of the Adult data set with 30718 objects and 13 features chosen. Specifically, the integer feature "fnlwgt" is discretized into different intervals (from 10 to 10000) to form distinct R ranging from 16 to 10000, since one of the existing categorial attributes "education" already has 16 values. The outcomes are shown in Figure 1(R), in which the horizontal axis refers to R, and the vertical axis indicates the relative complexity ratios in terms of $\xi(J/U)$, $\xi(I/J)$, and $\xi(I/U)$. From this figure, we observe all the ratios between 10% and 100%, which again verifies the complexity order for these four measures indicated in Section 5. Another issue is that all three curves decrease as R grows, which means the efficiency advantages of *IRSJ* upon *IRSU* (from 85.5% to 46.8%), *IRSI* upon *IRSJ* (from 78.2% to 40.2%), and *IRSI* upon *IRSU* (from 66.9% to 18.8%) all become more and more obvious with the increasing of R. The general trend of these ratios always falling comes from the fact that there is a higher probability of getting a join set smaller than the whole set, and an intersection set smaller than the join set, with larger R. The same conclusion also holds for the ratio $\xi(U/P)$, but this is due to the fact that $q^{-1}(x) = x/2^x$ is a strictly monotonously decreasing function when $x > 1$. We omit this ratio in Figure 1(R) since the denominator $|ICP^{(P)}|$ becomes exponentially large when R grows, e.g., it equals to 5.12×10^{83} when $R = 500$. Hence, *IRSI* is the least time-consuming intra-coupled similarity with regard to R.

In summary, all the above experiment results clearly show that *IRSI* outperforms *IRSP*, *IRSU*, and *IRSJ* in terms of the computational complexity. In particular, with the increasing numbers of either features or attribute values, *IRSI* demonstrates superior efficiency compared to the others. *IRSJ* and *IRSU* follow, with *IRSP* being the most time-consuming, especially for the large-scale data set.

6.2 Application

In this part of our experiments, we focus on the computational accuracy comparison. In the following, we evaluate the COD which is derived from (4.8):

$$COD(u_{i_1}, u_{i_2}) = \sum_{j=1}^{n} h_1(\delta_j^{Ia}(x_{i_1j}, x_{i_2j})) \cdot h_2(\delta_j^{Ie}(x_{i_1j}, x_{i_2j})), \tag{6.1}$$

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. Here, we consider $h_1(t) = 1/t - 1$ and $h_2(t) = 1 - t$ to reflect the complementarity of similarity and dissimilarity measures. In terms of the capability on revealing the relationship between data, the better the dissimilarity induced, the better is its similarity.

To demonstrate the effectiveness of our proposed COD in application, we compare two clustering methods based on two dissimilarity metrics on six data sets. Here, COD is used with the outperforming measure *IRSI*.

One of the clustering approaches is the k-modes (KM) algorithm [7], 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., spectral clustering (SC) [9], which makes use of the Laplacian Eigenmaps on dissimilarity matrix to perform dimensionality reduction for clustering prior to the k-means algorithm. In respect of feature dependency aggregations, however, Ahmad and Dey [1] evidenced that

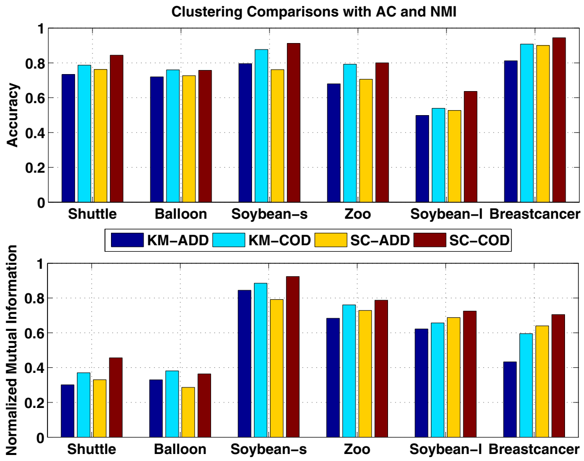

Figure 2: Clustering evaluation on six data sets

their proposed metric *ADD* outperforms *SMD* in terms of *KM* clustering. Thus, we aim to compare the performances of *ADD* [1] and *COD* (6.1) for further clustering evaluations.

We conduct four groups of experiments on the same data sets: *KM* with *ADD*, *KM* with*COD*, *SC* with *ADD*, and *SC* with*COD*. 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*) [3]. $AC \in [0, 1]$ 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. $AC = 1$ or $NMI = 1$ if the clusters and labels are identical, and $AC = 0$ or $NMI = 0$ if the two sets are independent. In fact, the larger *AC* or *NMI* is, the better the clustering is, and the better the corresponding dissimilarity metric is.

Figure 2 reports the results on six data sets with different $|U|$, ranging from 15 to 699 in increasing order. In terms of *AC* and *NMI*, the evaluations are conducted with *KM-ADD*, *KM-COD*, *SC-ADD*, and *SC-COD* individually. Followed by Laplacian Eigenmaps, the subspace dimensions are determined by the number of labels in *SC*. For each data set, the average performance is computed over 100 tests for *KM* and k-means in *SC* with distinct start points.

As can be clearly seen from Figure 2, the clustering methods with *COD*, whether *KM* or *SC*, outperform those with *ADD* in terms of both *AC* and *NMI* measures. That is to say, dissimilarity metric *COD* is better than *ADD* on clustering qualities. Specifically 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) 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), respectively; however, the latter rate belongs to [5.45% (Soybean-l), 38.12% (Shuttle)]. Since *AC* and *NMI* evaluate clustering quality from different aspects, they generally take minimal and maximal ratios on distinct data sets. Another significant observation is that *SC* mostly outperforms *KM* a little whenever it has the same dissimilarity metric; in fact, Luxburg [9] has indicated that *SC* very often outperforms k-means for numerical data.

We draw the following two conclusions: 1) intra-coupled relative similarity *IRSI* is the most efficient one when compared with *IRSP*, *IRSU* and *IRSJ*, especially for large-scale data; 2) our proposed object dissimilarity metric *COD* is better than others, such as dependency aggregation only *ADD*, for categorical data in terms of clustering qualities.

7. CONCLUSION

We have proposed *COS*, a novel coupled object similarity metric which involves both attribute value frequency distribution (intra-coupling) and feature dependency aggregation (inter-coupling) in measuring attribute value similarity for unsupervised learning of nominal data. Theoretical analysis and substantial experiments have shown that inter-coupled relative similarity measure *IRSI* significantly outperforms the others (*IRSP*, *IRSU*, *IRSJ*) in terms of efficiency, in particular on large-scale data, while maintaining equal accuracy. Moreover, our derived dissimilarity metric is more comprehensive and accurate in capturing the clustering qualities in accordance with substantial empirical results.

We are currently applying the *COS* measure with *IRSI* to feature discretization, clustering ensemble, and other data mining tasks. We are also considering extending the notion of "coupling" for the similarity of numerical data. Moreover, the proposed concepts *Inter-information Function* and *Information Conditional Probability* for the information table have potential for other applications.

8. ACKNOWLEDGMENT

This work is sponsored by Australian Research Council Grants (DP1096218, DP0988016, LP100200774, LP0989721), Tianjin Research Project (10JCYBJC07500), and QCIS (Center for Quantum Computation and Intelligent Systems).

9. REFERENCES

[1] A. Ahmad and L. Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering*, 63:503–527, 2007.

[2] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: a comparative evaluation. In *SDM 2008*, pages 243–254, 2008.

[3] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE TKDE*, 17(12):1624–1637, 2005.

[4] L. Cao, Y. Ou, and P. Yu. Coupled behavior analysis with applications. *IEEE Transactions on Knowledge and Data Engineering*, 2011.

[5] S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10(1):57–78, 1993.

[6] G. Das and H. Mannila. Context-based similarity measures for categorical databases. In *PKDD 2000*, pages 201–210, 2000.

[7] G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*. ASA-SIAM Series on Statistics and Applied Probability, VA, 2007.

[8] M. Houle, V. Oria, and U. Qasim. Active caching for similarity queries based on shared-neighbor information. In *CIKM 2010*, pages 669–678, 2010.

[9] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):1–32, 2007.

[10] D. Wilson and T. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.