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.

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 $f = \wedge_{j=1}^{n} f_j$ $(f_j : U \rightarrow V_j)$ 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 $f_j^* : 2^U \rightarrow 2^{V_j}$, $g_j : V_j \rightarrow 2^U$, and $g_j^* : 2^{V_j} \rightarrow 2^U$. 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 *SIF*s 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)*** *$\varphi_{j \rightarrow k} : V_j \rightarrow 2^{V_k}$ is defined:*

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

This *IIF* $\varphi_{j \rightarrow k}$ 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, $\varphi_{2 \rightarrow 1}(B_1) = \{A_1, A_2\}$.

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 $\delta_j^{Ia}(x, y)$ and ***Inter-coupled Attribute Value Similarity (IeAVS)*** measure $\delta_j^{Ie}(x, y)$ for feature a_j and $x, y \in V_j$, then *CAVS* $\delta_j^A(x, y)$ 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:*

$$\delta_j^A(x, y) = \delta_j^{Ia}(x, y) \cdot \delta_j^{Ie}(x, y) \tag{4.1}$$

where δ_j^{Ia} and δ_j^{Ie} 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:*

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

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 $1 \le |g_j(x)|, |g_j(y)| \le m$, then $\delta_j^{Ia} \in [1/3, m/(m + 2)]$. For example, in Table 2, both values B_1 and B_2 are observed twice, so $\delta_2^{Ia}(B_1, B_2) = 0.5$.

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:*

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

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 $\delta_{2|1}^P(B_1, B_2)$, 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 $\delta_{2|1}^P(B_1, B_2) = 0.5$.

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:*

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

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

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

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:*

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

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}:*

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

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 $1 \leq i_1, i_2 \leq m$, $1 \leq j \leq n$.*

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 *IRS**M***. 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 $2^R > R \geq P \geq Q$, 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 $Q \leq P \leq R$. 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.

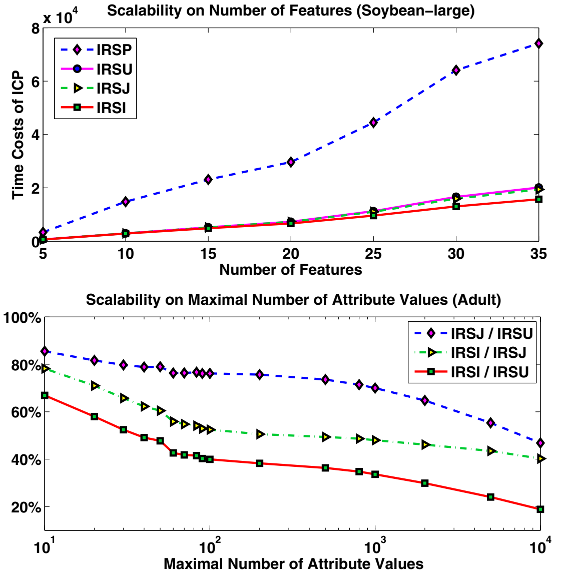

Figure 1: Scalability on $|A|$ and R respectively.

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

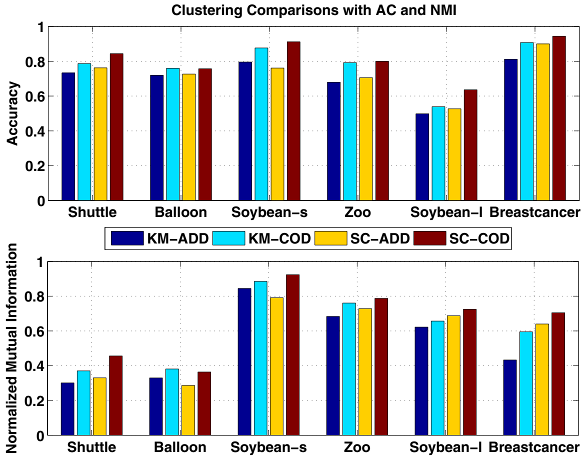

Figure 2: Clustering evaluation on six data sets

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.