A Weighting Similarity Learning on Categorical Data

Fuyuan Cao[a], Jie Wen[a]

[a]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*

Abstract

Attribute independence has been taken as a major assumption in the limited research that has been conducted on similarity analysis for categorical data. However, in real-world data sources,attribute are more or less associated with each other in terms of certain coupling relationships. This paper proposes a weighting distance 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 attribute as well as their integration on the object level. Substantial experiments on extensive UCI data sets verify the theoretical conclusions. The experimental results show that the similarity measure proposed in this paper has a good effect on clustering data clustering.

Keywords: Clustering, Coupled attribute similarity, Weighting similarity learning

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, document analysis, and image analysis. A typical aspect of these applications

*Corresponding author

Email addresses: cfy@sxu.edu.cn (Fuyuan Cao), 1967688145@qq.com (Jie Wen)

is clustering. The similarity between clusters is often built on top of the similarity between data objects. 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 lager is the similarity. The other similarity between attributes can also be converted into the difference of similarities between pairwise attribute values. Therefore, the similarity between attribute values plays a fundamental role in similarity analysis.

Compared with the intensive study on the similarity between two numerical variables, such as Euclidean and Minkowski distance, the similarity for categorical data has received much less attention. Only limited efforts have been made, including SMS, which uses 0s and 1s to distinguish the similarity between distinct and identical categorical values, occurrence frequency (OF) and information-theoretical similarity (Lin), 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 only deliver a local picture of the similarity. In addition, none of them provides a comprehensive similarity between categorical attributes by combining relevant aspects. A real database application example is described in Table 1.

Table 1: Instance Of The Movie Database

Movie	*Actor*	*Genre*	*Director*	*Class*
Godfaher II	De Niro	Crime	Scorsese	L1
Good Fellas	De Niro	Crime	Coppola	L1
Vertigo	Stewart	Thriller	Hitchcock	L2
Harvey	Stewart	Comedy	Koster	L2
N by NW	Grant	Thriller	Hitchcock	L2
Bishop's Wife	Grant	Comedy	Koster	L2

As shown in Table 1, six movie objects are divided into two classes with three nominal attributes. The SMS measure between the value of Actor of Vertigo's Stewart and the value of Actor of N by NW's Grant is 0, but Stewart and Grant are very similar. Another observation by following SMS is that the

similarity between Stewart and Grant is equal to that between De Niro and Stewart; however, the similarity of the former pair should be greater because both Actor belong of the same class L2.

The above examples show that it is much more complex to analyse the similarity between nominal variables than between continuous data. The SMS and its variants fail to capture a global picture of the genuine relationship for nominal data. Can Wang has put forward an effective algorithm (CADO) to improve the shortcomings of the above algorithms. She make a point about the data-driven intra-coupled similarity and inter-coupled similarity, as well as their global aggregation in unsupervised learning on nominal data.

However, the measure of intra-coupled similarity in her method does not show the similarity in the same class and the dissimilarity between different classes, and the measure of relationship between attributes is not given in the CASO algorithm. For example, the CADO measure similarity between the value of Actor of Godfather II's De Niro and the value of Actor of Good Fellas's De Niro is 0.5 and the similarity between the value of Actor of Good Fellas's De Niro and the value of Actor of Vertigo's Stewart is the same. However, since both Actor of the former pair belong of the same class L1, so the similarity should be greater.

In this paper, we explicitly discuss the distance measure for categorical data objects. This distance matrix takes into account the characteristics of the categorical values. The core idea is to measure the distance with the frequency probability of each attribute value in the whole data set. A new kind of weight named dynamic attribute weight has been presented to adjust the contribution of distance along each attribute to the whole object distance. Moreover, in order to utilize the useful relationship information accompanying with each pair of attributes well, the interdependence redundancy measure has been introduced to evaluate the dependence degree between different attributes. Subsequently, the distance between two values from one attribute is not only measured by their own frequency probabilities but also determined by the values of other attributes that are highly correlated with this one. The key contribution are as

follows.

- A dynamic weighting scheme for categorical attributes is presented, which assigns larger weights to the attributes with infrequent matching or mismatching value pairs as they can provide more important information.
- The dependence degree between each pair of attributes is introduced. The complete distance between two categorical values from one attribute is estimated with not only their own frequency probability but also the co-occurrent probability with other values from highly correlated attributes.

This paper is organized as follows. In Section 2, we specify preliminary definitions. The dynamic attribute weight and the relationship between pair of attributes are given in Section 3. Section 4 defines the intra-coupled similarity, inter-coupled similarity, and their aggregation. We describe the Weight-CADO algorithm in Section 5. The effectiveness of Weight-CADO is empirically studied in Section 6 and a flexible method to define dissimilarity metrics is also developed. Finally, we conclude this paper in Section 7.

2. Preliminary Definitions

Given the data set $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ with n objects represented by d categorical attributes $\{A_1, A_2, \cdots, A_d\}$. V_j is the set of attribute values from attribute $A_j (1 \leq i \leq d)$.

Definition 1 (p_r): The probability of attribute r in presenting value equal to x_{ir} in the object set X.

$$p(A_r = x_{ir}|X) = \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} \tag{1}$$

Here, the operation $\sigma_{\mathbf{A_r}=\mathbf{x_{ir}}}(\mathbf{X})$ counts the number of objects in the data set X that have the value $\mathbf{x_{ir}}$ for attribute $\mathbf{A_r}$ and the symbol NULL refers to the empty. $\mathbf{x_{ir}}$ is the categorical value of attribute $\mathbf{A_r}$ from data objects $\mathbf{x_i}$.

Definition 2 (p_r^-): The estimated probability of attribute r in presenting a value equal to x_{ir} in the object set X.

$$p^-(A_r = x_{ir}|X) = \frac{\sigma_{A_r=x_{ir}}(X) - 1}{\sigma_{A_r \neq NULL}(X) - 1} \tag{2}$$

For example, based on the attribute Actor in Table 1, $p(A_{Actor} = Stewart|X) = \frac{1}{3}$, $p^-(A_{Actor} = Stewart|X) = \frac{1}{5}$.

Definition 3 (ICP): The value subject $\acute{V_k}(\subseteq V_k)$ of attribute A_k, and the value $v_j(\in V_j)$ of attribute A_j, then the information conditional probability (ICP) of $\acute{V_k}$ with respect to v_j is $P_{k|j}(\acute{V_k}|v_j)$, defined as

$$P_{k|j}(\acute{V_k}|v_j) = \frac{\sigma_{A_k=\acute{V_k}} \wedge \sigma_{A_j=v_j}(X)}{\sigma_{A_j=v_j}(X)} \tag{3}$$

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_j fall in subset $\acute{V_k}$ and whose values of attribute a_j are exactly v_j as well. Hence, ICP quantifies the relative overlapping ratio of attribute values in terms of objects. for example, $P_{Actor|Genre}(Grant|Thriller) = 0.5$.

3. Proposed Weight Metric for Categorical Data

3.1. Dynamic Attribute Weight

It's a very common situation that each attribute have it's special features for categorical data in a data set. That is, unusual features generally can provide more information for the comparison between objects and we pay more attention to these special features they have. Considering this phenomenon, we can further adjust the distance metric according to following criterion. The contribution of the distance is inverse to the probability of these two value's situation in the whole data set. That is, if two data objects have different values along one attribute, then the contribution of the distance between these two values to the whole data distance is inverse to the probability that two data objects have different values along this attribute in the data set, and vice versa. Therefore, this kind of probability can be utilized as a dynamic weight of attribute distance.

For an attribute A_r with m_r possible values, the probability that two data objects form X have the same value along A_r is calculated by

$$p_s(A_r) = \sum_{j=1}^{m_r} p(A_r = a_{rj}|X)p^-(A_r = a_{rj}|X) \tag{4}$$

Correspondingly, the probability that two data objects from X have different values along A_r is given by

$$p_f(A_r) = 1 - p_s(A_r) \tag{5}$$

Subsequently, following the proposed criterion, the dynamic weight of attribute A_r should be:

$$\omega(A_r) = \begin{cases} p_s(A_r), & if \quad x_{ir} = x_{jr} \\ p_f(A_r), & otherwise \end{cases} \tag{6}$$

3.2. Relationship Between Categorical Attributes

Most existing distance or similarity metrics for categorical data treat each attribute individually. However, in real data, we often have some attributes that are highly dependent on each other. So, the computation of similarity or distance for categorical attribute should be considered based on frequently co-occurring items. That is, the similarity between two values from one attribute should be calculated by considering the other attributes that are highly correlated with this one. In particular, given the data set X, the dependence degree between each pair of attributes A_i and A_j $(i, j \in \{1, 2, \cdots, d\})$ can be quantified based on the mutual information between them, which is defined as

$$I(A_i; A_j) = \sum_{r=1}^{m_i} \sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log(\frac{p(a_{ir}, a_{jl})}{p(a_{ir})p(a_{jl})}) \tag{7}$$

Here, the items $p(a_ir)$ and $p(a_jl)$ stand for the frequency probability of the two attribute values in the while data set, which are calculated by

$$p(a_{ir}) = p(A_i = a_{ir}|X) = \frac{\sigma_{A_i=a_{ir}}(X)}{\sigma_{A_i \neq NULL}(X)} \tag{8}$$

$$p(a_{jl}) = p(A_j = a_{jl}|X) = \frac{\sigma_{A_j=a_{jl}}(X)}{\sigma_{A_j \neq NULL}(X)} \tag{9}$$

The expression $p(a_{ir}, a_{jl})$ is to calculate the joint probability of these two attribute values, i.e., the frequency probability of objects in X having $A_i = a_{ir}$ and $A_j = a_{jl}$, which is given by

$$p(a_{ir}, a_{jl}) = p(A_i = a_{ir} \wedge A_j = a_{jl}|X) = \frac{\sigma_{A_i=a_{ir}} \wedge \sigma_{A_j=a_{jl}}(X)}{\sigma_{A_i \neq NULL} \wedge \sigma_{A_j \neq NULL}(X)} \quad (10)$$

The mutual information between two attributes actually measures the average reduction in uncertainty about one attribute that results from learning the value of the other. A larger value of mutual information usually indicates greater dependence. However, a disadvantage of using this index is that its value increase with the number of possible values that can be chosen by each attribute. Therefore, Au et al. proposed to normalize the mutual information with a joint entropy, which yields the interdependence redundancy measure denoted as

$$R(A_i; A_j) = \frac{I(A_i; A_j)}{H(A_i; A_j)} \quad (11)$$

where the joint entropy $H(A_i, A_j)$ is calculated by

$$H(A_i; A_j) = -\sum_{r=1}^{m_i} \sum_{l=1}^{m_j} p(a_{ir}, a_{jl}) \log(p(a_{ir}, a_{jl})) \quad (12)$$

This interdependence redundancy measure evaluates the degree of deviation from independence between two attributes. In particular, $R(A_i; A_j) = 1$ means that the attributes A_i and A_j are strictly dependent on each other while $R(A_i; A_j) = 0$ indicates that they are statistically independent. If the value of $R(A_i; A_j)$ is between 0 and 1, we can say that these two attributes are partially dependent. Since the number of attribute values has no effect on the result of independence redundancy measure,it is perceived as a more ideal index to measure the dependence degree between different categorical attributes.

In the process of experiments, we maintain a $d * d$ relationship matrix R to store the dependence degree of each pair of attributes. Each element $R(i, j)$ of this matrix is given by $R(i, j) = R(A_i; A_j)$. It is obvious that R is a symmetric matrix with all diagonal elements equal to 1. To consider the independent attributes simultaneously in distance measure, for each attribute A_r, we find

out all the attributes that have obvious interdependence with it and store them in a set denoted as S_r. In particular, the set S_r is constructed by

$$S_r = \{A_i | R(A_r; A_i) > \beta, 1 \leq i \leq d\} \tag{13}$$

where β is a specific threshold.

4. Coupled Attribute Similarity

4.1. Intra-Coupled Interaction

According to CADO algorithm, the intra-coupled attribute similarity for values (IaASV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_r^{IaASV}(x_{ir}, x_{jr}) = \frac{\sigma_{A_r=x_{ir}}(X) \cdot \sigma_{A_r=x_{jr}}(X)}{\sigma_{A_r=x_{ir}}(X) + \sigma_{A_r=x_{jr}}(X) + \sigma_{A_r=x_{ir}}(X) \cdot \sigma_{A_r=x_{jr}}(X)} \tag{14}$$

For example, in Table 1, we have $\delta^{Ia}_{Actor}(Stewart, DeNiro) = \delta^{Ia}_{Actor}(DeNiro, DeNiro) = 0.5$ since both De Niro and Stewart appear twice.

However, the measure of intra-coupled similarity in CADO algorithm does not show the similarity in the same class and the dissimilarity between different classes. For instance, the similarity of the Godfather II's De Niro and Good Fellas's De Niro should be greater than the Good Fellas's De Niro and Harvey's Stewart because Godfather's Actor and Good Fellas's Actor belong to the same class L1.

Here, Wang consider $h_1(t) = 1/t - 1$ to reflect the complementarity between similarity and dissimilarity measures. In the algorithm proposed in this paper, we use it too. To overcome the shortcomings of CADO algorithm, we use the dynamic attribute weight we just described in Section 3. Subsequently, the weight intra-coupled attribute dissimilarity for values (Weight-IaADV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_r^{W-IaADV}(x_{ir}, x_{jr}) = \begin{cases} p_s(A_r) * (\frac{1}{IaASV} - 1), & if \quad x_{ir} = x_{jr} \\ p_f(A_r) * (\frac{1}{IaASV} - 1), & otherwise \end{cases} \tag{15}$$

4.2. Inter-Coupled Interaction

According to CADO algorithm, the inter-coupled attribute similarity for values (IeASV) between attribute value x_{ir} and x_{jr} of attribute A_r is

$$\delta_r^{IeASV}(x_{ir}, x_{jr}, \{V_k\}_{k\neq j}) = \sum_{k=1,k\neq j}^{d} \alpha_k \delta_{j|k}(x_{ir}, x_{jr}, V_k) \tag{16}$$

where α_k is the weight parameter for attribute A_k. In CADO algorithm, author assign $\alpha_k = \frac{1}{d-1}$. Here, Wang consider $h_2(t) = 1 - t$ to reflect the complementarity between similarity and dissimilarity measures.

However, this assignment method does not take into account the degree of correlation between the different columns. To overcome the shortcomings of CADO algorithm, we use the relationship matrix we just described in Section 3. Subsequently, the weight inter-coupled attribute similarity for values (Weight-IeASV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_r^{W-IeASV}(x_{ir}, x_{jr}, \{V_k\}_{k\neq j}) = \sum_{k=1,k\neq j}^{d} R(j,k) \delta_{j|k}(x_{ir}, x_{jr}, V_k) \tag{17}$$

In order to make dissimilarity measure satisfy nonnegativity, the weight inter-coupled attribute dissimilarity for values (Weight-IeADV) between values x_{ir} and x_{jr} for attribute A_r, that is, the convert between similarity and dissimilarity measure, is

$$\delta_r^{W-IeADV}(x_{ir}, x_{jr}, \{V_k\}_{k\neq j}) = \sum_{k=1,k\neq j}^{d} R(j,k) - \delta_r^{W-IeASV} \tag{18}$$

4.3. Coupled Interaction

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

The Weight-CADV between attribute values x_{ir} and x_{jr} of attribute A_r is

$$\delta_r^{W-CADV}(x_{ir}, x_{jr}, \{V_k\}_{k=1}^{n}) = \delta_r^{W-IaADV}(x_{ir}, x_{jr}) \cdot \delta_r^{W-IeADV}(x_{ir}, x_{jr}, \{V_k\}_{k\neq j}) \tag{19}$$

where $V_k(k \neq j)$ is a value set of attribute A_k different from A_j to enable the weight inter-coupled interaction. $\delta_r^{W-IaADV}$ and $\delta_r^{W-IeADV}$ are Weight-IaADV and Weight-IeADV.

As indicated in Eq.(19), we choose the multiplication of these two components. Weight-IaADV is associated with how often the value occurs, while Weight-IeADV 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. Alternatively, we could consider other combination forms of Weight-IaADV and Weight-IeADV according to the data structure, such as $\delta_r^{W-CADV}(x_{ir}, x_{jr}, \{V_k\}_{k=1}^n) = \alpha \cdot \delta_r^{W-IaADV}(x_{ir}, x_{jr}) + \gamma \cdot \delta_r^{W-IeADV}(x_{ir}, x_{jr}, \{V_k\}_{k\neq j})$, where $0 \leq \alpha, \gamma \leq 1(\alpha + \gamma = 1)$ are the corresponding weights. Thus, Weight-IaADV and Weight-IeADV can be controlled flexibly to display in which cases the intra-coupled interaction is more significant than the inter-coupled interaction, and vice versa.

5. Coupled Dissimilarity Algorithm

In previous sections, we have discussed the construction of Weight-CADV. In this section, a coupled dissimilarity between objects is built based on Weight-CADV. Below, we consider the sum of all these Weight-CADV measures.

Given the data set X, the Weight-CADO between object x_i and x_j is Weight-CADO(x_i,x_j)

$$Weight - CADO(x_i, x_j) = \sum_{r=1}^{d} \delta_r^{W-CADV}(x_{ir}, x_{jr}, \{V_k\}_{k=1}^n) \tag{20}$$

We then design an algorithm Weight-CADO, given in Algorithm 1, to compute the coupled object dissimilarity. The whole process of this algorithm is summarized as follows:

- Compute the Weight-IaADV for attributes (x_{ir} and x_{jr}) of attribute A_r;
- Compute the Weight-IeADV for attribute values (x_{ir} and x_{jr});
- Compute the Weight-CADV for attribute values (x_{ir} and x_{jr});

Algorithm 1 Weight Coupled Attribute Dissimilarity for Objects

Input: data set $X = \{x_1, x_2, \cdots, x_n\}$.

Output: $D(x_i, x_j)$ for $i, j \in \{1, 2, \cdots, n\}$.

Calculate $p_s(A_r)$ and $p_f(A_r)$ for each attribute A_r according to Eq.(4) and Eq.(5).

For each pair of attributes $(A_r, A_l)(r, l \in \{1, 2, \cdots, d\})$ calculate $R(A_r; A_l)$ according to Eq.(11).

Construct the relationship matrix R.

Get the index set S_r for each attribute A_r by $S_r = \{l | R(r, l) > \beta, 1 \leq l \leq d\}$.

Choose two objects x_i and x_j from X.

Let $D(x_i, x_j) = 0$.

for *attribute* a_r, $r = 1$ *to* n **do**

    // Compute the weight intra-coupled dissimilarity for two attribute values x_{ir} and x_{jr}

    Weight-IaADV = $\delta_r^{W-IaADV}(x_{ir}, x_{jr})$;

    for every value pair $(x_{ir}, x_{jr} \in [1, \delta_{A_r}])$ **do**

        //Compute the weight inter-coupled dissimilarity for two attribute values x_{ir} and x_{jr}

        Weight-IeADV = $\delta_r^{W-IeADV}(x_{ir}, x_{jr}, \{V_k\}_{k \neq j})$;

    end for

    //Compute coupled dissimilarity between two attribute values x_{ir} and x_{jr}

    Weight-CADV = Weight-IaADV $\cdot$ Weight-IeADV;

end for

//Compute coupled similarity between two objects x_i and x_j

Weight-CADO = sum(Weight-CADV);

$D(x_i, x_j)$ = Weight-CADO;

return $D(x_i, x_j)$;

- Compute the Weight-CADO for objects x_i and x_j;

6. Experiments

To investigate the effectiveness of the distance metric for the categorical data proposed in this paper, we mainly make some experiments on the five UCI data sets, Balloons data set, Soybean-small data set, Zoo data set, Congressional Voting Records data set and Breast Cancer data set. We firstly describe the preprocessing process of the five data sets. Then five evaluation indexes are introduced. Finally, we show the comparison results of the Weight-CADO algorithm with other algorithms.

In our experiments, the value of the threshold parameter β in the proposed metric was set equal to the average interdependence redundancy of all attribute pairs. That is, we let $\beta = \beta_0$, where β_0 is calculated by

$$\beta = \frac{1}{d^2} \sum_{i=1}^{d} \sum_{j=1}^{d} R(A_i; A_j), \tag{21}$$

6.1. Data Description

The information of the data sets we utilized is as follows.

- Balloons Data Set: There are 20 instances based on 4 attributes and each sample labeled with T or F.
- Soybean-small Data Set: There are 47 instances characterized by 35 multi-valued categorical attributes. According to the different kinds of diseases, all the instances should be divided into four groups.
- Zoo Data Set: This data set consists 101 instances represented by 16 attributes, in which each instance belongs to one of the seven animal categories.
- Congressional Voting Records Data Set: There are 435 votes based on 16 key features and each vote comes from one of the two different party affiliations.

- Breast Cancer Data Set: This data set has 699 instances described by nine categorical attributes with the values from 1 to 10. Each instance belongs to one of the two clusters labeled by benign and malignant.

6.2. Evaluation Indexes

To evaluate the effectiveness of the Weight-CADO algorithm, we used the following five external criterions: (1) adjusted rand index (ARI) [1], (2) normalized mutual information (NMI) [2], (3) accuracy (AC), (4) precision (PR) and (5) recall (RE) to compare the obtained cluster of each object with that provided by data label.

Let $\mathbf{X}$ be a matrix-object data set, $C = \{C_1, C_2, \cdots, C'_k\}$ be a clustering result of $\mathbf{X}$, $P = \{P_1, P_2, \cdots, P_k\}$ be a real partition in $\mathbf{X}$. The overlap between C and P can be summarized in a contingency table shown in Table 2, where n_{ij} denotes the number of objects in common between P_i and C_j, $n_{ij} = |P_i \bigcap C_j|$. p_i and c_j are the number of objects in P_i and C_j, respectively.

Table 2: The contingency table.

	C_1	C_2	$\cdots$	$C_{k'}$	$Sums$
P_1	n_{11}	n_{12}	$\cdots$	$n_{1k'}$	p_1
P_2	n_{21}	n_{22}	$\cdots$	$n_{2k'}$	p_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
P_k	n_{k1}	n_{k2}	$\cdots$	$n_{kk'}$	p_k
$Sums$	c_1	c_2	$\cdots$	$c_{k'}$	n

The five evaluation indexes are defined as follows:

$$ARI = \frac{\sum_{ij} C_{n_{ij}}^2 - [\sum_i C_{p_i}^2 \sum_j C_{c_j}^2]/C_n^2}{\frac{1}{2}[\sum_i C_{p_i}^2 + \sum_j C_{c_j}^2] - [\sum_i C_{p_i}^2 \sum_j C_{c_j}^2]/C_n^2},$$

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij} log(\frac{n_{ij}n}{p_i c_j})}{\sqrt{\sum_{i=1}^{k} p_i log(\frac{p_i}{n}) \sum_{j=1}^{k'} c_j log(\frac{c_j}{n})}},$$

$$AC = \frac{1}{n} \max_{j_1 j_2 \cdots j_k \in S} \sum_{i=1}^{k} n_{ij_i},$$

$$PE = \frac{1}{k} \sum_{i=1}^{k} \frac{n_{ij_i^*}}{p_i},$$

$$RE = \frac{1}{k'} \sum_{i=1}^{k'} \frac{n_{ij_i^*}}{c_i},$$

where $n_{1j_1^*} + n_{2j_2^*} + \cdots + n_{kj_k^*} = \max_{j_1 j_2 \cdots j_k \in S} \sum_{i=1}^{k} n_{ij_i}$ $(j_1^* j_2^* \cdots j_k^* \in S)$ and $S = \{j_1 j_2 \cdots j_k : j_1, j_2, \cdots, j_k \in \{1, 2, \cdots, k\},\ j_i \neq j_t \text{ for } i \neq t\ \}$ is a set of all permutations of $1, 2, \cdots, k$. For AC, PE, RE, k is equal to k' in general case. In addition, we consider that the higher the values of ARI, NMI, AC, PE and RE are, the better the clustering solution is.

6.3. Comparisons between CADO Alogrithm and Weight-CADO Alogrithm

One of the clustering approaches is the KM algorithm, 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 objects to the nearest mode. The other is a branch of graph-based clustering, i.e., SC, which makes use of Laplacian Eigenmaps on a dissimilarity matrix to perform dimensionality reduction for clustering before the k-means algorithm. Below, we aim to compare the performance of Weight-CADO against CADO as used in data cluster analysis for further clustering evaluation.

Table 3 reports the results on five data sets with different scale, ranging from 20 to 699 in the increasing order. For each data, the average performance is computed over 100 tests for KM and SC with distinct start points. Note that the highest measure score of each experimental setting is highlighted in boldface.

As Table 3 indicates, the clustering

Table 3: Clustering Evaluation On Five Data Sets

Evaluation Index	Data Set	KM		SC	
		CADO	Weight-CADO	CADO	Weight-CADO
AC	Balloons Data Set	label-2	label-3	label-4	label-5
	Soybean-small Data Set	label-2	label-3	label-4	label-5
	Zoo Data Set	label-2	label-3	label-4	label-5
	Voting Data Set	label-2	label-3	label-4	label-5
	Breast Cancer Data Set	label-2	label-3	label-4	label-5
NMI	Balloons Data Set	label-2	label-3	label-4	label-5
	Soybean-small Data Set	label-2	label-3	label-4	label-5
	Zoo Data Set	label-2	label-3	label-4	label-5
	Voting Data Set	label-2	label-3	label-4	label-5
	Breast Cancer Data Set	label-2	label-3	label-4	label-5
ARI	Balloons Data Set	label-2	label-3	label-4	label-5
	Soybean-small Data Set	label-2	label-3	label-4	label-5
	Zoo Data Set	label-2	label-3	label-4	label-5
	Voting Data Set	label-2	label-3	label-4	label-5
	Breast Cancer Data Set	label-2	label-3	label-4	label-5
PR	Balloons Data Set	label-2	label-3	label-4	label-5
	Soybean-small Data Set	label-2	label-3	label-4	label-5
	Zoo Data Set	label-2	label-3	label-4	label-5
	Voting Data Set	label-2	label-3	label-4	label-5
	Breast Cancer Data Set	label-2	label-3	label-4	label-5
RE	Balloons Data Set	label-2	label-3	label-4	label-5
	Soybean-small Data Set	label-2	label-3	label-4	label-5
	Zoo Data Set	label-2	label-3	label-4	label-5
	Voting Data Set	label-2	label-3	label-4	label-5
	Breast Cancer Data Set	label-2	label-3	label-4	label-5

References

[1] J. Liang, L. Bai, C. Dang, F. Cao, The k-means type algorithms versus imbalanced data distributions, IEEE Transactions on Fuzzy Systems 20 (4)

(2012) 728–745.

[2] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, The Journal of Machine Learning Research 3 (2003) 583–617.