A Weighted Coupled Attribute Distance Metric 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

Recently most studies of similarity analysis of categorical data are assume that attributes are mutual independent. However, in real data sets, attributes are more or less associated with certain coupling relationships. In this paper, we propose a weighted dissimilarity method for categorical objects. It involves the frequency-based weighted intra-coupled distance within an attribute and the weighted inter-coupled distance upon value co-occurrences between attributes. we prove the dissimilarity method is metric. The experimental results on UCI data sets show that the distance measure proposed in this paper has a good effect on categorical data.

Keywords: Clustering, Coupled attribute distance, Weighting distance

1. Introduction

In the unsupervised learning, the label of the training objects is unknown. The goal is to reveal the inherent nature and regularity of the data through the learning of the unlabeled training objects, and to provide the basis for further data analysis. The most widely used in these studies is clustering. Clustering analysis is an effective way to obtain the internal structure of data. The basic problem involved in the clustering algorithm is distance calculation, can also be

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

called similarity analysis. The rule between distance and similarity is the larger the distance between the objects, the smaller the similarity between them.

In the study of the similarity between two numerical variables, the distance measure methods include Euclidean and Minkowski distance. For categorical data, similarity measures have generally gone through two stages. Firstly, there are some distance measure methods that not consider coupling, includes SMS, which uses 0s and 1s to distinguish the similarity between distinct and identical categorical values, the occurrence frequency (OF) [4] and the information-theoretical similarity (Lin) [4], to discuss the similarity between categorical values. Secondly, there are some algorithms that only consider partial coupling, such as Can Wang proposed CADO algorithm [5].

The challenge is that these methods are too rough to accurately represent the similarity between categorical attribute values. In addition, none of them provides a comprehensive similarity between categorical attributes by combining relevant aspects. Even if the CADO algorithm considers partial coupling of attributes, the difference in the degree of coupling between attributes is not taken into account.

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. The key contribution are as follows.

- A intra-attribute 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. And it can adjust the contribution of distance along each attribute to the whole object distance. Moreover, the weight between two pairs of attribute values is related to the frequency of occurrence of two attribute values.

- A weighted coupled attribute distance metric between objects is proposed, which based on CADO algorithm [5]. By using the dynamic attribute weight and the relationship between categorical attributes [8] on the basis of the CADO algorithm, the characteristics of the more comprehensive response between objects are adopted.

This paper is organized as follows. In Section 2, we specify preliminary definitions. The intra-attribute weight and inter-attribute weight are given in Section 3. Section 4 proposes the weighted intra-coupled distance, the weighted inter-coupled distance, and their aggregation. We describe the W-CADO algorithm in Section 5. The effectiveness of W-CADO is empirically studied in Section 6. Finally, we conclude this paper in Section 7.

2. Preliminary Definitions

Given the data set $X = \{x_1, x_2, \cdots, x_n\}$ with n objects represented by d categorical attributes $\{A_1, A_2, \cdots, A_d\}$. V_r is the set of attribute values from attribute $A_r (1 \le r \le d)$, v_r is any attribute value of V_r. x_{ir} is the value of object x_i in attribute A_r.

Definition 1 ($p_{A_r}(x_{ir})$): The probability of attribute r in presenting a value equal to x_{ir} in the object set X [6].

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

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

Definition 2 ($p^{-}_{A_r}(x_{ir})$): The estimated probability of attribute r in presenting a value equal to x_{ir} in the object set X [6].

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

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

Definition 3 (ICP): The value subset $\acute{V_r}(\subseteq V_r)$ of attribute A_r, and the value $v_l(\in V_l)$ of attribute A_l, then the information conditional probability (ICP) of $\acute{V_r}$ with respect to v_l is $P_{A_r|A_l}(\acute{V_r}|v_l)$, defined as

$$P_{A_r|A_l}(\acute{V_r}|v_l) = \frac{\sigma_{A_r=\acute{V_r}}(X) \wedge \sigma_{A_l=v_l}(X)}{\sigma_{A_l=v_l}(X)} \quad (3)$$

Intuitively, when given all the objects with the value v_l of attribute A_l, ICP is the percentage of common objects whose values of attribute A_l fall in subset $\acute{V_r}$ and whose values of attribute A_l are exactly v_l as well. Hence, ICP quantifies the relative overlapping ratio of attribute values in terms of objects. for example, $P_{Actor|Genre}(\{Grant\}|Thriller) = 0.5$.

Definition 4: The inter-coupled relative similarity based on intersection set between values x_{ir} and x_{jr} of attribute A_r based on another attribute A_l is defined as

$$\lambda_{A_r|A_l}(x_{ir}, x_{jr}, V_l) = \sum_{v_l \in \cap} min\{P_{A_l|A_r}(\{v_l\}|x_{ir}), P_{A_l|A_r}(\{v_l\}|x_{jr})\} \quad (4)$$

where $v_l \in \cap$ denote $v_l \in$ the intersection of the set of value in attribute A_l corresponds with x_{ir} and the set of value in attribute A_l corresponds with x_{jr}

3. Proposed Weight Metric for Categorical Data

The attribute couplings include intra-attribute couplings and inter-attribute couplings. Below, the weight for intra-attribute and inter-attribute are formalized and exemplified.

Firstly, we illustrate the shortcomings of the CADO algorithm by a real application example.

A real data application example is described in Table 1. As shown in Table 1, six movie objects are divided into two classes with three categorical 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 SMS is that the similarity between Stewart and Grant

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

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 categorical variables than between continuous data. The SMS and its variants fail to capture a global picture of the real relationship for categorical data. Can Wang has put forward an effective algorithm [5] (CADO) to improve the shortcomings of the above algorithms. Wang make a point about the data-driven intra-coupled similarity and inter-coupled similarity, as well as their global aggregation in unsupervised learning on categorical data.

However, the measure of intra-coupled similarity in her method does not greater 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.

3.1. Intra-attribute Weighting

As we know, when we compare two objects, we usually pay more attention to the special features they have. In other words, unusual features generally can provide more information for the comparison between objects. Considering this phenomenon, we can further adjust the distance metric according to following criterion. The contribution of the distance between tow attribute values to the whole object distance is inverse to the probability of these two values' situation in the whole data set. That is, if two data objects have different values along one attribute, the contribution of the distance between these two values to the entire data distance is inverse to the probability that two data objects have different values along this attribute in the data set, and vice versa [8]. What's more, the contribution of distance between two different values of the same attribute should be related to the occurrence probability that two values. We give higher weight to the higher frequency of the attribute. Therefore, this kind of probability can be used as intra-attribute weight.

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

$$p_s(A_r) = \sum_{j=1}^{m_r} p_r(x_{jr}) p_r^-(x_{jr}) \tag{5}$$

For example, $p_s(Actor) = \frac{1}{5}$, $p_s(Director) = \frac{2}{15}$.

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

$$p_f(A_r) = 1 - p_s(A_r) \tag{6}$$

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

$$\eta(A_r) = \begin{cases} p_s(A_r), & if \quad x_{ir} = x_{jr} \\ p_f(A_r), & otherwise \end{cases} \tag{7}$$

In a word, $\eta(A_r)$ show the special features of attribute A_r among attributes.

For the two values x_{ir} and x_{jr} of the attribute A_r, the distance weight between these two values is calculated by

$$\theta(x_{ir}, x_{jr}) = \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} * \frac{\sigma_{A_r = x_{jr}}(X)}{\sigma_{A_r \neq NULL}(X)} \tag{8}$$

Alternatively, we could consider other forms of distance weight between two values of attribute A_r according to the data structure, such as $\theta(x_{ir}, x_{jr}) = \alpha \cdot \frac{\sigma_{A_r=x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} + \gamma \cdot \frac{\sigma_{A_r=x_{jr}}(X)}{\sigma_{A_r \neq NULL}(X)}$, where $0 \leq \alpha, \gamma \leq 1 (\alpha + \gamma = 1)$ are the corresponding weights. $\theta(x_{ir}, x_{jr})$ show the difference between two values of attribute A_r.

Subsequently, the intra-attribute weight is between value x_{ir} and x_{jr} for attribute A_r is

$$\omega(A_r, x_{ir}, x_{jr}) = \eta(A_r) * \theta(x_{ir}, x_{jr}) \tag{9}$$

Alternatively, we could consider other combination forms of η and θ according to the data structure, such as $\omega(A_r, x_{ir}, x_{jr}) = \alpha \cdot \eta(A_r) + \gamma \cdot \theta(x_{ir}, x_{jr})$, where $0 \leq \alpha, \gamma \leq 1(\alpha + \gamma = 1)$ are the corresponding weights. Thus, η and θ can be controlled flexibly to display in which cases the former is more significant than the latter, and vice versa.

3.2. Inter-attribute Weighting

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 [9]. 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 order to utilize the useful relationship information accompanying with each pair of attributes well, the interdependence redundancy measure [7] 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 by the values of other attributes that are highly relevant to this one. In especial, given the data set X, the dependence degree between each pair of attributes A_r and A_l $(r, l \in \{1, 2, \cdots, d\})$ can be quantified based on the mutual information

[10] between them, which is defined as

$$I(A_r;A_l) = \sum_{r=1}^{m_r}\sum_{l=1}^{m_l} p(x_{ir},x_{jl})\log(\frac{p(x_{ir},x_{jl})}{p(x_{ir})p(x_{jl})}) \tag{10}$$

Here, the items $p(x_{ir})$ and $p(x_{jl})$ stand for the frequency probability of the two attribute values in the while data set, which are calculated by

$$p(x_{ir}) = p(A_r = x_{ir}|X) = \frac{\sigma_{A_r=x_{ir}}(X)}{\sigma_{A_r\neq NULL}(X)} \tag{11}$$

$$p(x_{jl}) = p(A_l = x_{jl}|X) = \frac{\sigma_{A_l=x_{jl}}(X)}{\sigma_{A_l\neq NULL}(X)} \tag{12}$$

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

$$p(x_{ir},x_{jl}) = p(A_r = x_{ir} \wedge A_l = x_{jl}|X) = \frac{\sigma_{A_r=x_{ir}}(X) \wedge \sigma_{A_l=x_{jl}}(X)}{\sigma_{A_r\neq NULL}(X) \wedge \sigma_{A_l\neq NULL}(X)} \tag{13}$$

The mutual information between the two attributes actually measures the average reduction in the uncertainty of an attribute by learning the value of another attribute. A larger value of mutual information usually indicates a greater dependency. However, the 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. [7] proposed to normalize the mutual information with a joint entropy, which yields the interdependence redundancy measure denoted as

$$R(A_r;A_l) = \frac{I(A_r;A_l)}{H(A_r;A_l)} \tag{14}$$

where the joint entropy $H(A_r, A_l)$ is calculated by

$$H(A_r;A_l) = -\sum_{r=1}^{m_r}\sum_{l=1}^{m_l} p(x_{ir},x_{jl})\log(p(x_{ir},x_{jl})) \tag{15}$$

This interdependence redundancy measure evaluates the degree of deviation from independence between two attributes [7]. In particular, $R(A_r; A_l) = 1$ means that the attributes A_r and A_l are strictly dependent on each other while $R(A_r; A_l) = 0$ indicates that they are statistically independent. If the value of

$R(A_r; A_l)$ 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 ξ to store the dependence degree of each pair of attributes [8]. Each element $\xi(r, l)$ of this matrix is given by $\xi(r, l) = R(A_r; A_l)$. It is obvious that ξ 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 [8]. In particular, the set S_r is constructed by

$$S_r = \{A_l | R(A_r; A_l) > \beta, 1 \leq l \leq d\} \tag{16}$$

where β is a specific threshold.

4. Coupled Attribute Distance

4.1. The Weighted Intra-Coupled Interaction

According to CADO algorithm [5], the intra-coupled attribute similarity for values (IaASV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_{A_r}^{IaASV}(x_{ir}, x_{jr}) = \frac{\sigma_{A_r=x_{ir}}(X) \cdot \sigma_{A_r=x_{jr}}(X)}{\sigma_{A_r=x_{ir}}(X) + \sigma_{A_r=x_{jr}}(X) + \sigma_{A_r=x_{ir}}(X) \cdot \sigma_{A_r=x_{jr}}(X)} \tag{17}$$

For example, in Table 1, we have $\delta_{Actor}^{Ia}(Stewart, DeNiro) = \delta_{Actor}^{Ia}(DeNiro, DeNiro) =$ 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 above shortcomings of CADO algorithm, we use the dynamic attribute weight we just described in Section 3. Subsequently, the weight intra-coupled attribute distance for values (W-IaADV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_{A_r}^{W-IaADV}(x_{ir}, x_{jr}) = \omega(A_r, x_{ir}, x_{jr}) * (\frac{1}{IaASV} - 1) \tag{18}$$

4.2. The Weighted 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

$$\delta_{A_r}^{IeASV}(x_{ir}, x_{jr}, \{V_l\}_{l \neq r}) = \sum_{l=1, l \neq r}^{d} \alpha_l \lambda_{A_r|A_l}(x_{ir}, x_{jr}, V_l) \tag{19}$$

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 weighted inter-coupled attribute similarity for values (W-IeASV) between values x_{ir} and x_{jr} for attribute A_r is

$$\delta_{A_r}^{W-IeASV}(x_{ir}, x_{jr}, \{V_l\}_{l \neq r}) = \sum_{l=1, l \neq r}^{d} \xi(r, l) \lambda_{A_r|A_l}(x_{ir}, x_{jr}, V_l) \tag{20}$$

In order to make distance measure satisfy the object itself to its own distance is zero, the weighted inter-coupled attribute distance for values (W-IeADV) between values x_{ir} and x_{jr} for attribute A_r, that is, the convert between similarity and dissimilarity measure, is

$$\delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_l\}_{l \neq r}) = \sum_{l=1, l \neq r}^{d} \xi(r, l) - \delta_{A_r}^{W-IeASV} \tag{21}$$

4.3. Coupled Interaction

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

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

$$\delta_{A_r}^{W-CADV}(x_{ir}, x_{jr}, \{V_l\}_{l=1}^{d}) = \delta_{A_r}^{W-IaADV}(x_{ir}, x_{jr}) \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_l\}_{l \neq r}) \tag{22}$$

where $V_l(l \neq r)$ is a value set of attribute A_l different from A_r to enable the weight inter-coupled interaction. $\delta_{A_r}^{W-IaADV}$ and $\delta_{A_r}^{W-IeADV}$ are W-IaADV and W-IeADV.

As indicated in Eq.(22), we choose the multiplication of these two components. W-IaADV is associated with how often the value occurs, while W-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 W-IaADV and W-IeADV according to the data structure, such as $\delta_{A_r}^{W-CADV}(x_{ir}, x_{jr}, \{V_k\}_{k=1}^{d}) = \alpha \cdot \delta_{A_r}^{W-IaADV}(x_{ir}, x_{jr}) + \gamma \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_k\}_{k \neq j})$, where $0 \leq \alpha, \gamma \leq 1 (\alpha + \gamma = 1)$ are the corresponding weights. Thus, W-IaADV and W-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. Weighted Coupled Attribute Distance Algorithm

In previous sections, we have discussed the construction of W-CADV. In this section, a weighted coupled attribute distance between objects (W-CADO) is built based on W-CADV.

Given the data set X, the W-CADO between object x_i and x_j is

$$W-CADO(x_i, x_j) = \sum_{r=1}^{d} \delta_{A_r}^{W-CADV}(x_{ir}, x_{jr}, \{V_k\}_{k=1}^{d}) \tag{23}$$

We can prove that the dissimilarity measure $W-CADO(\cdot,\cdot)$ is a distance metric satisfying three properties as follows.

1) Nonnegativity: $W-CADO(x_i, x_j) \geq 0$ and $W-CADO(x_i, x_i) = 0$;

2) Symmetry: $W-CADO(x_i, x_j) = W-CADO(x_j, x_i)$;

3) Triangle inequality: $W-CADO(x_i, x_j) + W-CADO(x_j, x_k) \geq W-CADO(x_i, x_k)$.

Obviously, we can easily prove the first two properties according to the previous description. The triangle inequality as the third property is verified as follows.

Proof 1. *To prove the inequality*

$$W-CADO(x_i, x_j) + W-CADO(x_j, x_k) \geq W-CADO(x_i, x_k),$$

we only need to demonstrate

$$\sum_{r=1}^{d} \delta_{A_r}^{W-CADV}(x_{ir}, x_{jr}, \{V_l\}_{l=1}^{d}) + \sum_{r=1}^{d} \delta_{A_r}^{W-CADV}(x_{jr}, x_{kr}, \{V_l\}_{l=1}^{d}) \geq \sum_{r=1}^{d} \delta_{A_r}^{W-CADV}(x_{ir}, x_{kr}, \{V_l\}_{l=1}^{d}).$$

With Eq.(24), the inequality above can be rewritten as

$$\begin{aligned}
&\textstyle\sum_{r=1}^{d}(\delta_{A_r}^{W-IaADV}(x_{ir}, x_{jr}) \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_l\}_{l\neq r})) \\
&+ \textstyle\sum_{r=1}^{d}(\delta_{A_r}^{W-IaADV}(x_{jr}, x_{kr}) \cdot \delta_{A_r}^{W-IeADV}(x_{jr}, x_{kr}, \{V_l\}_{l\neq r})) \\
&= \textstyle\sum_{r=1}^{d}((\frac{1}{\sigma_{A_r=x_{ir}}(X)} + \frac{1}{\sigma_{A_r=x_{jr}}(X)}) \cdot \omega(A_r, x_{ir}, x_{jr}) \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_l\}_{l\neq r})) \\
&+ \textstyle\sum_{r=1}^{d}((\frac{1}{\sigma_{A_r=x_{jr}}(X)} + \frac{1}{\sigma_{A_r=x_{kr}}(X)}) \cdot \omega(A_r, x_{jr}, x_{kr}) \cdot \delta_{A_r}^{W-IeADV}(x_{jr}, x_{kr}, \{V_l\}_{l\neq r})) \\
&\geq \textstyle\sum_{r=1}^{d}((\frac{1}{\sigma_{A_r=x_{ir}}(X)} + \frac{1}{\sigma_{A_r=x_{kr}}(X)}) \cdot \omega(A_r, x_{ir}, x_{kr}) \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{kr}, \{V_l\}_{l\neq r})) \\
&= \textstyle\sum_{r=1}^{d}(\delta_{A_r}^{W-IaADV}(x_{ir}, x_{kr}) \cdot \delta_{A_r}^{W-IeADV}(x_{ir}, x_{kr}, \{V_l\}_{l\neq r})) \\
&= \textstyle\sum_{r=1}^{d} \delta_{A_r}^{W-CADV}(x_{ir}, x_{kr}, \{V_l\}_{l=1}^{d})
\end{aligned}$$

The above proof verifies that the triangle inequality property holds on all attribute. It follows that we have $W-CADO(x_i, x_j) + W-CADO(x_j, x_k) \geq$

Algorithm 1 Weighted Coupled Attribute Distance for Objects (W-CADO)

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.(5) and Eq.(6).
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.(14).
Construct the relationship matrix ξ.
Get the index set S_r for each attribute A_r by $S_r = \{l|\xi(r, l) > \beta, 1 \leq l \leq d\}$.
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**
  for every value pair $(x_{ir}, x_{jr} \in [1, \sigma_{A_r}])$ **do**
    // Compute the weight intra-coupled distance for two attribute values x_{ir} and x_{jr}
    W-IaADV $= \delta_{A_r}^{W-IaADV}(x_{ir}, x_{jr})$;
    //Compute the weight inter-coupled distance for two attribute values x_{ir} and x_{jr}
    W-IeADV $= \delta_{A_r}^{W-IeADV}(x_{ir}, x_{jr}, \{V_l\}_{l \neq r})$;
    //Compute coupled distance between two attribute values x_{ir} and x_{jr}
    W-CADV = W-IaADV $\cdot$ W-IeADV;
    //Compute coupled distance between two objects x_i and x_j
    W-CADO = sum(W-CADV);
  end for
end for
$D(x_i, x_j) =$ W-CADO;
return $D(x_i, x_j)$;

$W - CADO(x_i, x_k)$. Therefore, the dissimilarity measure $W - CADO(\cdot, \cdot)$ is a distance metric.

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

- Compute the W-IaADV for attributes (x_{ir} and x_{jr}) of attribute A_r;
- Compute the W-IeADV for attribute values (x_{ir} and x_{jr});
- Compute the W-CADV for attribute values (x_{ir} and x_{jr});
- Compute the W-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 W-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 [8]. That is, we let $\beta = \beta_0$, where β_0 is calculated by

$$\beta = \frac{1}{d^2} \sum_{r=1}^{d} \sum_{l=1}^{d} R(A_r; A_l), \tag{24}$$

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 object 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 W-CADO algorithm, we used the following five external criterions: (1) adjusted rand index (ARI) [12], (2) normalized mutual information (NMI) [13], (3) accuracy (AC), (4) precision (PR) and (5) recall (RE) to compare the obtained cluster of each object with that provided by data label.

As described in the Section 2, X represents a data set, $C = \{C_1, C_2, \cdots, C_k'\}$ be a clustering result of X, $P = \{P_1, P_2, \cdots, P_k\}$ be a real partition in 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:

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

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

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

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

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

where $n_{1j_1^*} + n_{2j_2^*} + \cdots + n_{kj_k^*} = \max_{j_1 j_2 \cdots j_k \in S} \sum_{i=1}^{k} n_{i j_i} \quad (j_1^* j_2^* \cdots j_k^* \in S)$ 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 W-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 distance matrix to perform dimensionality reduction for clustering before the k-means algorithm. Below, we aim to compare the performance of W-CADO against CADO as used in data cluster analysis for further clustering evaluation.

In the following tables report 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 50 tests for KM and SC with distinct start points.

Note that the highest measure score of each experimental setting is highlighted in boldface.

Table 3: Comparison on Balloons Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
K-Mode	CADO	0.73	0.3283	0.2280	0.7783	0.8417
	W-CADO	**0.76**	**0.3999**	**0.2943**	**0.81**	**0.8333**
SC	CADO	0.92	0.8404	0.7986	0.95	0.9333
	W-CADO	**0.96**	**0.9202**	**0.8993**	**0.9750**	**0.9667**

Table 4: Comparison on Soybean-small Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
K-Mode	CADO	0.7	0.6325	0.4422	0.8086	0.6784
	W-CADO	**0.7993**	**0.7509**	**0.6465**	**0.88030**	**0.7842**
SC	CADO	0.9894	0.9895	0.9797	0.9954	0.9875
	W-CADO	**1**	**1**	**1**	**1**	**1**

Table 5: Comparison on Zoo Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
K-Mode	CADO	0.7743	0.5113	0.4820	0.7963	0.5764
	W-CADO	**0.8158**	**0.5623**	**0.6570**	**0.8423**	**0.5764**
SC	CADO	0.8574	0.8158	0.7495	0.8335	0.7333
	W-CADO	**0.8693**	**0.7890**	**0.7334**	**0.8745**	**0.7446**

Table 6: Comparison on Congressional Voting Records Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
K-Mode	CADO	0.7621	0.2675	0.3011	0.7703	0.7375
	W-CADO	**0.8336**	**0.3869**	**0.4526**	**0.8387**	**0.8369**
SC	CADO	0.8782	0.4895	0.5710	0.8717	0.8897
	W-CADO	**0.8805**	**0.4994**	**0.5780**	**0.8743**	**0.8927**

As table listed above indicates, the clustering methods with W-CADO,

Table 7: Comparison on Breast Cancer Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
K-Mode	CADO	0.7497	0.2010	0.2191	0.8032	0.6516
	W-CADO	**0.8550**	**0.4879**	**0.5351**	**0.8514**	**0.8570**
SC	CADO	0.9399	0.6956	0.7729	0.9260	0.9512
	W-CADO	**0.9456**	**0.7126**	**0.7907**	**0.9276**	**0.9667**

whether KM or SC, outperform those with CADO on both AC, NMI, PR, RE and ARI. The reason is that the weight of the attribute added in our algorithm improves the similarity between similar objects and the differences between different classes of objects. Moreover, the consideration of a complete inter-coupled interaction leads to the largest improvement on clustering accuracy.

For K-Mode, the AC improving rate ranges from 4.0% (Balloons) to 14.2% (Soybean-small). With regard to SC, the AC rate takes the minimal and maximal radios as 0.61% (Breast Cancer) and 4.3% (Balloons). In short, it can be seen that the W-CADO algorithm is exactly better than the CADO algorithm. There is a significant observation that SC mostly outperforms K-Mode whenever it has the same distance metric. This is consistent with the finding in [11], indicating that SC very often outperforms k-means for numerical data.

7. Conclusion

We have proposed W-CADO, a weighted coupled attribute distance measure for objects incorporating both weighted intra-coupled attribute distance for values and weighted inter-coupled attribute distance for values based on CADO algorithm. By using the intra-attribute weight, the measure increase the intra-class aggregation and inter-class dissimilarity. Furthermore, the dependence degree between each pair of attribute is showed by the inter-attribute weight. Since consider inter-coupled interaction, W-CADO algorithm have improved the clustering accuracy largely. Experimental results on the five real data sets have shown that the W-CADO algorithm is better than the CADO algorithms in clustering categorical data.

References

[1] L.Wang, Y.Ou, P.S.Yu, Coupled behavior analysis with applications, IEEE Transactions on Fuzzy Systems 24 (8) (2012) 1378–1392.

[2] F.Figheiredo, L.Rocha, T.Couto, T.Salles, Word co-occurrence features for text classification, Information System 36 (5) (2011) 843–858.

[3] G.Wang, D.Hoiem, D.Forsyth, Learning image similarity from Flickr groups using fast kernel machines, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2177–2188.

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

[5] C.Wang, X. Dong, F. Zhou, L. Cao, Coupled Attribute Similarity Learning on Categorical Data, IEEE Transactions on Neural Network and Learning System 26 (4) (2015) 781–797.

[6] Tiago.R.L., dos.Santos, Luis E.Z, Categorical data clustering:What similarity measure to recommend, Expert System with Applications 42(2015) 1247-1260.

[7] Wai-Ho Au, K.C.C.Chan, A.K.C.Wong, Yang Wang, Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, IEEE Transactions on Computational Biology and Bioinformatics 2 (2) (2005) 83–100

[8] H.Jia, Y. Cheung, A New Distance Metric for Unsupervised Learning of Categorical Data, IEEE Transactions on Neural Network and Learning System 27 (5) (2016) 1065–1079.

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

[10] D.J.C.MacKay, Information Theory, Inference, and Learning Algorithms. Cambridge, U.K.: Cambridge Univ.Press, 2003

[11] U.Von Luxburg, A tutorial on spectral clustering, Statistics and Computing, 17 (4) (2007) 395–416

[12] 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.

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