A weighted distance metric for clustering 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

Measuring the distance or dissimilarity between two objects is a very difficult problem in clustering categorical data sets. The dissimilarity of categorical objects can be divided into three categories: attribute-irrelevance-based, attribute values co-occurrence-based and attribute coupling-based. We consider that each attribute has different importance for the contribution of dissimilarity between two objects. In this paper, we define a weighted dissimilarity measure between two 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. Then we prove the dissimilarity measure is a metric and give an algorithm AW-CADO. The experimental results on categorical data have shown the good effect of the proposed distance measure , comparing to other measures on several UCI data sets.

Keywords: Clustering, Coupled attribute distance, Weighting metric

1. Introduction

The aim of clustering is to divide data objects into homogeneous groups so that objects in the same group are similar and objects in different groups are different [**?**]. In clustering algorithms, how to define the dissimilarity between

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

objects is a very difficult problem, because different dissimilarity measures often result in different clustering results. For numerical data, there are many measures to compute the dissimilarity between two objects, such as Euclidean, Minkowski and Mahalanobis distance. As categorical data lacks of inherent geometrical property, distance measures for computing numerical data can not be applied to categorical data sets. To compute the dissimilarity between two categorical data, many measures have been defined. These measures can be divided into three categories:

- Attribute irrelevance-based measures: It is assumed that different attributes are mutual independent in most clustering algorithms. For example, simple matching similarity (abbr. SMS) uses 0s and 1s to distinguish the similarity between distinct and identical categorical values. The occurrence frequency similarity (abbr. OF) [4] is assigned with value equal to 1 when a match occurs. In the case of a mismatch, a real value that is greater than zero is assigned to the similarity. The information-theoretical similarity (abbr. Lin) [4] gives higher weight to matches on frequent values, and lower weight to mismatches on infrequent values.

- Attribute values co-occurrence-based measures: The dissimilarity measure between two categorical values of the same attribute depending upon co-occurrence probabilities of these two values with respect to every other attributes of data set. The representative measure is ALGO measure [1] proposed by Ahmad.

- Attribute coupling-based measures: They take into consideration both the frequency-based intra-attribute dissimilarity within an attribute and the inter-attribute dissimilarity between attributes. For example, Wang proposed the coupled attribute dissimilarity for objects (CADO) measure [12].

Below, we take a real example to compute the dissimilarity between two categorical objects to reveal the shortcomings of the above measures. A real data application example is described in Table 1.

Table 1: Objects 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 with three categorical attributes are divided into two classes. For movies Vertigo and Bishop's Wife, they similarity is 0 according to SMS measure, but Vertigo and Bishop's Wife are very similar because they belong to the same class L2. Another example by SMS is that the similarity between movies Vertigo and Bishop's Wife is equal to that between movies Vertigo and Good Fellas. However, the similarity of the former pair should be greater because Vertigo and Bishop's Wife belong to the same class L2. Similarly, the ALGO DISTANCE measure between movies Vertigo and Bishop's Wife is 0, which is equivalent to the distance between movies Vertigo and Good Fellas. This does not agree with the fact of Vertigo and Bishop's Wife belong to the same class L2, Vertigo and Good Fellas belong to different classes.

The above examples show that it is highly difficult to analyse the similarity between categorical variables. The SMS and the ALGO DISTANCE fail to capture a global picture of the real relationship for categorical data [12].

Wang put forward an effective algorithm (CADO) to improve the shortcomings of the above measures. Wang made 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 [12].

However, the measure of intra-coupled dissimilarity in CADO does not greatly show the high intra-class similarity and the low inter-class similarity. Moreover, the compute method of weight parameter for attributes is not given in the CADO algorithm. For example, the CADO measure similarity between Godfather II's De Niro and Good Fellas's De Niro is 0.5 and so is the similarity between Good Fellas's De Niro and Vertigo's Stewart. However, since both actor of the former pair belong to the same class L1, so the similarity should be greater.

The shortcomings of the above measures are obvious and we summarize them as three aspects.

- The difference of coupling between attributes is not taken into account. The CADO measure considers the inter-attribute coupling, but all of the weights for attributes are assumed to be the same.

- The difference of weight for an attribute is not considered in these measures when the dissimilarity or similarity between objects is computed. In real data sets, different attributes have their own features, so the contribution of the dissimilarity or similarity among attributes is different.

- For some measures, the intra-attribute dissimilarity between value pairs is independent of the probability of occurrence. For example, SMS and ALGO DISTANCE assume that the value pairs are mutual independent within an attribute.

To sum up, these methods have certain challenges to accurately reflect the dissimilarity or similarity on categorical data. The distance metric proposed in this paper aims at improving the above shortcomings.

In this paper, we explicitly define a measure on categorical data. The measure takes into account the characteristics of the categorical values, including the frequency probability within an attribute, the coupling inter-attribute and the feature of every attribute. The key contributions are as follows.

- An intra-attribute weighting scheme for categorical attributes is presented, which assigns different weight according to the different distribution of each attribute's values. The intra-attribute weighting not only takes into account the weight between different attributes, but also takes into consideration the occurrence frequency of the value pairs within an attribute.

- A weighted coupled attribute distance metric between objects (W-CADO) is proposed, which is based on CADO [12]. By using the intra-attribute weight and the inter-attribute weight [6] in the distance calculation, the comprehensive characteristics between objects are revealed.

This paper is organized as follows. In Section 2, we review some related definitions. The intra-attribute weight and inter-attribute weight are given in Section 3. Section 4 define the weighted intra-coupled distance, the weighted inter-coupled distance, and their integration. 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. Related Definitions

Given a data set $X = \{x_1, x_2, \ldots, x_n\}$ with n objects represented by d categorical attributes $\{A_1, A_2, \ldots, A_d\}$. Suppose that V_r is a set of attribute values from the attribute $A_r (1 \leq r \leq d)$ with m_r possible values and x_{ir} is the value of the object x_i in the attribute A_r. Obviously, $x_{ir} \in V_r$. $f = \cup_{r=1}^{d} f_r$,$f_r : X \rightarrow V_r (1 \leq r \leq d)$ is an information function that assigns a particular value of attribute A_r to every object.

Definition 1. *[10] ($p_{A_r}(x_{ir})$): For any $x_i \in X$, the probability of x_{ir} in the attribute A_r is defined as*

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

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

Definition 2. *[10] ($p^{-}_{A_r}(x_{ir})$): The estimated probability of attribute A_r in presenting a value equal to x_{ir} in X is defined as*

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

Obviously, we can obtain $p_{Actor}(Stewart) = \frac{1}{3}$, $p^{-}_{Actor}(Stewart) = \frac{1}{5}$ from Table 1.

Definition 3. *[12] (ICP): Given a subset $V'_r \subseteq V_r$ having m'_r possible values in the attribute A_r, and a value $v_l \in V_l$ in the attribute A_l, then the information conditional probability (ICP) of V'_r with respect to v_l is defined as*

$$P_{A_r|A_l}(V'_r|v_l) = \frac{\sum_{v_r \in V'_r} \sigma_{A_r=v_r \wedge A_l=v_l}(X)}{\sigma_{A_l=v_l}(X)}. \quad (3)$$

Intuitively, when given all the objects with the value v_l in the attribute A_l, ICP is the percentage of common objects whose values of attribute A_r fall in subset 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. *[12] (SIF): Two set information functions are defined as*

$$F_r : 2^X \rightarrow 2^{V_r}, F_r(X') = \{f_r(x_i)|x_i \in X'\} \quad (4)$$

$$G_r : 2^{V_r} \rightarrow 2^X, G_r(V'_r) = \{x_i|f_r(x_i) \in V'_r\} \quad (5)$$

where $1 \leq r \leq d, 1 \leq i \leq n, X' \subseteq X$, and $V'_r \subseteq V_r$.

These SIFs describe the relationships between objects and attribute values. Function $F_r(X')$ maps the object set X' to the associated value set of attribute A_r. Function $G_r(V'_r)$ maps the value set V'_r of attribute A_r to the dependent object set. In Table 1, $G_3(\{Hitchcock\}) = \{Vertigo, NbyNW\}$, $F_1(\{Vertigo, NbyNW\}) = \{Stewart, Grant\}$.

Definition 5. *[12] 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*

$$\lambda_{A_r|A_l}(x_{ir}, x_{jr}) = \sum_{v_l \in \varkappa} min\{P_{A_l|A_r}(\{v_l\}|x_{ir}), P_{A_l|A_r}(\{v_l\}|x_{jr})\}, \quad (6)$$

where $v_l \in \varkappa$ denote $v_l \in F_l(G_r(\{x_{ir}\})) \cap F_l(G_r(\{x_{jr}\}))$. In [12], Wang consider four measures for the inter-coupled similarity to calculate the similarity between two categorical values by considering their relationships with other attributes in terms of power set, universal set, joint set, and intersection set. Wang reveals the equivalent accuracy and superior efficiency of the measure based on the intersection set by theoretical analysis.

3. New weight-computing method for Categorical data

In [12], the attribute couplings include intra-attribute coupling and inter-attribute coupling. Since Wang doesn't consider the difference among attributes and the coupling between attributes, we propose a new weight-computing method. The weight for intra-attribute and inter-attribute are formalized and exemplified below.

3.1. Intra-attribute Weight

Recently most studies of similarity analysis of categorical data treat each attribute equally in data sets. However, it's not always reasonable in real data sets. As we know, unusual features generally can provide more information for the comparison between objects, so when we compare two objects, we usually pay more attention to the special features they have [6]. In other words, different attribute features should have different contributions in the distance calculation, furthermore, different value pairs within an attribute should have different weights. Considering this phenomenon, we can further adjust the weight according to the following criterion.

The contribution of the distance between two 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 greater the probability that two data objects have different values along this attribute in the data set, the less contribution of the distance between these two values to the entire data distance, and vice versa. What's more, distance metric should assign different weights according to different attribute features. According to the situation that two objects have the same or different value on an attribute, we regard the probability that have the same values or different values from this attribute as the weight of the attribute.

For an attribute A_r, the probability that two data objects from X have the same values along A_r is calculated by

$$p_s(A_r) = \sum_{v_r \in V_r} p_r(v_r) p_r^-(v_r). \tag{7}$$

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

$$p_f(A_r) = 1 - p_s(A_r). \tag{8}$$

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

$$\eta(A_r, x_{ir}, x_{jr}) = \begin{cases} p_s(A_r), & if \quad x_{ir} = x_{jr} \\ p_f(A_r), & otherwise \end{cases} \tag{9}$$

In a word, $\eta(A_r, x_{ir}, x_{jr})$ taking different values depend on whether x_{ir} and x_{jr} are equal, shows the special features of attribute A_r among attributes.

For two attribute values x_{ir} and x_{jr}, the weight between them is calculated by

$$\theta(x_{ir}, x_{jr}) = \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} * \frac{\sigma_{A_r = x_{jr}}(X)}{\sigma_{A_r \neq NULL}(X)}. \tag{10}$$

Here,we take OF measure as a reference that mismatches on less frequent values are assigned lower similarity and mismatches on more frequent values are assigned higher similarity [4]. Within a certain range, the bigger $\theta(x_{ir}, x_{jr})$ reflects the higher similarity of x_{ir} and x_{jr}. Alternatively, we could consider other forms of weight between two values of attribute A_r according to the data structure, such as $\theta(x_{ir}, x_{jr}) = \alpha \cdot \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} + \gamma \cdot \frac{\sigma_{A_r = x_{jr}}(X)}{\sigma_{A_r \neq NULL}(X)}$ or $\theta(x_{ir}, x_{jr}) = -\alpha \cdot \log \frac{\sigma_{A_r = x_{ir}}(X)}{\sigma_{A_r \neq NULL}(X)} - \gamma \cdot \log \frac{\sigma_{A_r = x_{jr}}(X)}{\sigma_{A_r \neq NULL}(X)}$, where $0 \leq \alpha, \gamma \leq 1 (\alpha + \gamma = 1)$ are the corresponding weights. $\theta(x_{ir}, x_{jr})$ show the weight between two values of attribute A_r and represent the distance between these value pair indirectly.

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

$$\omega(A_r, x_{ir}, x_{jr}) = \eta(A_r, x_{ir}, x_{jr}) * \theta(x_{ir}, x_{jr}). \tag{11}$$

Alternatively, we could consider other combination forms of η and θ according to the data structure, such as $\omega(A_r, x_{ir}, x_{jr}) = \alpha \cdot \eta(A_r, x_{ir}, x_{jr}) + \gamma \cdot \theta(x_{ir}, x_{jr})$, 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.

The intra-attribute weighting can adjust the contribution of distance along each attribute to the whole object distance. Moreover, the weight reflects the distance between different value pairs within an attribute on the basis of the occurrence frequency.

3.2. Inter-attribute Weight

Most existing distance or similarity measures for categorical data assume that each attribute is independent. However, in real data, we often have some attributes that are highly dependent on each other. Therefore, the computation of similarity or distance for categorical attribute should be considered based on frequently co-occurring items [5]. 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 [2] has been introduced to evaluate the dependence degree between different attributes. Subsequently, the distance between two values from one attribute is measured not only by their own frequency probabilities but also by the values of other attributes that are highly relevant to this one. In particular, given the data set X, the dependence degree between each pair of attributes A_r and A_l $(r, l \in \{1, 2, \ldots, d\})$ can be quantified based on the mutual information [9] between them, which is defined as

$$I(A_r; A_l) = \sum_{v_r \in V_r} \sum_{v_l \in V_l} p(v_r, v_l) \log(\frac{p(v_r, v_l)}{p_{A_r}(v_r) p_{A_l}(v_l)}). \tag{12}$$

Here, the items $p_{A_r}(v_r)$ and $p_{A_l}(v_l)$ stand for the frequency probability of the two attribute values in the while data set.

The expression $p(v_r, v_l)$ is to calculate the joint probability of these two attribute values, i.e., the frequency probability of objects in X having $A_r = v_r$ and $A_l = v_l$, which is given by

$$p(v_r, v_l) = p(A_r = v_r \wedge A_l = v_l | X) = \frac{\sigma_{A_r = v_r \wedge A_l = v_l}(X)}{\sigma_{A_r \neq NULL \wedge A_l \neq NULL}(X)}. \tag{13}$$

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 [9]. 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. [2] 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

$$H(A_r; A_l) = -\sum_{v_r \in V_r} \sum_{v_l \in V_l} p(v_r, v_l) \log(p(v_r, v_l)). \tag{15}$$

This interdependence redundancy measure evaluates the degree of deviation from independence between two attributes [2]. 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. 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 [6]. 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 Distance

According to CADO algorithm [12], 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 $\delta_{Actor}^{IaASV}(Stewart, DeNiro) = \delta_{Actor}^{IaASV}(DeNiro, DeNiro) = 0.5$ 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 considers $h_1(t) = 1/t - 1$ to reflect the complementarity between similarity and dissimilarity measures. We adopt the same complementarity in the algorithm proposed in this paper. In order to overcome the above shortcomings of CADO algorithm, we use the intra-attribute weight that is described in Section 3 to calculate the inter-coupled distance. Subsequently, the weight intra-coupled attribute distance for values (W-IaADV) between values x_{ir} and x_{jr} for attribute A_r is

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

For example, $\delta_{Actor}^{W-IaADV}(DeNiro, DeNiro) = \frac{1}{45}$, $\delta_{Actor}^{W-IaADV}(DeNiro, Stewart) = \frac{4}{45}$. They correspond to the fact that the distance between Good Fellas's De Niro and Vertigo's Stewart is larger than the distance between Godfather II's De Niro and Good Fellas's De Niro.

4.2. The Weighted Inter-Coupled Distance

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

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

where α_l is the weight parameter for attribute A_l. In CADO algorithm, Wang assign $\alpha_l = \frac{1}{d-1}$. 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 attributes. To overcome the shortcomings of CADO algorithm, we use the relationship matrix that is 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 meet the reflexivity, 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. Coupling Integration

So far, we have built 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 the frequency of the value, 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.(23), 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 $W-CADO(x_i,x_j)+W-CADO(x_j,x_k)\geq W-CADO(x_i,x_k)$. Therefore, the dissimilarity measure $W-CADO(\cdot,\cdot)$ is a distance metric.

We then design the AW-CADO algorithm, given in Algorithm 1, to compute the coupled object distance.

Algorithm 1 An algorithm based W-CADO for computing the distance between two Objects (AW-CADO)

Input: data set $X = \{x_1, x_2, \ldots, x_n\}$.
Output: $D(x_i, x_j)$ for $i, j \in \{1, 2, \ldots, n\}$.
Calculate $p_s(A_r)$ and $p_f(A_r)$ for each attribute A_r according to Eq.(7) and Eq.(8).
For each pair of attributes $(A_r, A_l)(r, l \in \{1, 2, \ldots, d\})$ calculate $R(A_r; A_l)$ according to Eq.(16).
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* d **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
$D(x_i, x_j)$ = W-CADO;
return $D(x_i, x_j)$;

6. Experiments on real data sets

To validate the effectiveness of W-CADO, we mainly conduct 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 [3]. We firstly describe the basic information of the five data sets. Then five evaluation indexes are introduced. Finally, we show the comparison results of the AW-CADO algorithm with the CADO algorithm.

6.1. Data sets

The information of the data sets we utilized is shown in Table 2.

Table 2: Data sets used in experiments

Data Sets	*Objects*	*Attributes*	k
Balloons Data Set	20	4	2
Soybean-small Data Set	47	35	4
Zoo Data Set	101	16	7
Congressional Voting Records Data Set	435	16	2
Breast Cancer Data Set	699	10	2

6.2. Evaluation indexes

To evaluate the effectiveness of the AW-CADO algorithm, we used the following five external criterions: (1) adjusted rand index (ARI) [7], (2) normalized mutual information (NMI) [11], (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, $C = \{C_1, C_2, \ldots, C'_k\}$ 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 3, 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.

The five evaluation indexes are defined as follows:

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

Table 3: The contingency table.

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

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

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

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

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

where $n_{1j_1^*} + n_{2j_2^*} + \ldots + n_{kj_k^*} = \max_{j_1 j_2 \ldots j_k \in S} \sum_{i=1}^{k} n_{ij_i} \quad (j_1^* j_2^* \ldots j_k^* \in S)$ and $S = \{j_1 j_2 \ldots j_k : j_1, j_2, \ldots, j_k \in \{1, 2, \ldots, k\},\ j_i \neq j_t \text{ for } i \neq t\ \}$ is a set of all permutations of $1, 2, \ldots, k$. 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 AW-CADO Alogrithm

In our experiments, the value of the threshold parameter β in the proposed metric is set equal to the average interdependence redundancy of all attribute pairs [6]. That is, β is calculated by

$$\beta = \frac{1}{d^2} \sum_{r=1}^{d} \sum_{l=1}^{d} \xi(r, l). \tag{24}$$

One of the clustering approaches is the k-Mode algorithm, designed to cluster categorical data sets. The main idea of k-Mode 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), 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.

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 k-Mode and SC with distinct start points. Note that the highest measure score of each experimental setting is highlighted in boldface.

Table 4: The Comparison of result on Balloons Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
k-Mode	CADO	0.7300	0.3283	0.2280	0.7783	0.8417
	W-CADO	**0.7600**	**0.3999**	**0.2943**	**0.8100**	**0.8333**
SC	CADO	0.9200	0.8404	0.7986	0.9500	0.9333
	W-CADO	**0.9600**	**0.9202**	**0.8993**	**0.9750**	**0.9667**

Table 5: The Comparison of result on Soybean-small Data Set

	Algorithm	*AC*	*NMI*	*ARI*	*PR*	*RE*
k-Mode	CADO	0.7000	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	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

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

Table 6: The Comparison of result 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 7: The Comparison of result 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**

Table 8: The Comparison of result 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**

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 AW-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 [8], indicating that SC very often outperforms k-means for numerical data.

7. Conclusions

We have proposed W-CADO, a weighted coupled attribute distance metric 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 increases 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 considering inter-coupled interaction, AW-CADO algorithm has improved the clustering accuracy largely. Experimental results on the five real data sets have shown that the AW-CADO algorithm is better than the CADO algorithms in clustering categorical data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (under grants 61573229, 61473194, 61432011 and U1435212), the Natural Science Foundation of Shanxi Province (under grant 2015011048), the Shanxi Scholarship Council of China (under grant 2016-003) and the National Key Basic Research and Development Program of China (973) (under grant 2013CB329404).

References

[1] Amir Ahmad and Lipika Dey. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*, 28(1):110–118, 2007.

[2] W. H. Au, K. C. Chan, A. K. Wong, and Y. Wang. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):83–101, 2005.

[3] K Bache and M Lichman. Uci machine learning repository [j/ol]. irvine, ca: School of information and computer science, university of california. 2014.

[4] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254. SIAM, 2008.

[5] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. Cactus-clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 73–83. ACM, 1999.

[6] Hong Jia, Yiu-ming Cheung, and Jiming Liu. A new distance metric for unsupervised learning of categorical data. *IEEE transactions on neural networks and learning systems*, 27(5):1065–1079, 2016.

[7] Jiye Liang, Liang Bai, Chuangyin Dang, and Fuyuan Cao. The -means-type algorithms versus imbalanced data distributions. *IEEE Transactions on Fuzzy Systems*, 20(4):728–745, 2012.

[8] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[9] MacKay and C Davidj. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[10] Tiago R. L. Dos Santos and Luis E. Zrate. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3):1247–1260, 2015.

[11] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[12] Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao, and Chi Hung Chi. Coupled attribute similarity learning on categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 26(4):781, 2015.