A space-structure based dissimilarity measure for categorical data

Kevin Alejandro Hernández, D. Cárdenas Peña, and Álvaro A. Orozco
Automatics Research Group, Faculty of Engineering, Universidad Tecnológica de Pereira (UTP), Colombia

ABSTRACT

The development of analysis methods for categorical data begun in 90's decade, and it has been booming in the last years. On the other hand, the performance of many of these methods depends on the used metric. Therefore, determining a dissimilarity measure for categorical data is one of the most attractive and recent challenges in data mining problems. However, several similarity/dissimilarity measures proposed in the literature have drawbacks due to high computational cost, or poor performance. For this reason, we propose a new distance metric for categorical data. We call it: weighted pairing (W-P) based on feature space-structure, where the weights are understood like a degree of contribution of an attribute to the compact cluster structure. The performance of W-P metric was evaluated in the unsupervised learning framework in terms of cluster quality index. We test the W-P in six real categorical datasets downloaded from the public UCI repository, and we make a comparison with the distance metric (DM3) method and hamming metric (H-SBI). Results show that our proposal outperforms DM3 and H-SBI in different experimental configurations. Also, the W-P achieves highest rand index values and a better clustering discriminant than the other methods.

Keywords: Categorical data, Clustering, Dissimilarity, Similarity, Space-structure

Corresponding Author: Kevin Alejandro Hernández, Faculty of Engineering, Universidad Tecnológica de Pereira (UTP), Pereira, Colombia. Email: kevin_loco@utp.edu.co

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

The augment of available datasets provides to the research community new resources to achieve scientific discoveries, optimizing industrial processes and it grants to find relations or characteristic patterns in data [1]. However, there are open issues, for example, determining a dissimilarity measure is one of the most attractive and recent challenges in data mining problems. This is because the performance of many algorithms for clustering, classification, dimensionality reduction, and outliers detection, depends on the metric used to measure similarity/dissimilarity among the data [2]. For this reason, it is convenient to establish an appropriate distance measure for a given data set, instead of using an arbitrary metric.

Choosing a metric for quantitative data (continuous) is relatively simple, since there are several developed metrics such as Euclidean, Cosine, Manhattan, among others. Also, with this type of data it can be used the standard methods of machine learning directly and performing numeric calculations without drawbacks or limitations [3, 4]. While, choosing a metric for categorical data (or nominal) is more complex, due to there is not an intrinsic similarity/dissimilarity measure established for categorical objects [5]. In addition, standard machine learning algorithms can not be applied directly in categorical data, because it is
not appropriate to calculate statistic descriptors (mean, standard deviation, etc) over a dataset with nominal or qualitative variables as if they were quantitative variables [6]. Besides, categorical data is highly overlapped.

This observation has motivated several researchers to work with categorical data, as in the case of [7] who carried out a study on distances for heterogeneous data (databases with mixed quantitative and qualitative variables), based on a supervised learning approach where each sample has additional information about the class to which it belongs. However, this approach can be extended to the unsupervised learning paradigm. Other works such as [8-12] was proposed to binarize the categorical information, resulting in samples that are assigned a 1 or a 0 as indicated by their original qualitative value, to later using a similarity/dissimilarity measures for binary data in cluster algorithms. Nevertheless, far from being a reliable solution these methods have a problem: they are only applicable to categorical databases whose variables have only two possible states, which in general is not the case. Further, these algorithms need to handle a large number of binary attributes when the datasets have features with many more categories, classes or groups. The above in increases the computational cost and memory storage of the algorithm [13].

Alternatively, in the work developed by [14], the authors evaluate the performance of a variety of similarity measures in the literature like overlap, inverse occurrence frequency, occurrence frequency, among others. This is done in the context of outliers detection, but their experiments showed a very poor performance and unstable results with increased standard deviation, and suggest that there is no one best performing similarity measure, and it is necessary to understand how a similarity measure handles the different attributes of categorical datasets [15]. In a recent research [16] was proposed a new metric to measure the distance between categorical type objects, based on the frequency probability of each attribute value in the whole dataset and the degree of dependence among different attributes. However, the probability distribution of the attributes must be taken into account, and the inherent structure of the data in the feature space is not considered [14, 17].

Given the previously pointed out, in this paper, we propose a new approach to determine the similarity/dissimilarity measure between qualitative data based on the number of possible states of a categorical variable, to assign the degree of relevance or degree of contribution to the compact cluster structure of each attribute. We call our method: weighted pairing (W-P) based on feature space-structure. The effectiveness of the proposed W-P metric is demonstrated by performing experiments in real categorical databases obtained from the public UCI machine learning repository [18]. We compare our proposal with the distance metric method (DM3) and hamming (H-SBI). We analyze the performance of the W-P distance metric by embedding it into the framework of the K-modes algorithm, which is the most popular distance-based clustering method for purely categorical data [16, 19], using a centroid initialization method proposed in [20].

2. WEIGHTED PAIRING DISTANCE (W-P)

In this section, we introduce the dissimilarity measure for categorical data in the paradigm of unsupervised learning as a weighted pairing distance learned according to the attribute compactness within the data structure. Let a set of $N$ objects $X = \{x_n; n \in [1, N]\}$, each of them expressed as the vector $x_n = [x_{n1}, \ldots, x_{np}, \ldots, x_{np}]$, where $P$ stands for the number of attributes. For categorical data, the object $n$ at attribute $p$, $x_{np} \in A_p$, takes one value from the unordered, discrete set $C_p = \{c_{pd}; d \in [1, D_p]\}$ of $D_p$ possible values [21]. In order to establish a similarity measure between categorical objects, the simple matching function aggregates the number matching values [3]:

$$v(x_n, x_m) = \frac{1}{P} \sum_{p=1}^{P} \delta(x_{np}, x_{mp})$$

(1)

being $\delta(x_{np}, x_{mp})$ the delta function that equals to 1 if $x_{np} = x_{mp}$ or 0 in otherwise. Equation (1) determines how many attribute values a couple of samples have in common, and $1/P$ is a normalization factor. However, the simple matching lacks of an attribute ranking required for understanding the categorical data [15, 22].

Aiming to overcome above issue, we propose a dissimilarity measure that considers the relevance of each attribute, termed weighted pairing (W-P) distance, as follows:

$$d_w(x_n, x_m) = 1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp})$$

(2)

$$s.t. \sum_{p=1}^{P} w_p = 1$$

(3)

with the normalized relevance weights $w_p \in [0, 1]$ satisfying $\sum_{p=1}^{P} w_p = 1$, and $d_w(x_n, x_m) \in [0, 1]$. 

A space-structure based dissimilarity measure for categorical data (Kevin Alejandro Hernández)
To determine the attribute relevance, we assume that the more possible values an attribute can take, the more dispersed the objects in the feature space, as Figure 1 illustrates. Then, we account for the data compactness in the relevance weights in terms of the attribute cardinality as:

$$w_p = \frac{|C_p|^{-1}}{\sum_{q=1}^{|C_q|} |C_q|^{-1}}$$  \hspace{1cm} (4)

where operator $|\cdot|$ determines the number of objects within a set. Therefore, the most relevant attributes are those contributing to the most to data compactness, since the attribute relevance becomes inversely proportional to its number of possible values.

![Attribute space example for categorical data, with $L_1 = 2$ and $L_2 = 6$](image)

Figure 1. Attribute space example for categorical data, with $L_1 = 2$ and $L_2 = 6$

### 2.1. Weighted pairing distance properties

If the attribute weights satisfy (3), the dissimilarity function in (2) becomes a distance.

**Lemma 1.** $d_w(x_n, x_m)$ is non-negative.

**Proof.** Let $d_w(x_n, x_m) \geq 0$:

$$1 - \sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) \geq 0$$

$$\sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) \geq 1$$

$$\max \sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) = 1$$

which holds if $\delta(\cdot; \cdot) \leq 1$, and $w_p$ satisfies (3).

**Lemma 2.** $d_w(x_n, x_m)$ is symmetric.

**Proof.** Let $d_w(x_n, x_m) = d_w(x_m, x_n)$:

$$1 - \sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) = 1 - \sum_{p=1}^P w_p \delta(x_{np}, x_{mp})$$

$$\sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) = \sum_{p=1}^P w_p \delta(x_{np}, x_{mp})$$

due to $\delta(x_{np}, x_{mp}) = \delta(x_{np}, x_{mp})$ the property is satisfied.

**Lemma 3.** $d_w(x_n, x_m) = 0$ if and only if $x_n = x_m$.

**Proof.** Let $d_w(x_n, x_m) = 0$:

$$1 - \sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) = 0$$

$$\sum_{p=1}^P w_p \delta(x_{np}, x_{mp}) = 1$$
which is the maximum value that the linear combination takes according to Lemma 1. Such a maximum is achieved if and only if \( w_p \geq 0 \) satisfies Equation (3) and \( \delta(x_{np}, x_{mp}) = 1 \) for all \( p \), that is, \( x_n = x_m \).

**Lemma 4.** \( d_w(x_n, x_m) \) satisfies the triangle inequality.

**Proof.** The triangle inequality states that for any three objects \( x_n, x_m, x_o \), it must be satisfied that \( d_w(x_n, x_m) \leq d_w(x_n, x_o) + d_w(x_o, x_m) \). Then:

\[
1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) \leq 1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{op}) + 1 - \sum_{p=1}^{P} w_p \delta(x_{op}, x_{mp})
\]

\[
\sum_{p=1}^{P} w_p \left( \delta(x_{np}, x_{op}) + \delta(x_{op}, x_{mp}) - \delta(x_{np}, x_{mp}) \right) \leq 1
\]

To prove by contradiction, (5) is not satisfied when \( \delta(x_{np}, x_{op}) + \delta(x_{op}, x_{mp}) - \delta(x_{np}, x_{mp}) > 1 \) for any \( p \), that is

\[
\delta(x_{np}, x_{op}) = \delta(x_{op}, x_{mp}) = 1
\]

\[
\delta(x_{np}, x_{mp}) = 0
\]

Equation (6) implies \( x_{np} = x_{op} \) and \( x_{op} = x_{mp} \). Therefore, \( x_{np} = x_{op} = x_{mp} \) yields \( \delta(x_{np}, x_{mp}) = 1 \), which by contradicting (7), proves (5).

### 2.2. Distance implementation

The computation of the proposed W-P distance is a two-step process. Firstly, we need to learn the weights \( w_p \) given a dataset as shown in Algorithm 1. Secondly, given the weights and an object pair, we compute the W-P distance following Algorithm 2.

**Algorithm 1** Distance weights computation

```plaintext
function WP WEIGHT(Dataset X)
  for \( p = 1 \) to \( P \) do
    \( L_p = |e_p| \)
  end for
  \( L = \frac{1}{\sum_{p=1}^{P} L_p} \)
  for \( p = 1 \) to \( P \) do
    \( w_p = \frac{1}{L_p} \)
  end for
end function
```

**Algorithm 2** WP dissimilarity measure

```plaintext
function WP METRIC(\( w_p \), Dataset X, \( \mu \))
  for \( n = 1 \) to \( N \) do
    for \( k = 1 \) to \( K \) do
      for \( p = 1 \) to \( P \) do
        if \( x_{np} = \mu_{kp} \) then
          \( \delta = 1 \)
        else
          \( \delta = 0 \)
        end if
      end for
      \( d_w = 1 - w_p \delta \)
    end for
  end for
end function
```

where \( x_n \) is a 1-by-\( P \) vector containing a single observation, \( \mu \) is an \( K \)-by-\( P \) matrix containing all centroids, \( d_w \) is an \( N \)-by-\( K \) matrix of distances between all observations \( x \) and all centroids \( \mu \).

According to Algorithm 1, W-P demands the computation of \( P \) cardinality values and a scaling factor \( L \), that yields a time cost of \( \sim O(4P) \). Besides, Algorithm 2 verifies an attribute matching \( P \) times for \( NK \) object pair with a complexity of \( \sim O(3NK) \). Therefore, the computational complexity of W-P is linear on the number of attributes, \( O(2P) \).
3. EXPERIMENTAL SETUP

We analyze the performance of the W-P distance in clustering tasks by embedding it into the K-modes algorithm [23] and initialized cluster centroids with [20] method. Table 1 summarizes the details of the six UCI machine learning datasets considered for evaluating the proposed distance [18].

The first dataset, Congressional Voting Records, includes 16 key votes identified by the Congressional Quarterly Almanac of the US House of Representatives, grouped into Democrat or Republican. The Breast Cancer Wisconsin (Original) collection holds 699 samples periodically collected from 1989 to 1991 as the clinical cases of Dr. Wolberg and labeled as benign or malignant. The Mushroom dataset corresponds to 23 species of gilled mushrooms in the Agaricus and Lepiota families. The dataset describes each specimen in terms of its physical characteristics and classifies it as poisonous or edible. Soybean (Small) contains 35 categorical attributes, among nominal and ordered, and four disease classes. The Car Evaluation, a dataset derived from a hierarchical decision model, labels 1728 cars according to six different aspects into unacceptable, acceptable good, and very good quality. Lastly, the Zoo dataset contains sixteen boolean-valued attributes to group specimens into seven different animal classes.

<table>
<thead>
<tr>
<th>Data sets</th>
<th># of samples</th>
<th># of features</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>435</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>WBCD</td>
<td>699</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Mushroom</td>
<td>8124</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Soybean</td>
<td>47</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>Car</td>
<td>1728</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Zoo</td>
<td>101</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>

For each considered dataset, we assess the clustering performance of W-P in a 50 fold bootstrap scheme, and report average and standard deviation for the following cluster quality indexes, the average intracluster/intercluster distance, cluster discrimination index, rand index and normalized mutual information index. We compare our proposal with the state-of-the-art methods, distance metric (DM3) and Hamming metric with support based initialization (H-SBI), methods proposed in [16] and [20] respectively.

Cluster discrimination index (CDI): Given K clusters $X_k \subset X$ with $\bigcup_k X_k$ and $X_k \cap X_{k'} = \emptyset$, the CDI computes the performance according to the average intracluster distances (AID) as [16]:

$$CDI = \frac{1}{K} \sum_{k=1}^{K} \frac{AID(X_k, X_k)}{\sum_{k \neq k'} AID(X_k, X_{k'})}$$

$$AID(X_k, X_{k'}) = \frac{\sum_{x_k \in X_k} \sum_{x_{k'} \in X_{k'}} d_w(x_k, x_{k'})}{N_k N_{k'}}$$

where $N_k$ stands for the cardinality of the k-th cluster. Therefore, $CDI \geq 0$, and the smaller its value, the more distant the clusters and the closer the objects within each cluster. Where $\mu_k$ denotes the centroid of the k-th cluster resulting from the K-modes algorithm, and $\Delta_k$ stands for the average distance between $\mu_k$ and all the k-th cluster objects.

Rand index (RI): is a similarity measure based on the overlap in class agreement, compared to the class disagreement, is defined as [24]:

$$RI = \frac{\text{Tr}(CM)}{N}$$

being $\text{Tr}()$ the trace operator and $CM \in [0, N]^{K \times C}$ corresponds to the permuted confusion matrix between the cluster algorithm output $l_k$ and gold standard label $y_c$.

Normalized mutual information (NMI): The NMI score relies on the shared object membership, is a symmetric measure for the degree of dependency between $l_k$ and $y_c$. Unlike correlation, mutual information also takes higher order dependencies into account [25]:

$$NMI = \frac{\sum_{k} \sum_{c} N_{kc} \log \left( \frac{N_{kc} N_k}{N_c N_{kc}} \right)}{\sqrt{\left( \sum_{k} N_k \log \left( \frac{N_k}{N} \right) \right) \left( \sum_{c} N_c \log \left( \frac{N_c}{N} \right) \right)}}$$

Int J Elec & Comp Eng, Vol. 11, No. 1, February 2021 : 620 - 627
where $N_c$, $N_k$, $N_{kc}$ denote the number of objects labeled as $y_n = c$, the number of objects in the $k$-th cluster, and the number of objects grouped in the $k$-th cluster that belong to $c$-th ground truth label, respectively. NMI is a positive value with a maximum of 1 achieved when the ground-truth and the resulting clustering perfectly match.

4. RESULTS AND DISCUSSION

In order to determine the best metric performance between DM3, H-SBI and W-P, we validate our proposal metric distance in terms of intracluster/intercluster distance, CDI, RI and NMI index in the unsupervised learning framework. The first evaluated index was CDI, a smaller CDI value indicates better discrimination on the cluster structure of the dataset. In the Table 2 we compare our W-P metric with DM3 and H-SBI. We see that for three of the six datasets we obtain better CDI values, this may be a consequence of the allocation of the weights, giving greater relevance to the attributes that make the structure of the data more compact.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>W-P</th>
<th>DM3</th>
<th>H-SBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>0.4286</td>
<td>0.4342</td>
<td>0.4459</td>
</tr>
<tr>
<td>WBCD</td>
<td>0.7330</td>
<td>0.3374</td>
<td>0.7787</td>
</tr>
<tr>
<td>Mushroom</td>
<td>0.6991</td>
<td>0.7150</td>
<td>0.6111</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.2826</td>
<td>0.2086</td>
<td>0.3220</td>
</tr>
<tr>
<td>Car</td>
<td>0.7895</td>
<td>0.7919</td>
<td>0.8152</td>
</tr>
<tr>
<td>Zoo</td>
<td>0.2024</td>
<td>0.2656</td>
<td>0.2112</td>
</tr>
</tbody>
</table>

As we can see that for distance-based clustering on categorical data, the K-modes algorithm with the proposed distance metric has a competitive advantage in terms of clustering rand index, in Table 3 the distance W-P metric obtain a better results in four of six datasets in comparison with DM3 and H-SBI, specifically in the Voting and Mushroom datasets the RI index increase drastically. In addition to making an exhaustive evaluation of the W-P metric introduced in the K-modes clustering algorithm with support based initialization, the NMI index was evaluated. And as can see in Table 4 we obtained better results to DM3 and H-SBI, we exceeded them in four of the six datasets, the results indicate that our proposed distance metric is more appropriate for the unsupervised categorical data analysis.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>W-P</th>
<th>DM3</th>
<th>H-SBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>0.8992±0.0056</td>
<td>0.7823±0.0016</td>
<td>0.8639±0.0181</td>
</tr>
<tr>
<td>WBCD</td>
<td>0.9192±0.0130</td>
<td>0.8827±0.0752</td>
<td>0.4996±0.0737</td>
</tr>
<tr>
<td>Mushroom</td>
<td>0.7662±0.0156</td>
<td>0.6732±0.0880</td>
<td>0.6573±0.0940</td>
</tr>
<tr>
<td>Soybean</td>
<td>0.9861±0.0284</td>
<td>0.9314±0.0758</td>
<td>0.7800±0.1769</td>
</tr>
<tr>
<td>Car</td>
<td>0.4366±0.0148</td>
<td>0.3059±0.0123</td>
<td>0.1670±0.1350</td>
</tr>
<tr>
<td>Zoo</td>
<td>0.7424±0.0425</td>
<td>0.9064±0.0450</td>
<td>0.5224±0.1395</td>
</tr>
</tbody>
</table>

The average intracluster distance of each cluster and the average intercluster distance between each pair of clusters has been presented in Table 5 to Table 10, where we can see that for Voting, WBCD, and Car datasets the W-P intercluster distance increase in comparison with DM3 results, while Mushroom, Soybean, and Zoo datasets the average W-P intracluster distance decrease in comparison with DM3, this is reasonable because by (4) we can deduce that for datasets with larger number of attributes the weights become small, so the distance between samples is short. Moreover, in Voting and Car datasets the difference between the average intercluster and intracluster distance with W-P metric is greater than DM3 metric results, in the rest of datasets, the difference between the average intercluster and intracluster distance is very similar.
In this work, we introduced a new similarity/dissimilarity measure for categorical data based on the feature space structure. This distance metric is a variation of pairing matching but weighted. We call our method weighted pairing (W-P) based on feature space-structure. The weights are determined for the number of states that each feature has, indicating which attribute contributes more to the cluster’s compact structure. The performance of W-P metric was evaluated in terms of intracluster/intercluster distance, CDI, RI, and NMI index into a K-modes algorithm with support-based initialization in the unsupervised learning framework, and we compare with the distance metric (DM3) and H-SBI methods. The obtained results showed a better performance for W-P than DM3 and H-SBI, we demonstrated that this way of computing a distance is effective in recovering the inherent clustering structures from categorical data when such structures exist, and this can be attributed to the fact that our approach is space-structure based.

Table 5. Average intracluster/intercluster distance obtained by the different metrics on the voting data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.3542</td>
<td>0.2909</td>
</tr>
<tr>
<td>C2</td>
<td>0.6380</td>
<td>0.7295</td>
</tr>
</tbody>
</table>

Table 6. Average intracluster/intercluster distance obtained by the different metrics on the WBCD data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.1699</td>
<td>0.7946</td>
</tr>
<tr>
<td>C2</td>
<td>0.6380</td>
<td>0.8789</td>
</tr>
</tbody>
</table>

Table 7. Average intracluster/intercluster distance obtained by the different metrics on the mushroom data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.3882</td>
<td>0.3578</td>
</tr>
<tr>
<td>C2</td>
<td>0.5774</td>
<td>0.4450</td>
</tr>
<tr>
<td>C3</td>
<td>0.3752</td>
<td>0.2643</td>
</tr>
</tbody>
</table>

Table 8. Average intracluster/intercluster distance obtained by the different metrics on the soybean data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.1095</td>
<td>0.0597</td>
</tr>
<tr>
<td>C2</td>
<td>0.6149</td>
<td>0.2081</td>
</tr>
<tr>
<td>C3</td>
<td>0.5233</td>
<td>0.2339</td>
</tr>
<tr>
<td>C4</td>
<td>0.5651</td>
<td>0.1891</td>
</tr>
</tbody>
</table>

Table 9. Average intracluster/intercluster distance obtained by the different metrics on the car data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.5165</td>
<td>0.6316</td>
</tr>
<tr>
<td>C2</td>
<td>0.5688</td>
<td>0.7446</td>
</tr>
<tr>
<td>C3</td>
<td>0.5753</td>
<td>0.7474</td>
</tr>
<tr>
<td>C4</td>
<td>0.5881</td>
<td>0.7409</td>
</tr>
</tbody>
</table>

Table 10. Average intracluster/intercluster distance obtained by the different metrics on the zoo data set

<table>
<thead>
<tr>
<th>Clusters</th>
<th>DM3</th>
<th>W-P metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.17</td>
<td>0.10</td>
</tr>
<tr>
<td>C2</td>
<td>0.73</td>
<td>0.37</td>
</tr>
<tr>
<td>C3</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>C4</td>
<td>0.67</td>
<td>0.62</td>
</tr>
<tr>
<td>C5</td>
<td>0.57</td>
<td>0.51</td>
</tr>
<tr>
<td>C6</td>
<td>0.72</td>
<td>0.29</td>
</tr>
<tr>
<td>C7</td>
<td>0.77</td>
<td>0.35</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this work, we introduced a new similarity/dissimilarity measure for categorical data based on the feature space structure. This distance metric is a variation of pairing matching but weighted. We call our method weighted pairing (W-P) based on feature space-structure. The weights are determined for the number of states that each feature has, indicating which attribute contributes more to the cluster's compact structure. The performance of W-P metric was evaluated in terms of intracluster/intercluster distance, CDI, RI, and NMI index into a K-modes algorithm with support-based initialization in the unsupervised learning framework, and we compare with the distance metric (DM3) and H-SBI methods. The obtained results showed a better performance for W-P than DM3 and H-SBI, we demonstrated that this way of computing a distance is effective in recovering the inherent clustering structures from categorical data when such structures exist, and this can be attributed to the fact that our approach is space-structure based.
ACKNOWLEDGEMENTS

This work was developed under the framework of the research project “Desarrollo de una metodología para la identificación de perfiles de los consumidores del servicio público utilizando técnicas de aprendizaje de máquina” (number 2-20-8) funded by Vice-Rectory for Research of Universidad Tecnológica de Pereira.

REFERENCES