ConDist: A Context-Driven Categorical Distance Measure

Markus Ring[1(✉)], Florian Otto[1], Martin Becker[2], Thomas Niebler[2], Dieter Landes[1], and Andreas Hotho[2]

[1] Faculty of Electrical Engineering and Informatics, Coburg University of Applied Sciences and Arts, 96450 Coburg, Germany
{markus.ring,florian.otto,dieter.landes}@hs-coburg.de
[2] Data Mining and Information Retrieval Group, University of Würzburg, 97074 Würzburg, Germany
{becker,niebler,hotho}@informatik.uni-wuerzburg.de

Abstract. A distance measure between objects is a key requirement for many data mining tasks like clustering, classification or outlier detection. However, for objects characterized by categorical attributes, defining meaningful distance measures is a challenging task since the values within such attributes have no inherent order, especially without additional domain knowledge. In this paper, we propose an unsupervised distance measure for objects with categorical attributes based on the idea that categorical attribute values are similar if they appear with similar value distributions on correlated context attributes. Thus, the distance measure is automatically derived from the given data set. We compare our new distance measure to existing categorical distance measures and evaluate on different data sets from the UCI machine-learning repository. The experiments show that our distance measure is recommendable, since it achieves similar or better results in a more robust way than previous approaches.

Keywords: Categorical data · Distance measure · Heterogeneous data · Unsupervised learning

1 Introduction

Distance calculation between objects is a key requirement for many data mining tasks like clustering, classification or outlier detection [13]. Objects are described by a set of attributes. For continuous attributes, the distance calculation is well understood and mostly the Minkowski distance is used [2]. For categorical attributes, defining meaningful distance measures is more challenging since the values within such attributes have no inherent order [4]. The absence of additional domain knowledge further complicates this task.

However, several methods exist to address this issue. Some are based on simple approaches like checking for equality and inequality of categorical values, or create a new binary attribute for each categorical value [2]. An obvious drawback of these two approaches is that they cannot reflect the degree of similarity

A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 251–266, 2015.
DOI: 10.1007/978-3-319-23528-8_16

or dissimilarity between two distinct categorical values. Yet, more sophisticated methods incorporate statistical information about the data [6–8].

In this paper, we take the latter approach. In contrast to previous work, we take into account the quality of information that can be extracted from the data, in form of correlation between attributes. The resulting distance measure is called ConDist (Context based Categorical Distance Measure): We first derive a distance measure for each attribute separately. To this end, we take advantage of the fact that categorical attributes are often correlated, as shown in an empirical study [8], or by the fact that entire research fields exist which detect and eliminate such correlations, e.g. feature selection or dimensionality reduction. In order to calculate the distances for the values within a *target attribute*, we first identify the correlated context attributes. The distance measure on target attributes is then based on the idea that attribute values are similar if they appear with similar value distributions on their corresponding set of correlated context attributes. Finally, we combine these distance measures on separate attributes to calculate the distance of objects, again taking into account correlation information. We argue that incorporating the correlation of context attributes and the target attribute itself are important in order to maximize the relevant distance information extracted from the data and mitigate the possibly incorrect influence of uncorrelated attributes.

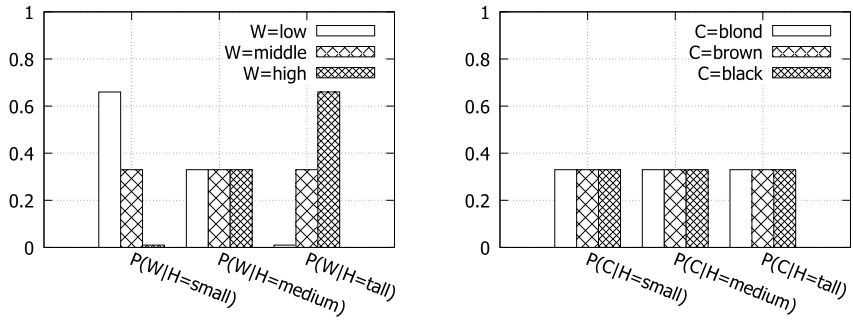
Table 1 shows a sample data set. Let us assume, we want to calculate the distance between the different values of attribute *height*, i.e., *height* is our target attribute. As mentioned above, our distance measure calculates its distance based on the value distributions of other attributes. For the attributes *weight* and *haircolor* these distributions ($P(X|H = small)$, $P(X|H = medium)$ and $P(X|H = tall)$) are shown in Figure 1. In the case of *weight* the distributions are different. Thus, they will add information to our distance calculations. However, the distributions for *haircolor* are the same for all values of the target attribute. Thus, they will not contribute information to our distance measure. At the same time, we can see that *weight* is correlated to *height*, since higher weight implies greater height with a high probability. For *haircolor* on the other hand, there is no correlation, since *haircolor* does not imply *height*. Since we also take this correlation information into account, we exclude uncorrelated attributes from the distance measure. Therefore, context attribute *haircolor* will not be taken into account when calculating distances between the values of *height*.

Overall, we propose an unsupervised distance measure for objects described by categorical attributes. Our new distance measure *ConDist* calculates distances by identifying and utilizing relevant statistical relationships from the given data set in form of correlations between attributes. This way, *ConDist* tries to compensate for the lack of inherent orders within categorical attribute domains.

The rest of the paper is organized as follows: Related work on categorical distance measures is discussed in Section 2. Section 3 describes the proposed distance measure *ConDist* in detail. Section 4 gives an experimental evaluation and the results are discussed in Section 5. The last section summarizes the paper.

Table 1. Example data set which describes nine people with three categorical attributes. The attributes *height* and *weight* have natural orders. Whereas the attribute *haircolor* has no natural order. *Height* and *weight* are correlated to each other while the attribute *haircolor* is uncorrelated to the other two attributes. *ConDist* uses such correlations between attributes to extract relevant information for distance calculation.

#	height	weight	haircolor
1	small	low	blond
2	small	low	brown
3	small	middle	black
4	medium	low	black
5	medium	middle	brown
6	medium	high	blond
7	tall	middle	blond
8	tall	high	brown
9	tall	high	black



(a) CPD of *weight* (b) CPD of *haircolor*

Fig. 1. This figure shows the conditional probability distributions (CPDs) of context attributes *weight* and *haircolor*, given the different values of the target attribute *height* based on Table 1. W stands for *weight*, C for *haircolor* and H for *height*. *ConDist* uses the differences between CPDs of context attributes to calculate the distance of target attribute values. Thus, *weight* can be used to calculate a meaningful distance between the values of *height*, while *haircolor* will yield the same distance for all three target attribute values.

2 Related Work

This section reviews related work on categorical distance measures. Distance measures can be divided into supervised and unsupervised. In the supervised setting, the class membership of the objects is provided and this information is exploited by the distance measures. In the unsupervised setting, distance measures are based exclusively on assumptions and statistics of the data. Since the proposed distance measure is unsupervised, the following review considers only

unsupervised categorical distance measures. We categorize them into distance calculation (I) without considering context attributes, (II) considering all context attributes, (III) considering a subset of context attributes and (IV) based on entire objects instead of individual attributes.

Boriah et al. [4] give a comprehensive overview of distances measures from category (I). In contrast to *ConDist*, these distance measures ignore available information that could be extracted from context attributes. For example, the distance measure *Eskin* only uses the cardinality of the target attribute domain to calculate distances. [4] evaluated these distance measures for outlier detection and observed that no specific distance measure dominates all others.

Category (II) includes distance measures that employ all context attributes without distinguishing between correlated and uncorrelated. Li and Ho [8] compute the distance between two categorical values as the sum of dissimilarities between the context attributes' conditional probability distribution (CPD) when the target attribute takes these two values. However, they do not recommend their distance measure for data sets with highly independent attributes. Similary, [1] calculates the distance between two values using the co-occurrence probabilities of these two values and the values of the context attributes.

Category (III) selects a subset of context attributes for each target attribute. DILCA [6] is a representative of this category and uses Symmetric Uncertainty (SU) [15] for selecting context attributes. SU calculates the correlation between two attributes. In contrast to our work, all selected context attributes are weighted equally for the distance calculation. Consequently, the potentially differing suitability of the selected context attributes is not reflected in the distance calculation process.

Category (IV) aims to compute distances between entire objects instead of distances between different values within an attribute. Consequently, the distance between different values within an attribute varies in dependence of the whole objects. Recently, Jia and Cheung [7] proposed such a distance measure for cluster analysis. Their basic assumption is that two categorical values with high frequencies should have a higher distance than two categorical values with low frequencies. Therefore, they select and weigh a set of correlated context attributes for each target attribute using the normalized mutual information [3]. Jia and Cheung [7] compared their distance measure with the *Hamming Distance* on four data sets. They conclude that their distance measure performs better than the *Hamming Distance* on the evaluated data sets.

The proposed distance measure *ConDist* neither ignores context attributes (category I) nor simply includes all context attributes (category II). Instead, it follows the approach of the third category but extends the subset selection with a weighting scheme for context attributes. Furthermore, the target attribute itself is included in the distance computation. In contrast to the fourth category, two particular values within an attribute have always the same distance, independent of the corresponding objects. This allows *ConDist* to calculate a distance matrix for each attribute.

3 The Distance Measure ConDist

This section introduces *ConDist*, a new distance measure. Section 3.1 presents the underlying ideas and the core formula. Since *ConDist* first calculates the distance between single attributes before combining them, it requires adjusted distance functions for each attribute. In Section 3.2, we explain how these attribute distance functions are derived. When combining attribute-wise distances, *ConDist* uses a specific weighting scheme which is explained in Section 3.3. Both, the attribute distance functions as well as the weighting scheme use a set of correlated context attributes. Section 3.4 defines how this set is derived and how an impact factor is calculated which accounts for the varying amount of information dependent on different correlation values. Finally, we address the issue of how *ConDist* can be applied to objects characterized by continuous and categorical attributes in Section 3.5.

3.1 Definition of ConDist

This section provides the core formula of *ConDist*, calculating the distance between two objects characterized by attributes.

Let A and B be two objects in the data set D and let each object be characterized by n attributes. Furthermore, let the value of attribute X for object A be denoted by A_X. *ConDist* follows a two-step process: First, it calculates the distance between each of the attributes of the objects A and B and then it combines them using attribute specific weights. Formally, *ConDist* defines the distance for two objects A and B as the weighted sum over all attribute distances:

$$ConDist(A, B) = \sum_X w_X \cdot d_X(A, B), \tag{1}$$

where w_X denotes the weighting factor assigned to attribute X (defined in Section 3.3) and $d_X(A, B)$ denotes the distance of the values A_X and B_X of attribute X in the objects A and B (defined in Section 3.2).

The distance function d_X on the values of each attribute X needs to be calculated beforehand and is based on the idea that attribute values with similar distributions of values in a set of correlated context attributes are similar. The weighting factor w_X accounts for differences in the number of context attributes and the degree of their correlation with the target attribute X. Both, the attribute distance functions d_X as well as the weighting factors w_X incorporate correlation information in order to maximize the relevant information that can be extracted from the data set and mitigate the possibly incorrect influence of uncorrelated attributes. For an example on differently correlated attributes and their influence on distribution based distance measures, please refer to Section 1 as well as Table 1 and Figure 1.

3.2 Attribute Distance d_X

As mentioned in Section 3.1, the distance d_X of values of a single attribute X is based on the idea that attribute values $x \in dom(X)$ are similar if they appear

with similar distributions of values in a set of correlated context attributes. Thus, when comparing two objects A and B in attribute X, we first calculate the Euclidean distance between the two conditional probability distributions $P(Y|X = A_X)$ and $P(Y|X = B_X)$ for each attribute $Y \in context_X$ from the set of correlated context attributes $context_X$ of the target attribute X. Then, we weight them using an individual impact factor $impact_X(Y)$ (Section 3.4) and add up these distances for all attributes $Y \in context_X$. The impact factor accounts for the fact that the amount of information about the target attribute X in a context attribute Y decreases with both increasing and decreasing correlation $cor(X|Y)$ as explained in Section 3.4. The resulting formula is:

$$\hat{d}_X(A,B) = \sum_{Y \in context_X} impact_X(Y) \sqrt{\sum_{y \in dom(Y)} \Big(p(y|A_X) - p(y|B_X)\Big)^2}, \quad (2)$$

where $dom(Y)$ is the domain of attribute Y, $p(y|A_X) = p(y|X = A_X)$ denotes the probability that value y of context attribute Y is observed under the condition that value A_X of attribute X is observed in the data set D.

As mentioned above, the attribute distance d_X relies on a set of correlated context attributes $context_X$ as defined in Section 3.4. Because every attribute is correlated to itself, the target attribute is also added to the set of context attributes. The motivation for including the target attribute is two-fold: First, it ensures that the list of context attributes is not empty even if all attributes are independent. Second, the distance between two distinct values is always larger than 0. Thus, if no correlated context attributes can be identified, *ConDist* calculates the maximum distance for each distinct value-pair in target attribute X. In this case, *ConDist* reduces to the distance measure *Overlap* and distinguishes only between equality and inequality of categorical values.

It should be noted that *ConDist* normalizes the attribute distance by the maximum distance value $d_{X,max}$ between any two values $x, u \in dom(X)$ of attribute X:

$$d_X(A,B) = \frac{\hat{d}_X(A,B)}{d_{X,max}} \quad (3)$$

The proof that *ConDist* is a distance measure closely follows the proof of the Euclidean metric and exploits the fact that a linear combination of distance measures is also a distance measure. For brevity reasons, we omit the proof.

3.3 Attribute Weighting Function w_X

ConDist compares objects based on the distances between each of the attribute values associated with the objects it compares (see Equation (1)). Each of these attributes is weighted differently by an individual weighting factor w_X. This section explains why these weights w_X are necessary and how they are calculated.

The weight w_X is especially necessary for data sets in which some attributes depend on each other, while others do not: refer back to the example in Table 1.

For attribute *haircolor*, no correlated context attribute can be identified. Consequently, only the attribute *haircolor* itself is used for distance calculation and no additional information can be extracted from context attributes. Therefore, the normalized results of Equation (2) always results in the maximum distance 1 for any pair of non-identical values. In contrast, the attribute *weight* is a correlated context attribute for attribute *height*, and vice versa. Consequently, *ConDist* is able to calculate more meaningful distances for both attributes and these attributes should be weighted higher than *haircolor*.

However, average distances in attribute *haircolor* are larger than in attributes *weight* and *height*. Consequently, distinct values in attribute *haircolor* have implicitly larger relative weight than distinct values in attributes *height* and *weight*.

To solve this issue, the weighting factor w_X assigns a weight to each attribute X based on (I) the amount of identified context attributes and (II) their impact on the target attribute X:

$$w_X = 1 + \frac{\sum_{Y \in context_X} impact_X(Y)}{n \cdot c}, \tag{4}$$

where $context_X$ and $impact_X(Y)$ are defined as in section 3.4, n is the number of attributes in the data set D and c denotes a normalization factor defined as the maximum of the impact function (see Section 3.4) which is independent of the attributes X and Y and amounts to $\frac{8}{27}$.

3.4 Correlation, Context and Impact

The attribute distance measures d_X (Section 3.2) and the weighting scheme w_X (Section 3.3) use the notion of correlation on categorical distance measures as well as a correlation related impact factor. Both are defined here.

Correlation $cor(X|Y)$. A measure of correlation is required to determine an appropriate set of context attributes. For this purpose, we build a correlation measure on the basis of the *Information Gain* (IG) which is widely used in information theory [9]. The IG is calculated as follows:

$$IG(X|Y) = H(X) - H(X|Y), \tag{5}$$

where $H(X)$ is the entropy of an attribute X, and $H(X|Y)$ is the conditional entropy of attribute X given attribute Y. According to this measure, the attribute X is more correlated with attribute Y than attribute W if $IG(X|Y) > IG(X|W)$. The information gain $IG(X|Y)$ is always less than or equal to the entropy $H(X)$. Based on this observation, the function $cor(X|Y)$ is defined as:

$$cor(X|Y) = \frac{IG(X|Y)}{H(X)} \tag{6}$$

and describes a correlation measure which is normalized to the interval $[0, 1]$. The quality of possible conclusions from the given attribute Y to the target attribute X can differ from the quality of conclusions from given attribute X to target attribute Y. This aspect is considered in the asymmetric correlation function $cor(X|Y)$ and allows us to always extract the maximum amount of useful information for each target attribute X.

Context $context_X$. For both, the attribute distance d_X (Section 3.2) in *ConDist* as well as for the weighting scheme w_X (Section 3.3), the notion of a set of correlated context attributes $context_X$ is used. This set is defined using the previously defined correlation function $cor(X|Y)$ and a user-defined threshold θ. That is, context attributes Y are included in $context_X$ only if their correlation with target attribute X is equal to or exceeds the threshold θ:

$$context_X = \{Y \mid cor(X|Y) \geq \theta\} \tag{7}$$

Impact $impact_X(Y)$. Again, for both, the attribute distance d_X (Section 3.2) as well as for the weighting scheme w_X (Section 3.3), a so called impact factor $impact_X(Y)$ is used. This factor accounts for the fact that the amount of information about the target attribute X in a context attribute Y decreases with both increasing and decreasing correlation $cor(X|Y)$.

A high correlation value means that a value of a context attribute $Y \in context_X$ implies the value of a target attribute X with a high probability. For example, when we know that someone is heavy, it is more likely that this person is tall than small (see Table 1). Thus, in the extreme case of perfectly correlated attributes, the conditional probability distributions $P(Y|X = A_X)$ and $P(Y|X = B_X)$, for $A_X \neq B_X$ do not overlap. This means that using the Euclidean distance to calculate the similarity of those two CPDs (as in Formula (2)) limits the distance information gained from the context attribute to values of 0 (for $A_X = B_X$) and 1 (for $A_X \neq B_X$) after normalization in Formula (3).

A low correlation value means that a value of a context attribute $Y \in context_X$ implies little to no preference for a particular value of a target attribute X. This means that the similarity between the conditional probability distributions $P(Y|X = A_X)$ and $P(Y|X = B_X)$ may be random, thus, possibly conveying incorrect distance information.

Consequently, non-correlated attributes are excluded to avoid introducing incorrect information. In contrast, perfectly correlated attributes are still used, because they contribute at least no incorrect information. However, since they deliver exclusively high distances for distinct values, their impacts should be reduced. Otherwise, the distances calculated by the other context attributes would be blurred.

Therefore, we choose a weighting function that (I) increases fast at the onset of correlation between attributes, (II) increases more slowly with existing, but partial correlation, and (III) decreases at nearly perfect correlation. In particular, we propose the impact function as depicted in Figure 2 and defined as:

$$impact_X(Y) = cor(X|Y)\Big(1 - 0.5 \cdot cor(X|Y)\Big)^2. \tag{8}$$

In general, this impact function can be replaced by other functions respecting the three properties introduced above.

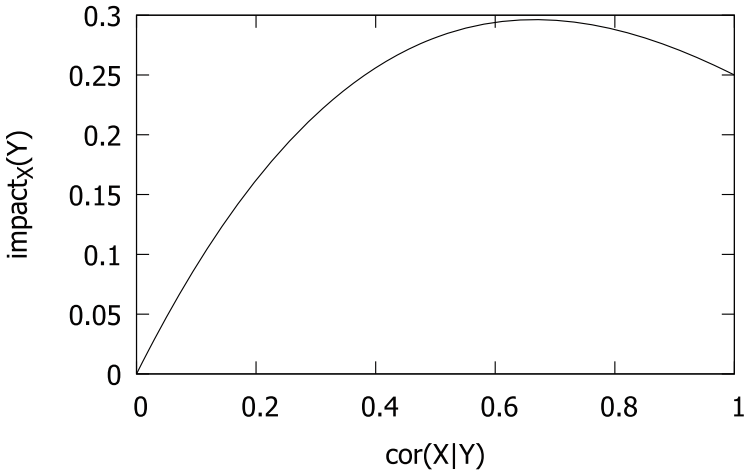


Fig. 2. Graph of the impact function $impact_X(Y)$ as defined in (8).

3.5 Heterogeneous Data Sets

Many real-world data sets contain both continuous and categorical attributes. To apply *ConDist* to such data sets, two situations have to be distinguished: either the target attribute is continuous or the context attribute is continuous.

If the target attribute is continuous, no context attributes are necessary. The Minkowski distance can be used, but should be normalized to the interval $[0, 1]$. Since meaningful distances can be calculated for continuous attributes, the attribute weight w_X (see Section 3.3) should be maximized. If the context attribute is continuous, the continuous value range should be discretized. We propose to use the discretization algorithm TUBE [11], because it does not require any parameters. Other discretization algorithms can be used as well.

4 Experiments

This section presents an experimental evaluation of *ConDist* in the context of classification and clustering. We compared *ConDist* with *DILCA* [6], *JiaCheung* [7], *Quang* [8] and several distance measures presented in [4], namely *Eskin*, *Gambaryan*, *Occurrence Frequency (OF)* and *Overlap*. For *DILCA*, we used the non-parametric approach $DILCA_{RR}$ as described in [6] and for *JiaCheung* we set the threshold parameter β to 0.1 as recommended by the authors [7].

Table 2. Characteristics of the data sets. The column *Correlation* contains the average correlation between each pair of attributes in the data set, calculated by the function $cor(X|Y)$, see Equation (6). The value ranges from 0 if no correlation exists to 1 if all attributes are perfectly correlated. The data sets are separated in three subsets from highly correlated to uncorrelated based on their average correlation.

Data Sets	Instances	Attributes	Classes	Correlation
Teaching Assistant Evaluation	151	5	3	0.336
Soybean Large	307	35	19	0.263
Breast Cancer Winconsin	699	10	2	0.216
Mushroom-Extended	8416	22	2	0.162
Mushroom	8124	22	2	0.161
Dermatology	366	34	6	0.098
Lymphography	148	18	4	0.070
Soybean Small	47	35	19	0.070
Breast Cancer	286	9	2	0.054
Audiology-Standard	226	69	24	0.044
Hayes-Roth	160	4	3	0.045
Post-Operative Patient	90	8	3	0.031
TicTacToe	958	9	2	0.012
Monks	432	6	2	0.000
Balance-Scale	625	4	3	0.000
Car	1728	6	4	0.000
Nursey	12960	8	5	0.000

4.1 Evaluation Methodology

Classification. A k-Nearest-Neighbour classifier is used to compare *ConDist* with existing categorical distance measures in the context of classification. For simplification, we fix $k = 7$ in all tests. We evaluate by 10-fold-cross validation and use the classification accuracy as evaluation measure. To reduce confounding effects of the generated subsets, the 10-fold cross-validation is repeated 100 times with different subsets for each data set. We finally compare the averages of the classification accuracies over all executions.

Clustering. The hierarchical WARD algorithm [14] is used to evaluate the performance of *ConDist* in the context of clustering. *ConDist* and its competitors are used to calculate the initial distance matrix as input for WARD. For simplification, the clustering process is terminated when the number of clusters is equal to the number of classes in the data sets. Performance is measured by *Normalized Mutual Information (NMI)* [12] which ranges from 0 for poor clustering to 1 for perfect clustering with respect to the predefined classes.

Data Sets. For the evaluation of *ConDist* the *multivariate categorical data sets for classification* from the UCI machine learning repository [10] are chosen. We exclude data sets with less than 25 objects (e.g., *Balloons*) or mainly binary

Table 3. Classification accuracies for various thresholds θ in *ConDist*. Each column contains the results in percent for particular thresholds θ.

	Threshold θ								
Data Set	0.00	0.01	0.02	0.03	0.05	0.10	0.20	0.50	1.00
Soybean Large	91.74	91.74	91.79	91.80	**91.82**	89.75	89.36	89.63	91.30
Lymphography	**83.36**	**83.36**	83.30	83.24	83.01	81.99	82.01	81.24	81.26
Hayes-Roth	68.11	68.36	68.51	68.60	**69.21**	64.47	64.47	64.47	64.47
TicTacToe	**99.99**	**99.99**	**99.99**	99.98	94.74	94.74	94.74	94.74	94.74
Balance-Scale	77.35	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**
Car	88.98	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**
Average	84.92	85.45	**85.47**	**85.47**	84.67	83.36	83.30	83.22	83.50

attributes (e.g., *Chess*). Furthermore, we include some *multivariate mixed data sets for classification* from the UCI machine learning repository which mainly consist of categorical attributes and some integer attributes with a small set of distinct values (e.g. an integer attribute that contains the number of students in a course): *Teaching Assistant Evaluation, Breast Cancer Winconsin, Dermatology* and *Post-Operative Patient*. Since not all competitors have an explicit way to process integer attributes, we treated all integer attributes as categorical. The final set of data sets is given in Table 2. The data sets are divided in three subgroups: highly-correlated (Correlation ≥ 0.05), weakly-correlated (Correlation > 0) and uncorrelated (Correlation $= 0$).

4.2 Experiment 1 – Context Attribute Selection

Experiment 1 analyzes the effects of varying threshold θ (see Section 3.4) in *ConDist*'s context attribute selection. The threshold θ defines the minimum value of the function $cor(X|Y)$ that a candidate attribute Y has to reach in order to be selected as context attribute for the target attribute X. The higher the threshold θ, the fewer context attributes are used. In the extreme case of $\theta = 0$, all context attributes are used for distance calculation. In the other extreme case $\theta = 1$, only the target attribute itself is used. For this experiment, a representative subset of two highly-correlated (*Soybean Large* and *Lymphography*), two weakly-correlated (*Hayes-Roth* and *TicTacToe*) and two uncorrelated (*Balance-Scale* and *Car*) data sets are used. The results can be seen in Table 3.

The average classification accuracy (I) increases with low thresholds θ, (II) reaches a peak at $\theta = 0.02$ and $\theta = 0.03$, (III) decreases slowly with medium high thresholds, (IV) reaches the minimum at $\theta = 0.5$ and (V) slowly increases with high thresholds again. For nearly all data sets, the classification accuracy stabilizes with increasing thresholds. The lower the attribute correlation within the data set, the faster this effect is reached. For uncorrelated data sets like *Car* and *Balance-Scale*, it can already be observed with thresholds greater than or equal to $\theta = 0.01$. Due to the peak at $\theta = 0.02$, this value is used for the further experiments in this paper.

Table 4. Comparison of categorical distance measures in the context of classification. Each column contains the classification accuracies in percent for a particular distance measure. The data sets are separated in three subsets from highly correlated to uncorrelated based on their average correlation.

Data Set	ConDist	DILCA	Eskin	JiaCheung	Gambaryan	OF	Overlap	Quang
Teaching Assistant. E.	49.85	**50.68**	48.79	49.54	49.44	39.16	45.84	44.48
Soybean Large	91.79	91.48	89.83	89.45	87.18	89.61	91.30	**92.01**
Breast Cancer W.	96.13	95.55	95.67	95.08	92.84	72.47	95.25	**96.28**
Dermatology	96.76	**97.97**	94.91	97.39	91.69	61.12	95.90	96.64
Lymphography	83.30	82.09	79.17	**83.95**	80.72	72.77	81.26	81.53
Breast Cancer	73.85	72.94	73.18	74.30	**74.55**	68.32	74.06	70.45
Audiology Standard	**66.44**	64.80	63.24	60.95	66.16	51.87	61.27	55.56
Hayes-Roth	68.50	67.59	46.71	68.27	60.84	58.71	61.74	**71.19**
Post-Operative Patient	69.62	68.22	68.36	67.28	**69.69**	69.44	68.59	68.69
TicTacToe	**99.99**	90.65	94.74	99.93	98.25	76.80	94.74	99.65
Car	**90.56**	90.25	90.03	90.01	90.25	87.83	**90.56**	88.25
Nursey	**94.94**	92.61	93.29	93.32	93.24	94.65	**94.94**	94.72
Monks	94.50	90.76	87.29	87.34	86.61	**98.67**	94.50	96.66
Balance-Scale	**78.66**	78.43	**78.66**	78.65	77.13	78.54	**78.66**	77.51
Average	**82.49**	81.00	78.85	81.10	79.90	72.85	80.62	80.97

4.3 Experiment 2 – Comparison in the Context of Classification

Experiment 2 compares *ConDist* with several categorical distance measures in the context of classification. All data sets from Table 2 are used. The results are given in Table 4, except for the data sets *Mushroom-Extended*, *Mushroom* and *Soybean Small*. These data sets are omitted in the table since all distance measures reach 100 percent classification accuracy. Consequently, these data sets would only blur the differences between the categorical distance measures.

ConDist achieves the highest average classification accuracy of all distance measures. In the case of highly- and weakly-correlated data sets, context based categorical distance measures (*ConDist, DILCA, JiaCheung* and *Quang*) achieve mostly better results than other distance measures. In the case of uncorrelated data, previous context based categorical distance measures are inferior to *ConDist* and non-context based categorical distance measures.

Statistical Significance Test. In this test, we want to evaluate if the differences in Table 4 are statistically significant. Demšar [5] deals with the statistical comparison of classifiers over multiple data sets. They recommend the Wilcoxon Signed-Ranks Test for the comparison of two classifiers and the Friedman-Test for the comparison of multiple classifiers. Therefore, we use the Friedman-Test to compare all distance measures and the Wilcoxon Signed-Ranks Test for post-hoc tests. The Friedman-Test is significant for $p < 0.001$; thus we can reject the null hypothesis that all distance measures are equivalent. Consequently, we applied

Table 5. Results of the Wilcoxon Signed-Ranks Test comparing the classification accuracies of *ConDist* with each other distance measure. The first row contains the calculated p-value, the second row contains the result of the Wilcoxon Signed-Ranks Test: *yes*, if *ConDist* performs statistically different, *no* otherwise.

	DILCA	Eskin	JiaCheung	Gambaryan	OF	Overlap	Quang
p-value	0.016	0.002	0.045	0.002	0.002	0.008	0.096
significant	yes	yes	yes	yes	yes	yes	no

the Wilcoxon Signed-Ranks Test with $\alpha = 0.05$ on the classification accuracies of Table 4. Table 5 shows that there is a significant difference between *ConDist* and the distance measures *Eskin, JiaCheung, Gambaryan, OF* and *Overlap*. However, the test fails for *ConDist* and *Quang*.

4.4 Experiment 3 – Comparison in the Context of Clustering

Experiment 3 compares *ConDist* with several categorical distance measures in the context of clustering. All data sets from Table 4 are used. The results are given in Table 6.

For some data sets (*Teaching Assistang Evaluation, Lymphography, Breast Cancer, Hayes-Roth, Post-Operative Patient, TicTacToe, Monks, Balance-Scale, Nursey* and *Car*) the clustering fails to reconstruct the predefined classes. For the remaining data sets, no distance measure dominates. However, most distance measures perform poorly on single data sets, whereas *ConDist* achieves more stable results.

Statistical Significance Test. In analogy to Section 4.3, we first apply the Friedman-Test on the results shown in Table 6. Here, the null hypothesis that all distance measures are equivalent cannot be rejected. Nevertheless, we applied the Wilcoxon Signed-Ranks Test ($\alpha = 0.05$) between *ConDist* and the other distance measures. Except for *Eskin* and *Quang*, the results of the Wilcoxon Signed-Ranks Test show no statistically significant differences.

5 Discussion

5.1 Experiment 1 – Context Attribute Selection

Table 3 shows that many useful context attributes are discarded if threshold θ is too high. This is especially the case for weakly correlated data sets, e.g. *Hayes-Roth* and *TicTacToe*. For *Hayes-Roth*, the decrease of classification accuracy is observed for $\theta > 0.05$, and for *TicTacToe* the decrease of classification accuracy is already observed for $\theta > 0.02$. In contrast to this, if the threshold θ is too low, independent context attributes are added which may contribute noise to the distance calculation. This is especially the case for uncorrelated data sets,

Table 6. Comparison of categorical distance measures in the context of clustering. Each column contains the *NMI* of the clustering results found by the *WARD* algorithm where the initial distance matrix is calculated with the particular distance measure. *NMI* assigns low values to poor clusterings and high values to good clusterings with respect to the predefined classes. The data sets are also separated in three subsets based on their average correlation.

Data Set	ConDist	DILCA	Eskin	JiaCheung	Gambaryan	OF	Overlap	Quang
Teaching Assistant Eva.	.078	**.085**	**.085**	**.085**	**.085**	.060	.044	.042
Soybean Large	.803	.785	.758	.735	.772	**.805**	.793	.778
Breast Cancer Winconsin	.785	.557	.749	.656	.601	.217	.621	**.798**
Mushroom Extended	**.597**	**.597**	.317	.223	**.597**	**.597**	**.597**	.245
Mushroom	**.594**	**.594**	.312	**.594**	**.594**	.312	**.594**	.241
Dermatology	.855	**.946**	.832	.879	.863	.292	.847	.859
Lymphography	.209	.303	.165	.207	.163	.243	.226	**.320**
Soybean Small	.687	.690	.687	**.701**	.692	.690	.689	.692
Breast Cancer	.063	**.068**	.031	.074	.001	.002	.100	.001
Audiology-Standard	.661	.612	.623	**.679**	.620	.439	.568	.582
Hayes-Roth	.017	.027	.004	.012	.007	**.166**	.006	.029
Post-Operative Patient	**.043**	.017	.018	.025	.017	.032	.019	.033
TicTacToe	**.087**	.003	.003	.082	.085	.001	.033	.039
Monks	.001	.000	.000	.000	.000	**.081**	.001	.003
Balance-Scale	**.083**	.036	.064	.067	.064	.064	**.083**	.036
Car	.062	.036	**.150**	**.150**	**.150**	.062	.062	.036
Nursey	.048	.006	.037	.037	.037	**.098**	.048	.006
Average	**.334**	.315	.284	.306	.315	.245	.314	.279

e.g. for $\theta = 0$ in *Balance-Scale* and *Car*. However, *ConDist's* impact function $impact_X(Y)$ accounts for this effect in highly-correlated data sets.

Consequently, the concrete value of the threshold θ is not too crucial as long as two conditions are fulfilled: (I) θ must be large enough to ensure that context attributes are purged whose correlations are caused by too small data sets and (II) θ must be small enough to ensure that context attributes with significant correlations are retained. Therefore, we recommend $\theta = 0.02$ for *ConDist*, because the experiments show that this threshold achieves the best overall results.

5.2 Experiment 2 – Comparison in the Context of Classification

For highly correlated data sets, distance measures using context attributes outperform other distance measures. However, for those data sets no best distance measures can be identified among the context based distance measures.

For uncorrelated data sets, previous context-based distance measures (*DILCA*, *Quang* and *JiaCheung*) achieved inferior results in comparison to *ConDist* and non-context based distance measures. This is because, e.g., *DILCA*

and *Quang* use only context attributes for the distance calculation which results in random distances if all context attributes are uncorrelated.

In contrast, *ConDist* achieved acceptable results because not only correlated context attributes, but also the target attributes are considered. This effect is also illustrated by the comparison between *ConDist* and *Overlap*. *ConDist* is equal to *Overlap* if no correlated context attributes can be identified, see uncorrelated data sets (*Monks, Balance-Scale, Nursey* and *Car*) in Table 4. However, for weakly- and highly-correlated data sets, *ConDist's* consideration of context attributes turns into an advantage, leading to better results than *Overlap*. The improvement of *ConDist* can be statistically confirmed by the Wilcoxon Signed-Ranks Test (see Table 5).

5.3 Experiment 3 – Comparison in the Context of Clustering

Table 6 shows that the majority of the different distance measures reach, by and large, similar outcomes for individual data sets. This is because the clustering algorithm and its ability to reconstruct the given classes have much higher impact on the results than the distance measure used to calculate the initial distance matrix. However, it can be seen that the performance of single distance measures strongly decreases for individual data sets. For example, *JiaCheung* which often achieves good results, performs very poorly in the *Mushroom* data set. Similar observations can be made for *OF*, *Eskin*, *Quang* and *DILCA*, mainly in the data sets *Breast Cancer Winconsin*, *Mushroom*, *Mushroom Extended*, *Dermatology* and *Audiology-Standard*. In contrast, *ConDist* is almost always among the best results and shows the most stable results for the different data sets.

The Friedman-Test fails for Experiment 3 and the Wilcoxon Signed-Ranks Test shows also no statistically significant differences in the performance of *ConDist* and the compared distance measures, except for *Eskin* and *Quang*. However, the results of Experiment 3 lead to the assumption that *ConDist* may be a more robust distance measure than its competitors.

6 Summary

Categorical distance calculation is a key requirement for many data mining tasks. In this paper, we proposed *ConDist*, an unsupervised categorical distance measure based on the correlation between the target attribute and context attributes. With this approach, we aim to compensate for the lack of inherent orders within categorical attributes by extracting statistical relationships from the data set.

Our experiments show that *ConDist* is a generally usable categorical distance measure. In the case of correlated data sets, *ConDist* is comparable to existing context based categorical distance measures and superior to non-context based categorical distance measures. In the case of weakly and uncorrelated data sets, *ConDist* is comparable to non-context based categorical distance measures and superior to context based categorical distance measures. The overall improvement of *ConDist* can be statistically confirmed in the context of classification. In the context of clustering, this improvement could not be statistically confirmed.

In the future, we want to extend the proposed distance measure so that it can automatically infer the parameter θ from the data sets. Additionally, we want to transform categorical attributes to continuous attributes with aid of the proposed distance measure.

Acknowledgments. This work is funded by the Bavarian Ministry for Economic affairs through the WISENT project (grant no. IUK 452/002) and by the DFG through the PoSTS II project (grant no. STR 1191/3-2). We appreciate the support of our project partners HUK COBURG, Coburg, Germany and Applied Security, Großwallstadt, Germany.

References

1. Ahmad, A., Dey, L.: 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. Alamuri, M., Surampudi, B.R., Negi, A.: A survey of distance/similarity measures for categorical data. In: Proc. of IJCNN, pp. 1907–1914. IEEE (2014)
3. Au, W.H., Chan, K.C., Wong, A.K., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **2**(2), 83–101 (2005)
4. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. In: Proc. SIAM Int. Conference on Data Mining, pp. 243–254 (2008)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)
6. Ienco, D., Pensa, R.G., Meo, R.: Context-Based Distance Learning for Categorical Data Clustering. In: Adams, N.M., Robardet, C., Siebes, A., Boulicaut, J.-F. (eds.) IDA 2009. LNCS, vol. 5772, pp. 83–94. Springer, Heidelberg (2009)
7. Jia, H., Cheung, Y.M.: A new distance metric for unsupervised learning of categorical data. In: Proc. of IJCNN, pp. 1893–1899. IEEE (2014)
8. Le, S.Q., Ho, T.B.: An association-based dissimilarity measure for categorical data. Pattern Recognition Letters **26**(16), 2549–2557 (2005)
9. Lehmann, E., Romano, J.: Testing Statistical Hypotheses, Springer Texts in Statistics. Springer (2005)
10. Lichman, M.: Uci machine learning repository (2013). http://archive.ics.uci.edu/ml
11. Schmidberger, G., Frank, E.: Unsupervised Discretization Using Tree-Based Density Estimation. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 240–251. Springer, Heidelberg (2005)
12. Strehl, A., Ghosh, J.: Cluster ensembles–a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3 (2003)
13. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley Boston (2006)
14. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)
15. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. ICML **3**, 856–863 (2003)