Automatic Threshold Calculation for the Categorical Distance Measure ConDist

Markus Ring[1], 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,dieter.landes}@hs-coburg.de,
[2] Data Mining and Information Retrieval Group, University of Würzburg, 97074 Würzburg, Germany
{hotho}@informatik.uni-wuerzburg.de

Abstract. The measurement of distances between objects described by categorical attributes is a key challenge in data mining. The unsupervised distance measure *ConDist* approaches this challenge based on the idea that categorical values within an attribute are similar if they occur with similar value distributions on correlated context attributes. An impact function controls the influence of the correlated context attributes in *ConDist's* distance calculation process.
ConDist requires a user-defined threshold to purge context attributes whose correlations are caused by noisy, non-representative or small data sets. In this work, we propose an automatic threshold calculation method for each pair of attributes based on their value distributions and the number of objects in the data set. Further, these thresholds are also considered when applying *ConDist's* impact function. Experiments show that this approach is competitive with respect to well selected user-defined thresholds and superior to poorly selected user-defined thresholds.

Keywords: categorical data, distance measure, unsupervised learning

1 Introduction

Distance calculation between objects is a key requirement for many data mining tasks like clustering, classification or outlier detection [15]. Objects are described by a set of attributes which can be divided into continuous and categorical attributes. For continuous attributes, distance calculation is well understood and mostly uses the Minkowski distance [2]. For categorical attributes, defining meaningful distance measures is more challenging since the values within such attributes have no inherent order [4]. However, several methods exist to address this issue. A comprehensive overview of categorical distance measures is given

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

in [4]. Yet, more sophisticated categorical distance measures incorporate statistical information like correlations about the data [8,9,11,14]. *ConDist* (Context based Categorical Distance Measure) [14] is such an unsupervised categorical distance measure. For distance calculation, *ConDist* extracts available information from correlations between the target attribute (the attribute for which distances shall be calculated) and the correlated context attributes. *ConDist* uses a correlation measure based on the information gain. Each context attribute whose correlation exceeds a user-defined threshold θ is used for distance calculation. This threshold θ must be large enough to ensure that context attributes are purged whose correlations are caused by noisy, non-representative or too small data sets. Simultaneously, the threshold θ must be small enough to retain context attributes with significant correlations.

In this paper, we propose a data-driven method for calculating *ConDist's* threshold. In [14], the user has to define a single threshold for all attributes. In contrast to this approach, the proposed method calculates an individual threshold $\theta_{X|Y}$ for each combination of target attribute X and context attribute Y. These thresholds $\theta_{X|Y}$ can be better adapted to the specific correlation requirements of two concrete attributes than a single threshold θ. We consider the number of objects in the data set and the value distributions of target attribute X and context attribute Y when calculating the individual thresholds $\theta_{X|Y}$. The calculated thresholds $\theta_{X|Y}$ are also taken into account when applying *ConDist's* impact function. The impact function controls the influence of the correlated context attributes in *ConDist's* distance calculation process and considers the varying amount of information that can be extracted from a correlated context attribute. The proposed method for the automatic threshold calculation makes *ConDist* parameterless and simplifies the application of the distance measure.

The rest of the paper is organized as follows: Related work on categorical distances measures and their approaches for identifying correlated context attributes are discussed in Section 2. Section 3 gives a short description of the categorical distance measure *ConDist*. Section 4 introduces the proposed method for the automatic threshold calculation. Section 5 gives an experimental evaluation of the proposed automatic threshold calculation method and the results are discussed in Section 6. The last section summarizes the paper.

2 Related Work

Unsupervised categorical distance measures may be divided into distance calculation (I) without considering context attributes and (II) considering context attributes.

Boriah et al. [4] give a comprehensive overview of distances measures from category (I). These distance measures ignore 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.

Distance measures from category (II) consider context attributes in the distance calculation process [1,8,9,10,11,14]. For example, the distance measures

proposed in [1] and [11] use all context attributes for distance calculation without distinguishing between correlated and uncorrelated. Conversely, the proposed distance measures in [8] and [9] use only a subset of context attributes for distance calculation. Jia and Cheung [9] use a normalized version of the mutual information (NMI) [3], whereas DILCA [8] relies on Symmetric Uncertainty (SU) [17] to determine the correlation between two attributes. For both, NMI and SU, the user has to define a threshold for the selection of correlated context attributes. The distance measure *CBDL* [10] uses the *Pearson's chi-squared test* χ^2 [12] for identifying correlated context attributes. Yet, the user needs to provide a significance level alpha for the *Pearson's chi-squared test* χ^2.

Like [9], *ConDist* [14] only uses correlated context attributes for distance calculation. It measures the correlation between attributes based on the information theoretical concept of entropy. In [14], the user has to define a threshold θ for the selection of correlated context attributes. In this work, we propose an automatic threshold calculation method for *ConDist*, which is based on the value distribution of the attributes and the number of objects in the data set.

3 The Distance Measure ConDist

In this section, we give a short description of the categorical distance measure *ConDist* [14]. The core idea is presented in Section 3.1. Since *ConDist* uses correlated context attributes in the distance calculation process, we explain in Section 3.2 how the set of correlated context attributes is derived. Section 3.3 describes the impact function of *ConDist* which accounts for the varying amount of information that can be extracted from a correlated context attribute.

3.1 ConDist

The distance between two objects A and B is calculated as the sum of distances in each attribute and defined as follows:

$$ConDist(A,B) = \sum_{X} w_X \cdot \frac{d_X(A,B)}{d_{X,max}}, \tag{1}$$

where w_X denotes a weighting factor assigned to attribute X. Since w_X is not relevant for threshold calculation, the reader is referred to [14] for further details on w_X. The function $d_X(A,B)$ denotes the distance of the values A_X and B_X of the objects A and B in attribute X. The maximum distance between any two values $x, u \in dom(X)$ of attribute X is given by $d_{X,max}$ and is used to normalize all attribute distances to the interval $[0,1]$.

The distance $d_X(A,B)$ between two values A_X and B_X within an attribute X is calculated according to the following formula:

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

where $dom(Y)$ is the domain of attribute Y, and $p(y|A_X) = p(y|X = A_X)$ 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 data set D. The set of correlated context attributes for a specific target attribute X is given by $context_X$ (see Section 3.2). The function $impact_X(Y)$ controls the influence of context attribute Y on target attribute X and is described in Section 3.3.

3.2 Selection of Context Attributes

ConDist uses an asymmetric function $cor(X|Y)$ to measure the correlation between a target attribute X and a context attribute Y. The function $cor(X|Y)$ is defined as follows:

$$cor(X|Y) = \frac{IG(X|Y)}{H(X)}, \tag{3}$$

where $H(X)$ is the entropy of the target attribute X and $IG(X|Y)$ is the information gain of target attribute X given context attribute Y. The information gain $IG(X|Y)$ is the difference between the entropy $H(X)$ of attribute X and the conditional entropy $H(X|Y)$ of attribute X given attribute Y:

$$IG(X|Y) = H(X) - H(X|Y) \tag{4}$$

Consequently, the function $cor(X|Y)$ is normalized to the interval $[0, 1]$. The higher the value of the correlation function $cor(X|Y)$, the higher the correlation between the two attributes. In [14], all context attributes whose correlations exceed a user-defined threshold θ are added to the set of correlated context attributes $context_X$ for target attribute X:

$$context_X = \{Y \mid cor(X|Y) \geq \theta\} \tag{5}$$

Note that the target attribute X itself is always in the set of correlated context attributes $context_X$ since $cor(X|X) = 1$.

3.3 The Impact of Context Attributes

ConDist uses an impact function $impact_X(Y)$ to control the influence of a correlated context attribute Y on target attribute X in the distance calculation process. This function accounts for the fact that the varying amount of extractable information depends on the degree of correlation between the attributes X and Y. In general, the quality of the extracted information grows with the strength of the correlation. However, for highly correlated attributes, the amount of extractable information decreases. In the extreme case of a perfectly correlated context attribute Y, no further information about distinct values in target attribute X can be extracted since Y predicts the values of X. To be precise, *ConDist* uses the impact function as defined as:

$$impact_X(Y) = cor(X|Y)\Big(1 - \frac{1}{2}cor(X|Y)\Big)^2, \tag{6}$$

where $cor(X|Y)$ is the correlation function introduced in Section 3.2.

4 Automatic Threshold Calculation Method

In this section, we propose a data-driven approach to replace the user-defined threshold θ of Section 3.2. In principle, *ConDist's* impact function should control automatically the influence of the context attributes without additional thresholds. However, the experiments in [14] showed that an additional threshold θ is necessary, especially for non-correlated data sets.

In Section 4.1, we use an example to explain the reason why an additional threshold is necessary. Based on that example, we propose a way how this threshold could be calculated from the data set in Section 4.2. The proposed automatic threshold calculation method involves an additional adjustment of *ConDist's* impact function which is described in Section 4.3.

4.1 Problem Description by Example

The impact function $impact_X(Y)$ (Section 3.3) controls the influence of context attributes in the distance calculation process and depends on the value of the correlation function $cor(X|Y)$ (Section 3.2). We give an example when these two functions fail to control the influence of context attributes without additional threshold θ.

Table 1: Example data set which describes eight people with three categorical attributes *sex*, *height* and *haircolor*.

#	sex	haircolor	height
1	male	brown	tall
2	male	blond	tall
3	male	black	medium
4	male	brown	medium
5	female	blond	medium
6	female	black	small
7	female	brown	small
8	female	blond	small

Consider the example data set in Table 1. Let us assume, we want to calculate distances for the attribute *height*. In this case, *sex* and *haircolor* are the context attributes. Further, we may assume that in the considered population attributes *haircolor* and *height* are independent of each other, while attributes *height* and *sex* are correlated. When applying *ConDist's* correlation function $cor(X|Y)$ and impact function $impact_X(Y)$, we achieve the following results:

$$cor(height|sex) = \frac{IG(height|sex)}{H(height)} \approx \frac{1.561 - 0.906}{1.561} \approx 0.420 \tag{7}$$

$$cor(height|haircolor) = \frac{IG(height|haircolor)}{H(height)} \approx \frac{1.561 - 1.439}{1.561} \approx 0.122 \tag{8}$$

$$impact_{height}(sex) \approx 0.262 \tag{9}$$

$$impact_{height}(haircolor) \approx 0.108 \tag{10}$$

As expected, the context attribute *sex* has higher impact on the target attribute *height* than context attribute *haircolor*. However, the context attribute *haircolor* has also a small impact on the target attribute *height*. Since we have also a highly correlated context attribute *sex*, the small impact of context attribute *haircolor* is almost negligible.

However, if we would have only the context attribute *haircolor*, the small impact factor would lead to small differences for distinct values in target attribute *height*. These small differences originate from the fact that the estimated probability density functions used in $cor(X|Y)$ are not representative due to the small training data set. Consequently, the differences are conceptually not intended since, given the particular population of our example, the context attribute *haircolor* is independent from *height*. In this case, it would be preferable to use only the target attribute itself for distance calculation. Therefore, a threshold θ is necessary to purge such context attributes.

4.2 Data-Driven Threshold Calculation

The example in Section 4.1 shows that too small data sets are problematic for the correlation function $cor(X|Y)$. This follows from the fact that $cor(X|Y)$ requires the information gain $IG(X|Y)$, which in turn requires the entropy of attribute X and the conditional entropy of attribute X given attribute Y. The entropy $H(X)$ and the conditional entropy $H(X|Y)$ are defined as follows:

$$H(X) = -\sum_{x \in dom(X)} p(x) \log_2 \big(p(x)\big) \text{ and} \tag{11}$$

$$H(X|Y) = -\sum_{y \in dom(Y)} p(y) \sum_{x \in dom(X)} p(x|y) \log_2 \big(p(x|y)\big), \tag{12}$$

where $p(x)$ is the probability of value x and $p(x|y)$ is the conditional probability of value x given value y in data set D. Consequently, the probability density functions $p(X)$ and $p(Y)$ of the attributes X and Y are necessary for calculating $H(X)$ and $H(X|Y)$. These two functions can be estimated more accurately if the data set is large. Consequently, the smaller the data set, the higher the possibility of errors in the results delivered by the correlation function.

Further, two attributes X and Y are non-correlated in *ConDist's* correlation function $cor(X|Y)$, if and only if the following equation holds:

$$H(X) = H(X|Y) \tag{13}$$

Equation (13) requires that the conditional probability density functions of attribute X given a value $y \in dom(Y)$ are all identical and equal to the probability density function of attribute X. The larger the cardinality of $dom(X)$ and $dom(Y)$, the more objects are necessary to fulfill this requirement in the case

of non-correlated attributes since the value distributions and conditional value distributions must be estimated from the data set. Consequently, the cardinality and the distribution of $dom(X)$ and $dom(Y)$ should be considered in the threshold calculation process as well. Both factors are reflected in the entropy of an attribute.

Therefore, we calculate the threshold $\theta_{X|Y}$ based on these two aspects:

$$\theta_{X|Y} = \frac{H(X) \cdot H(Y)}{n}, \tag{14}$$

where n is the number of objects in the data set. This threshold decreases with an increasing number of objects and increases with increasing attribute entropies $H(X)$ and $H(Y)$. The threshold $\theta_{X|Y}$ may be viewed as an estimate of the portion of correlation that is due to estimating the probability density functions $p(X)$ and $p(Y)$ from the data set. The calculation of $\theta_{X|Y}$ is easy and no user-defined parameter is necessary.

If we apply the automatic calculation of the threshold $\theta_{X|Y}$ to the example in the Section 4.1, we can observe the following results:

$$\theta_{height|sex} = \frac{H(height) \cdot H(sex)}{n} \approx \frac{1.561 \cdot 1}{8} \approx 0.195 \text{ and} \tag{15}$$

$$\theta_{height|haircolor} = \frac{H(height) \cdot H(haircolor)}{n} \approx \frac{1.561 \cdot 1.561}{8} \approx 0.305. \tag{16}$$

The correlation value of attribute *sex* (see Equation (7)) exceeds the calculated threshold $\theta_{height|sex}$, whereas the correlation value of attribute *haircolor* (see Equation (8)) does not exceed the threshold $\theta_{height|haircolor}$. Applying the proposed context-sensitive threshold $\theta_{X|Y}$ would imply that only the attribute *sex* would be added to the set of correlated context attributes $context_{height}$ for target attribute *height*.

4.3 Adjustment of the Impact Function

In Section 4.2, we interpreted the threshold $\theta_{X|Y}$ as the amount of correlation which is caused by estimating probability density functions from the data set. Consequently, this amount of correlation should also be considered in the impact function $impact_X(Y)$. To that end, we adjust *ConDist's* impact function $impact_X(Y)$ as follows:

$$impact_X(Y) = \begin{cases} 0 & \text{if } cor(X|Y) \leq \theta_{X|Y} \\ cor_\theta(X|Y)\Big(1 - \frac{1}{2} cor_\theta(X|Y)\Big)^2 & \text{if } cor(X|Y) > \theta_{X|Y} \end{cases}, \tag{17}$$

where $cor_\theta(X|Y)$ is the adjusted correlation value rescaled to the interval $[0, 1]$:

$$cor_\theta(X|Y) = \frac{cor(X|Y) - \theta_{X|Y}}{1 - \theta_{X|Y}}. \tag{18}$$

If we apply the new impact function to the example in the Section 4.1, we can observe the following results:

$$cor_\theta(height|sex) \approx \frac{0.420 - 0.195}{1 - 0.195} \approx 0.280 \tag{19}$$

$$impact_{height}(sex) \approx 0.280\big(1 - \frac{1}{2} \cdot 0.280\big)^2 \approx 0.207 \tag{20}$$

$$impact_{height}(haircolor) = 0 \tag{21}$$

Using the proposed approach, only the context attribute *sex* has an impact on target attribute *height*.

5 Experimental Evaluation

This section presents an experimental evaluation of the automatic calculation of the threshold $\theta_{X|Y}$ (Section 4). We compare our new approach with the user-defined threshold method presented in [14] and with the categorical distance measure *DILCA* [8], which is the most serious competitor in [14]. For *DILCA*, we used the non-parametric approach $DILCA_{RR}$ as described in [8].

5.1 Evaluation Methodology

We evaluate the different threshold calculation methods for *ConDist* in the context of classification. A k-Nearest-Neighbor classifier is used to compare the different categorical distance measures (*DILCA* and *ConDist*) and the different methods for threshold calculation in *ConDist*. For simplification, we do not try to optimize the selection of the parameter k of the k-Nearest-Neighbor classifier. Instead we fix the number of neighbors $k = 7$ in all tests in order to create an equal base for the different configurations. We evaluate by 10-fold-cross validation and use the classification accuracy as evaluation measure. To reduce confounding effects of the generated subsets, 10-fold cross-validation is repeated 100 times with different subsets for each data set.

For evaluation, the *multivariate categorical data sets for classification* from the UCI machine learning repository [13] are chosen. We exclude data sets with less than 25 objects (e.g., *Balloons*) or mainly binary attributes (e.g., *Chess*). Furthermore, we include some *multivariate mixed data sets for classification* from [13] 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 Wisconsin, Dermatology* and *Post-Operative Patient*. All integer attributes are treated as categorical. The final set of data sets is given in Table 2. The column *Correlation* contains the average correlation between each distinct pair of attributes, calculated by the function $cor(X|Y)$, see Equation (3). The value ranges from 0 if no correlation exists to 1 if all attributes are perfectly correlated. The data sets are separated in two groups: correlated (Correlation > 0) and non-correlated (Correlation $= 0$).

Table 2: Characteristics of the data sets.

Data Sets	Instances	Attributes	Classes	Correlation
Teaching Assistant Evaluation	151	5	3	0.336
Soybean Large	307	35	19	0.263
Breast Cancer Wisconsin	699	10	2	0.216
Dermatology	366	34	6	0.098
Lymphography	148	18	4	0.070
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

5.2 Experimental Setup and Results

This experiment compares the automatic calculated threshold $\theta_{X|Y}$ with various user-defined thresholds θ in *ConDist* and with the categorical distance measure *DILCA*. The threshold θ expresses the minimum value of the function $cor(X|Y)$ that a context attribute Y has to achieve in order to be selected as correlated 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. The automatic calculated threshold $\theta_{X|Y}$ follows the approach of Section 4. The results of this experiment are summarized in Table 3, where each column contains the average classification accuracies for a particular threshold.

Table 3 shows that the automatic calculation of the threshold $\theta_{X|Y}$ achieves the best average classification accuracy. The user-defined thresholds $\theta = 0.01$ and $\theta = 0.02$ achieve similar good results. Without any threshold $\theta = 0$, a decreasing classification accuracy can be observed for non-correlated data sets. For too high user-defined thresholds θ, the average classification accuracies decrease. Compared with *DILCA*, the proposed approach $\theta_{X|Y}$ is comparable for highly correlated data sets and superior for weakly- and non-correlated data sets.

Statistical Significance Test. This test aims at examining if the differences in Table 3 are statistically significant. Demšar [5] deals with the statistical comparison of classifiers over multiple data sets. They recommend the Wilcoxon Signed-Ranks Test [16] for the comparison of two classifiers and the Friedman-Test [6,7] for the comparison of multiple classifiers. Following this line, we use the Friedman-Test to compare all different configurations and the Wilcoxon Signed-Ranks Test for post-hoc tests. The Friedman-Test is significant for $p < 0.05$; thus we can reject the null hypothesis that all threshold calculation methods in *ConDist* and

Table 3: Classification accuracies for the proposed automatic threshold calculation (column $\theta_{X|Y}$), various user-defined thresholds and *DILCA*. Each column contains the results for a specific threshold, e.g. the column 0.02 contains the results for $\theta = 0.02$.

		ConDist							*DILCA*
Data Set	$\theta_{X\|Y}$	0	0.01	0.02	0.05	0.1	0.2	1.0	$DILCA_{RR}$
Teaching A. E.	49.93	49.85	49.85	49.85	49.71	48.74	48.74	45.84	**50.86**
Soybean Large	91.76	91.74	91.74	91.79	**91.82**	89.75	89.36	91.30	91.48
B. C. Wisconsin	96.17	96.13	96.13	96.13	96.13	96.15	**96.25**	95.25	95.55
Dermatology	96.70	96.74	96.74	96.76	96.81	96.35	96.23	95.90	**97.97**
Lymphography	**83.36**	**83.36**	**83.36**	83.30	83.01	81.99	82.01	81.26	82.09
Hayes-Roth	68.59	68.11	68.36	68.50	**69.21**	64.47	64.47	61.74	67.59
Audiology-Std.	66.22	66.33	66.27	66.27	**66.56**	65.41	61.81	61.35	62.31
Postoperative P.	69.71	**69.83**	69.81	69.62	**69.83**	68.27	68.58	68.59	68.22
TicTacToe	**99.99**	**99.99**	**99.99**	**99.99**	94.74	94.74	94.74	94.74	90.65
Car	**90.56**	88.98	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**	**90.56**	90.25
Monks	**97.32**	95.16	**97.32**	**97.32**	**97.32**	**97.32**	**97.32**	**97.32**	92.06
Balance-Scale	**78.66**	77.35	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**	**78.66**	78.43
Nursey	**94.94**	94.43	**94.94**	**94.94**	**94.94**	**94.94**	**94.94**	**94.94**	92.61
Average	**83.38**	82.92	83.36	83.36	83.02	82.10	81.82	81.34	81.53

DILCA are equivalent. Subsequently, we applied the Wilcoxon Signed-Ranks Test with $\alpha = 0.05$ on the classification accuracies of Table 3.

Table 4 shows significant differences between $\theta_{X|Y}$ and *DILCA* and between $\theta_{X|Y}$ and the user-defined thresholds $\theta = 0.1$, $\theta = 0.2$ and $\theta = 1.0$. For the remaining user-defined thresholds θ, the Wilcoxon Signed-Ranks Test shows no statistically significant differences.

6 Discussion

For correlated data sets, high user-defined thresholds θ lead to decreasing results, e.g. $\theta = 0.1$, $\theta = 0.2$ or $\theta = 1.0$ for the data sets *Teaching Assistant Evaluation* and *Lymphography*. For these thresholds, many useful correlated context attributes are discarded. The same observation can be made for weakly-correlated data sets at lower thresholds. Consider the decreasing classification accuracy for the data set *TicTacToe* at threshold $\theta = 0.05$. For non-correlated data sets, nearly all threshold methods achieve the same results. Only the absence of any threshold ($\theta = 0$) leads to inferior results. In this case, non-correlated context attributes are added to the set of context attributes $context_X$, which may contribute noise to the distance calculation process.

The proposed automatic calculation of the threshold $\theta_{X|Y}$ achieves good results for correlated and non-correlated data sets. As a consequence, the proposed method achieves the best average classification accuracy. The average classification accuracies for user-defined thresholds $\theta = 0.01$, $\theta = 0.02$ and $\theta = 0.05$ are

Table 4: Results of the Wilcoxon Signed-Ranks Test comparing the classification accuracies of the automatic calculation of the threshold $\theta_{X|Y}$ with various user-defined thresholds θ and with $DILCA$. The first row contains the calculated p-values, the second row contains the result of the Wilcoxon Signed-Ranks Test: *yes*, if $\theta_{X|Y}$ performs significantly different, *no* otherwise.

	$\theta = 0$	$\theta = 0.01$	$\theta = 0.02$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 1$	$DILCA$
p-value	0.0830	0.7998	0.2070	1	0.0092	0.0113	0.0092	0.0231
significant	no	no	no	no	yes	yes	yes	yes

marginally worse. The Wilcoxon-Signed Ranks Test confirms that there are no statistical significant differences between them. In contrast to this, statistically significant differences can be observed for too high user-defined thresholds.

These observations indicate that the proposed automatic calculation of the threshold $\theta_{X|Y}$ is superior to poorly selected user-defined thresholds and competitive to well selected user-defined thresholds. Consequently, $\theta_{X|Y}$ is preferable to the user-defined approach in [14], since the user-defined parameter θ is omitted and the quality of results does not deteriorate.

For highly correlated data sets, the results of the proposed approach $\theta_{X|Y}$ and $DILCA$ are comparable. For weakly- and non-correlated data sets, $DILCA$ achieves inferior results in comparison to *ConDist*. This is because $DILCA$ uses only context attributes for distance calculation which results in random distances if all context attributes are non-correlated.

7 Summary

Categorical distance calculation is a key requirement for many data mining tasks. In this paper, we propose an extension for the unsupervised categorical distance measure *ConDist* [14]. *ConDist* uses the correlation between attributes to extract available information for distance calculation. In [14], the user has to define a threshold θ for the selection of correlated context attributes. This threshold θ has to purge context attributes whose correlations are caused by noisy, non-representative or too small data sets.

In this work, we proposed an automatic threshold calculation method for the distance measure *ConDist*. This approach calculates for each pair of target attribute X and context attribute Y an individual threshold instead of using a single user-defined threshold θ. The calculated thresholds $\theta_{X|Y}$ depend on the number of objects in the data set and the entropies of the attributes. Consequently, these individual thresholds can be better adapted to the specific correlation requirements of each pair of attributes. Further adjustments were made to *ConDist's* impact function $impact_X(Y)$.

The proposed extension makes *ConDist* parameterless and simplifies the application of the distance measure. Our experiments show that the automatic threshold calculation method is competitive to well selected user-defined thresh-

olds θ and superior to poorly selected user-defined thresholds θ. For these two reasons, the proposed approach is preferable to the user-defined approach in [14].

Acknowledgements 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. HO 2586/2-2).

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. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200), 675–701 (1937)
7. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics 11(1), 86–92 (1940)
8. Ienco, D., Pensa, R.G., Meo, R.: Context-based distance learning for categorical data clustering. In: Advances in Intelligent Data Analysis VIII, pp. 83–94. Springer (2009)
9. Jia, H., Cheung, Y.M.: A new distance metric for unsupervised learning of categorical data. In: Proc. of IJCNN. pp. 1893–1899. IEEE (2014)
10. Khorshidpour, Z., Hashemi, S., Hamzeh, A.: Cbdl: Context-based distance learning for categorical attributes. Int. J. Intell. Syst. 26(11), 1076–1100 (2011)
11. Le, S.Q., Ho, T.B.: An association-based dissimilarity measure for categorical data. Pattern Recognition Letters 26(16), 2549–2557 (2005)
12. Lehmann, E., Romano, J.: Testing Statistical Hypotheses. Springer Texts in Statistics, Springer (2005)
13. M. Lichman: Uci machine learning repository (2013), `http://archive.ics.uci.edu/ml`
14. Ring, M., Otto, F., Becker, M., Niebler, T., Landes, D., Hotho, A.: Condist: A context-driven categorical distance measure. In: Machine Learning and Knowledge Discovery in Databases. pp. 251–266. Springer (2015)
15. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley Boston (2006)
16. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics bulletin 1(6), 80–83 (1945)
17. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. vol. 3, pp. 856–863 (2003)