Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition Letters 26 (2005) 2549–2557

Pattern Recognition Letters

www.elsevier.com/locate/patrec

An association-based dissimilarity measure for categorical data

Si Quang Le, Tu Bao Ho *

School of Knowledge Science, Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa 923-1292, Japan

Received 25 September 2004; received in revised form 18 May 2005
Available online 28 July 2005

Communicated by A. Fred

Abstract

In this paper, we propose a novel method to measure the dissimilarity of categorical data. The key idea is to consider the dissimilarity between two categorical values of an attribute as a combination of dissimilarities between the conditional probability distributions of other attributes given these two values. Experiments with real data show that our dissimilarity estimation method improves the accuracy of the popular nearest neighbor classifier.

Keywords: Dissimilarity measures; Categorical data; Conditional probability distribution; Hypothesis testing; Nearest neighbor

1. Introduction

Measuring the (dis)similarity between data objects is one of the primary tasks for distance-based techniques in data mining and machine learning, e.g., distance-based clustering and distance-based classification. In this task, measuring (dis)similarity in categorical data is a challenging problem because the categorical data do not have any structures, and thus only an identical comparison operation can be applied.

The most common similarity measures for categorical data are binary vector-based methods (Liebetrau, 1983; Krantz et al., 1971; Baulieu, 1989; Gower, 1971; Gower and Legendre, 1986; Albert, 1983; Jaccard, 1912; Batagelj and Bren, 1995; Hubálek, 1982). These methods transform each data object into a binary vector, at which each bit indicates the presence or absence of a possible attribute value. Then the similarity between two objects is estimated by the similarity between two corresponding binary vectors. The most popular measures for binary vectors belong to two

* Corresponding author. Tel./fax: +81 761 51 1730.
E-mail addresses: quang@jaist.ac.jp (S.Q. Le), bao@jaist.ac.jp (T.B. Ho).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.06.002

families S_θ and T_θ introduced by Gower and Legendre (1986). These methods are simple, but they have two main drawbacks: (1) the transformation of data objects into binary vectors, in which making the similarity between two values either 0 or 1 may leave out many subtleties of the data; (2) they do not take into account the correlations between attributes that typically exist in real-life data and are potentially concerned with the difference among attribute values.

In addition to the binary vector-based methods, similarity measure methods for mixed numerical data (Gowda and Diday, 1991a,b, 1992; de Carvalho, 1994; de Carvalho, 1998; Goodall, 1966; Ichino and Yaguchi, 1994) can also be applied to categorical data. In (Goodall, 1966), Goodall proposed a statistical approach, in which uncommon attribute values make greater contributions to the overall similarity between two objects than common attribute values. The overall similarity is estimated by combining similarities between values pairs by using Lancaster's method (Lancaster, 1949). Setting aside the statistical approach, algebraic methods have been also proposed (Gowda and Diday, 1991a,b, 1992; de Carvalho, 1994; de Carvalho, 1998; Ichino and Yaguchi, 1994). In (Gowda and Diday, 1991a,b, 1992), the similarity between two values of an attribute is based on three factors: (1) the relative position of two values, position; (2) the relative sizes of two values without referring to common parts, span; (3) the common parts between two values, content. Similarly, the sizes of the union (the joint operation $\otimes$) and the intersection (the meet operation $\oplus$) of two attribute values are also taken into account (de Carvalho, 1994; de Carvalho, 1998; Ichino and Yaguchi, 1994). Subsequently, similarities of all attributes are integrated into the similarity between objects by using Minkowski distance.

In principle, the methods mentioned above can be considered direct methods because the dissimilarity between two attribute values is synthesized directly from the values. In this paper, we present a novel indirect method to measure the dissimilarity for categorical data. It is called indirect in the sense that the dissimilarity between two values of an attribute is indirectly estimated by using relations between other attributes under the condition of giving these two values. The method is composed of two iterative steps. First, the dissimilarity between two values of an attribute is estimated as as the sum of the dissimilarities between conditional probability distributions of other attributes given these two values. Then, the dissimilarity between two data objects is the sum of dissimilarities of their attribute value pairs. We investigate the efficiency of the proposed method in terms of theoretical properties and experiments with real data. Both theoretical proofs and experiments show that the method is not proper for data sets with independent attributes. Fortunately, experiments with real data show that attributes are typically correlated.

The rest of this paper is organized as follows. In Section 2 we describe the proposed measure in detail. In Section 3 we investigate the proposed measure's properties and its computational complexity. Experiments with real data are presented in Section 4. Conclusions, suggestions for drawbacks and further work are given lastly.

2. Association-based dissimilarity

2.1. Similarity measure

In the following we introduce some notations: Let $A_1, \ldots, A_m$ be m categorical attributes and $\mathrm{dom}(A_i)$ be the domain of attribute A_i. Let $D \subseteq A_1 \times \cdots \times A_m$ denotes a data set and $\boldsymbol{x} = (x_1, \ldots, x_m)$ where $x_i \in \mathrm{dom}(A_i)$ denote a data object of D. Let $p(A_j = v_j | A_i = v_i)$ be the conditional probability of $A_j = v_j$ given that $A_i = v_i$. More generally, let $\mathrm{cpd}(A_j | A_i = v_i)$ be the conditional probability distribution of attribute A_j given that attribute A_i holds value v_i.

The first, and perhaps the most important step, is to estimate the dissimilarity between two values of an attribute. To motivate the method, consider a data set D with n objects described by two attributes: Shape $= \{\square, \diamond, \triangle\}$ and Color $= \{R, G, B\}$. We suppose that n is large enough that conditional probabilities $p(A_j = v_j | A_i = v_i)$ and conditional probability distributions $\mathrm{cpd}(A_j | A_i = v_i)$ can be approximately estimated from data set D as shown in Table 1. Now in considering the relation between the two attributes Shape and Color,

Table 1
Example: The correlation between attribute Color (dom(Color) = $\{R, G, B\}$), and attribute Shape (dom(Shape) = $\{\square, \diamond, \triangle\}$)

	Contingent table				Contingent probability table			
	R	G	B		$p(R\|.)$	$p(G\|.)$	$p(B\|.)$	
□	50	40	10	100	0.5	0.4	0.1	1.0
◇	21	21	28	70	0.3	0.3	0.4	1.0
△	24	24	72	120	0.2	0.2	0.6	1.0

an important observation is that conditional probability distribution cpd(Color|Shape = □) is closer to cpd(Color|Shape = ◇) than cpd(Color|Shape = △). On the other hand, the nature of observable dissimilarities also indicates that the dissimilarity between □ and ◇ is somehow smaller than the dissimilarity between □ and △. These observations suggest that the dissimilarity between two values of attribute Shape can be estimated from the conditional probability distributions of the attribute Color given these two attributes.

Now we are ready to define the dissimilarity between two values v_i and v_i' of attribute A_i given that the data set D is composed of m different attributes.

Definition 1. The dissimilarity between two values v_i and v_i' of attribute A_i, denoted by $\phi_{A_i}(v_i, v_i')$, is the sum of dissimilarities between conditional probability distributions of other attributes given that attribute A_i holds values v_i and v_i':

$$\phi_{A_i}(v_i, v_i') = \sum_{j,j\neq i} \psi(\mathrm{cpd}(A_j|A_i = v_i), \mathrm{cpd}(A_j|A_i = v_i')), \tag{1}$$

where $\psi(.,.)$ is a dissimilarity function for two probability distributions.

Definition 1 means that the dissimilarity between two values v_i and v_i' of attribute A_i is directly proportional to dissimilarities between the conditional probability distributions of other attributes given these values. Thus, the great (small) dissimilarity between these conditional probability distributions leads to the great (small) dissimilarity between v_i and v_i'.

To date, several dissimilarity measures between probability distributions have been proposed (Lin, 1991; Rached et al., 2001; Kullback, 1959; Kullback and Leibler, 1951). In this paper, we use the most popular one, the Kullback–Leibler divergence method (Kullback, 1959; Kullback and Leibler, 1951)

$$\mathrm{KL}(P, P') = \sum_x \left(p(x) \lg \frac{p(x)}{p'(x)} + p'(x) \lg \frac{p'(x)}{p(x)} \right), \tag{2}$$

where lg is a logarithm having base 2.

To illustrate our method, the dissimilarities of value pairs (□, ◇) and (□, △) as given in Table 1 are computed as follows:

$$\phi_{\mathrm{Shape}}(\square, \diamond) = .5 \lg \frac{.5}{.3} + .3 \lg \frac{.3}{.5} + .4 \lg \frac{.4}{.3} + .3 \lg \frac{.3}{.4} + .1 \lg \frac{.1}{.4} + .4 \lg \frac{.4}{.1} = .8$$

$$\phi_{\mathrm{Shape}}(\square, \triangle) = .5 \lg \frac{.5}{.2} + .2 \lg \frac{.2}{.5} + .4 \lg \frac{.4}{.2} + .2 \lg \frac{.2}{.4} + .1 \lg \frac{.1}{.6} + .6 \lg \frac{.6}{.1} = 1.9$$

Having defined the dissimilarity between values of an attribute, now dissimilarities of different attributes are combined to estimate the dissimilarity between two data objects.

Definition 2. The dissimilarity between two data objects $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted by $\phi(\boldsymbol{x}, \boldsymbol{y})$, is the sum of dissimilarities of their attribute value pairs:

$$\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} \phi_{A_i}(x_i, y_i). \tag{3}$$

Definition 2 means that the smaller the dissimilarities of attribute value pairs of $\boldsymbol{x}$ and $\boldsymbol{y}$ are, the smaller the dissimilarity between $\boldsymbol{x}$ and $\boldsymbol{y}$.

2.2. Algorithm for computing similarities between data objects

In this subsection, we present a three-step algorithm to measure the dissimilarities of all pairs of data objects of a data set D (see Fig. 1). At the first step, all conditional probabilities $p(A_j = v_j|$

```
Algorithm for computing similarities between data objects
Input: Data set $D$
Output: Dissimilarities between data objects.

BEGIN
    Step 1. Estimate all conditional probabilities $p(A_j = v_j | A_i = v_i)$.
    Step 2. For any pair of values $v_i$ and $v'_i$ of attribute $A_i$, compute
        $\phi_{A_i}(v_i, v'_i) = \sum_{v_j \in dom(A_j), j \neq i} \left( p(v_j|v_i) \lg \frac{p(v_j|v_i)}{p(v_j|v'_i)} + p(v_j|v'_i) \lg \frac{p(v_j|v'_i)}{p(v_j|v_i)} \right)$
    Step 3. For any data object pairs $(\mathbf{x},\mathbf{y})$, compute
        $\phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1} \phi_{A_i}(x_i, y_i)$
END
```

Fig. 1. Algorithm for computing similarities between data objects.

$A_i = v_i$) are estimated from data set D. Then the dissimilarities of the value pairs are computed based on based on the probabilities $p(A_j = v_j|A_i = v_i)$ conditioned on these values. Finally, the dissimilarities of data object pairs are determined using Eq. (3).

Let us now turn our attention the complexity of the algorithm, given that data set D consists of n objects which are composed of m attributes. At the first step, estimating all conditional probabilities $p(A_j = v_j|A_i = v_i)$ is done in $O(nm^2)$ time. Then it takes $O(m_v^3)$ time to compute the dissimilarities of all pairs of attribute values where m_v is the number of attribute values. Finally, all dissimilarities between data objects are determined in $O(n^2m)$ time. Overall, the complexity of the algorithm is $O(nm^2) + O(m_v^3) + O(n^2m) = O(n^2m)$ because m and m_v are typically smaller than n. Thus, this complexity is at the same level as other commonly used direct measure methods.

3. Characteristics

In this section, we investigate the properties of the proposed measure.

Proposition 1. *For any data object pair* $(\boldsymbol{x},\boldsymbol{y})$, *it holds true for*:

(1) $\phi(\boldsymbol{x},\boldsymbol{y}) \geqslant 0$
(2) $\phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{y},\boldsymbol{x})$
(3) $\phi(\boldsymbol{x},\boldsymbol{x}) = 0$

Proof

(1) $\phi(\boldsymbol{x},\boldsymbol{y}) \geqslant 0$: Since KL dissimilarity between two probability distributions is non-negative, the dissimilarity of two values x_i and y_i is non-negative

$$\phi_{A_i}(x_i, y_i) = \sum_{j=1, i \neq j}^{m} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = x_i), \mathrm{cpd}(A_j|A_i = y_i)) \geqslant 0.$$

This implies that

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \phi_{A_i}(x_i, y_i) \geqslant 0.$$

(2) $\phi(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{y},\boldsymbol{x})$: Since KL dissimilarity between two probability distributions is symmetric, the dissimilarity between two values x_i and y_i is also symmetric

$$\phi_{A_i}(x_i, y_i) = \sum_{j=1}^{m} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = x_i), \mathrm{cpd}(A_j|A_i = y_i))$$
$$= \sum_{j=1}^{m} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = y_i), \mathrm{cpd}(A_j|A_i = x_i)) = \phi_{A_i}(y_i, x_i).$$

It means that

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \phi_{A_i}(x_i,y_i) = \sum_{i=1}^{m} \phi_{A_i}(y_i,x_i) = \phi(\boldsymbol{y},\boldsymbol{x}).$$

(3) $\phi(\boldsymbol{x},\boldsymbol{x}) = 0$: Since KL dissimilarity between two identical probability distributions is equal to 0, the dissimilarity between two identical values is equal to 0.

$$\phi_{A_i}(x_i,x_i) = \sum_{j=1}^{m} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = x_i), \mathrm{cpd}(A_j|A_i = x_i)) = 0.$$

It means that

$$\phi(\boldsymbol{x},\boldsymbol{x}) = \sum_{i=1}^{m} \phi_{A_i}(x_i,x_i) = 0. \qquad \square$$

Proposition 2. *The dissimilarity between two values v_i and v_i' of attribute A_i is zero if and only if the conditional probability distributions of other attributes, given that attribute A_i holds values v_i and v_i', are identical.*

$$\phi_v(v_i,v_i') = 0 \iff \mathrm{cpd}(A_j|A_i = v_i) \equiv \mathrm{cpd}(A_j|A_i = v_i') \quad \textit{for } j = 1 \ldots m,\ j \neq i.$$

Proof. Since KL dissimilarity between two probability distributions is non-negative, and equal to 0 if and only if the distributions are identical, the dissimilarity between two values v_i and v_i' is equal to 0 if and only if the conditional probability distributions of other attributes when A_i holds values v_i and v_i' are identical. It implies that

$$\phi_{A_i}(v_i,v_i') = \sum_{j=1,j\neq i}^{m} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = v_i), \mathrm{cpd}(A_j|A_i = v_i')) = 0,$$

which is equivalent to

$$\mathrm{cpd}(A_j|A_i = v_i) \equiv \mathrm{cpd}(A_j|A_i = v_i') \quad \text{for } j = 1 \ldots m, j \neq i. \qquad \square$$

Proposition 3. *If all attribute pairs are independent, dissimilarities between data objects are all equal to zero.*

Proof. Since A_i and A_j are independent for all i and j,

$$P(A_j = v_j|A_i = v_i) = P(A_j = v_j) = p(A_j = v_j|A_i = v_i'), \quad \forall v_i, v_j, v_i'.$$

It means that $\mathrm{cpd}(A_j|A_i = v_i)$ and $\mathrm{cpd}(A_j|A_i = v_i')$ is identical. It leads to

$$\mathrm{KL}(\mathrm{cpd}(A_j|A_i = x_i), \mathrm{cpd}(A_j|A_i = y_i)) = 0, \quad \forall x_i, y_i$$

and therefore

$$\phi_{A_i}(x_i,y_i) = \sum_{j} \mathrm{KL}(\mathrm{cpd}(A_j|A_i = x_i), \mathrm{cpd}(A_j|A_i = y_i)) = 0.$$

That is equivalent to

$$\phi(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \phi_{A_i}(x_i,y_i) = 0, \quad \forall \boldsymbol{x},\boldsymbol{y}. \qquad \square$$

Two points can be induced from Proposition 3 are

- When attributes can be divided into groups such that attributes of different groups are independent, the dissimilarity between two objects can be referred to as dissimilarities between these two objects with respect to groups individually. If we deem each of the attribute groups an independent aspect of objects, the dissimilarity between objects is considered with respect to aspects independently. This leads to an idea of replacing each of the attribute groups by one or a few attributes that can have more discriminating power. Investigating this idea however, is out of the scope of this paper.
- The proposed measure cannot be applied to databases whose attributes are absolutely independent. Discussions about this drawback are given in Section 5.

4. Evaluations

In this section we show the merit of our approach when it is applied to real data. To this end, we carried out two experiments: the first experiment analyzes the dependency between attributes of real data sets to investigate its impact to our method. In the second experiment, we compare the proposed measure with measures based on the two following ideas. The first one (Gower and Legendre, 1986) is to consider the similarity between objects based on the number of identical and nonidentical attribute value pairs. This idea includes most of popular similarity measures such as Russell and Rao (1940), Jaccard (1912), Dice (1945), and Sokal and Sneath (1963). The second idea is to consider the similarity between objects based on index probability introduced by Goodall in (Goodall, 1966). Similarity measures of these two ideas estimate the similarity between attribute values directly and are different from the proposed measures where the similarity between two attribute values are estimated indirectly based on relations of other attributes given these two values. To compare similarity measures, we combine these measures with the popular distance-based data mining method, nearest neighbor classifier (NN) (Cover and Hart, 1967), and analyze the accuracies of NN.

4.1. T_θ, S_θ and Goodall

Let $A_1,\ldots,A_m$ be m attributes, $U=\bigcup_{i=1}^{m}A_i$ be the set of all possible attribute values and $M=|U|$.

Denote $\boldsymbol{x}=(x_1,\ldots,x_m)$ and $\boldsymbol{y}=(y_1,\ldots,y_m)$ where $x_i, y_i \in A_i$ two data objects, and X and Y corresponding binary vectors with the length M, respectively. Clearly both X and Y have m 1 bits and $M-m$ 0 bits. Denote $\overline{X}$ the complementary of X and

- $a=XY$—the number of values which X and Y share.
- $b=X\overline{Y}$—the number of values which X has and Y lacks.
- $c=\overline{X}Y$—the number of values which X lacks and Y has.
- $d=\overline{XY}$—the number of values which both X and Y lacks.

Obviously, $a+b+c+d=M$.

In (Gower and Legendre, 1986), Gower and Legendre introduced two families of similarities

$$T_\theta=\frac{a}{a+\theta(b+c)} \quad\text{and}\quad S_\theta=\frac{a+d}{a+d+\theta(b+c)},$$

where $\theta>0$ to avoid negative values. These families of similarities contains many popular similarity measures for binary vectors such as Russell and Rao (1940), Jaccard (1912), Dice (1945), and Sokal and Sneath (1963).

Proposition 4. *T_θ and S_θ are increasing functions with a.*

Proof

- T_θ is a increasing function with a.
  Clearly, $a+b=m$, $a+c=m$. Thus $b=c=m-a$. So, T_θ is rewritten as
  $$T_\theta=\frac{a}{a+2\theta(m-a)}.$$
  We have
  $$T'_\theta(a)=\frac{2m\theta}{(a(\theta-1)-m\theta)^2}>0.$$
  Thus, T_θ is an increasing function with a.
- S_θ is an increasing function with a.
  Since $a+b+c+d=M$, $d=2m-M$. S_θ is rewritten as
  $$S_\theta=\frac{2a+M-2m}{2a+M-2m+2\theta(m-a)}.$$
  We have
  $$S'_\theta(a)=\frac{2M\theta}{(2a(\theta-1)+2m(1-\theta)-M)^2}>0.$$
  Thus, S_θ is an increasing function with a. □

Proposition 4 implies that the closest neighbors of a data object with respect to any similarity measures of family T_θ or S_θ are identical. It means that NN products the same accuracy when using any similarity measures of family T_θ or S_θ.

4.1.1. Goodall similarity measure

In (Goodall, 1966), Goodall introduced the idea of using index and probability to estimate the similarity between data objects. For each attributes, an order relationship between pairs of values is established so that the similarity between common identical values is less than that of the uncommon identical values. Having defined the order relation, the similarity of one value pair is defined as the probability of picking up randomly a value pair that is less than or equally similar to this pair. After that the similarities of attribute value pairs of two objects are integrated by Fisher or Lancaster transforms (Fisher, 1950; Lancaster, 1949) into the similarity of these objects.

4.2. Data sets

We used 30 diverse data sets from UCI (Blake and Merz, 1998), for which numerical attributes are automatically discretized using the data mining system CBA (Liu et al., 1998). Details of these data sets can be found in Table 2.

4.3. Experimental methodology

4.3.1. Dependency analysis

For each data set D, we estimate dependency between attributes by a dependency factor $\rho(D)$ that is the proportion of the number of dependent attribute pairs and the total number of attribute pairs

$$\rho(D) = \frac{|\{(A_i, A_j)\ :\ A_i \text{ and } A_j \text{ are dependent}\}|}{m(m-1)},$$

where $\rho(D)$ is directly proportional to the dependency between attributes of D. Thus, $\rho(D)$ is 100% when all attribute pairs are dependent and 0% when they are all independent.

To estimate the dependency of two attributes, we used the χ^2 test with a 95% significance level.

4.3.2. Accuracy analysis

We compared the accuracy of NN in combination with the proposed measure (denoted μ_1) with the accuracy of NN when combined with a similarity measure of family T_θ or S_θ, or Goodall (denoted μ_0), using the 10-time 10-fold cross-validation strategy and hypothesis testing as follows:

$$H_0 : \mu_0 = \mu_1 \quad \text{versus} \quad H_1 : \mu_0 < \mu_1.$$

Since each 10-time 10-fold cross-validation result contains 100 trials, the difference between μ_0 and μ_1 follows the normal distribution

$$z = \frac{\mu_1 - \mu_0}{\sqrt{\frac{\delta_0^2}{100} + \frac{\delta_1^2}{100}}},$$

where δ_0 and δ_1 are the deviations of the accuracy test results of NN with a measure of S_θ or T_θ (or Goodall) and NN with the proposed measure. The significance probability for $H_1(P_{\text{value}})$ is $\text{Norm}(Z < z)$ where $\text{Norm}(\cdot)$ is the standard normal distribution.

4.4. Experimental results and discussion

Experiment results are presented in Table 2, including:

- Data set information: name of the data set (name), number of objects (n), number of attributes (m), number of attribute values (m_v).
- Results of the first experiment: dependency factors $\rho(D)$.
- Results of the second experiment: average accuracy μ_1 of NN with our method ($\phi(.,.)$), average accuracy (μ_0) of NN with a measure of T_θ or S_θ (Goodall), significant probability (P_{value}) that indicates how accurate of NN with our method is in comparison to the accuracy of NN and a measure of T_θ or S_θ (Goodall).

As can be seen from Table 2, for almost all data sets, attributes are strongly dependent on each other. In particular, there are 14 data sets whose dependency factors are greater than 90%, and only one data set whose dependency factor is less than 50%. This proves the experimentally applicability of the proposed measure to real data.

A more important observation is that in 27 and 24 out of 30 cases, the combination of NN and the proposed method achieves a higher accuracy than the combination of NN and a measure of T_θ or S_θ and Goodall. In addition, NN with our method is significantly more accurate than NN with a measure of T_θ or S_θ and Goodall at 27 and 21 out of 30 cases, respectively (P-values are greater than 95%).

Table 2
Experiment results

Data sets						$\phi(.,.)$	T_θ or S_θ		Goodall	
No.	Name	n	m	m_v	$\rho(D)$	μ_1	μ_0	P_{value}	μ_0	P_{value}
1	Splice	3190	61	296	100	85	66	100	47	100
2	ttt	958	10	27	94	97	69	100	90	100
3	Zoo	101	18	136	92	98	88	100	92	100
4	Wine	178	14	37	97	99	92	100	91	100
5	Waveform	5000	22	106	82	77	70	100	65	100
6	Bridges	106	13	199	100	71	61	100	60	100
7	crx	690	16	60	98	83	77	100	78	100
8	Lymphography	148	19	59	92	85	76	100	76	100
9	Promoters	106	58	228	86	82	73	100	54	100
10	Anneal	898	39	100	58	99	98	100	96	100
11	Flare	1066	13	42	80	70	65	100	67	100
12	Hypo	3163	26	61	92	99	98	100	98	100
13	Krvskp	3196	37	73	93	88	80	100	87	100
14	Post-operative	90	9	24	82	61	52	100	55	100
15	Sick	2800	30	66	78	97	95	100	96	100
16	Hepatitis	155	20	51	81	85	80	100	81	99
17	Cleve	303	14	31	71	79	77	99	77	99
18	Pima	768	9	17	57	74	71	100	73	96
19	Breast	699	11	31	80	96	93	100	96	96
20	Glass	214	10	22	61	64	60	98	61	96
21	Diabetes	768	9	17	57	67	65	100	66	94
22	Mushroom	8124	23	117	91	96	92	100	95	91
23	Iris	151	5	12	100	92	88	100	91	67
24	Vehicle	846	19	71	99	66	63	100	66	64
25	Primary-tumor	339	18	42	92	33	31	98	33	50
26	Heart	270	14	22	54	76	74	97	76	26
27	AllbpCleand	2800	30	70	77	96	97	20	97	2
28	Vote	435	17	48	100	94	90	100	95	1
29	German	1000	21	63	59	68	69	4	70	0
30	Monks	432	7	17	0	50	56	0	36	100

Moreover, Table 2 shows that for data sets with high dependency between attributes (e.g., data sets ttt, spice), NN with the proposed measure are much more accurate than NN with a measure of T_θ or S_θ and Goodall (e.g., ttt: 97% versus 69% and 90%, splice: 85% versus 66% and 47%). However, for some data sets with low dependency factors (e.g., monks, german), the combination of NN and proposed method is slightly worse than the combination of a measure of T_θ or S_θ (e.g., monks: 50% versus 56% and 36%, german: 68% versus 69% and 70%).

5. Conclusions

We introduced a novel dissimilarity measure for categorical data. Our method measures the dissimilarity between two values of an attribute based on relations between the attribute and the other attributes. Experiments with real data show that categorical attributes are often correlated to each other, thus, the proposed measure is appropriate for many real applications. Moreover, applying the proposed measure to real data significantly boosts the accuracy of NN in comparison to combining it with other popular methods, e.g., Jaccard, Dice, Russell and Goodall. However, there are some data sets whose attributes are highly independent, and for these, other measures are slightly better than our method. Note, the proposed method is not suitable for data sets whose attributes are significantly independent. Therefore, one should test the independence of attributes before deciding the suitable method. If the attributes are highly correlated, the proposed method is recommended. Otherwise, one should chose other measures such as measures of families S_θ and T_θ. To release the assumption about the dependence of the attributes, we suggest considering the dissimilarity between two values not only based on dissimilarities between conditional probability distributions of other attributes when given these values but also dissimilarities between the conditional probability distributions and the distributions of attributes. However, the idea is beyond the scope of this paper.

Acknowledgments

We appreciate Le Sy Vinh at the Heinrich–Heine University of Duesseldorf, Germany, and Judith Steeh at Japan Advanced Institute of Science and Technology, Japan for helpful comments on the manuscript.

References

Albert, M., 1983. Measures of Association Quantitative Applications in the Social Sciences, vol. 32. Sage, Newbury Park, CA.

Batagelj, V., Bren, M., 1995. Comparing resemblance measures. J. Classif. 12 (1), 73.

Baulieu, F., 1989. Classification of presence/absence based dissimilarity coefficients. J. Classif. (6), 233–246.

Blake, C., Merz, C., 1998. (uci) repository of machine learning databases. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13, 21–27.

de Carvalho, F., 1994. Proximity coefficients between boolean symbolic objects. In: Diday, E. et al. (Eds.), New Approaches in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organisation, vol. 5. Springer-Verlag, Berlin, pp. 387–394.

de Carvalho, F.A.T., 1998. Extension based proximities between constrained boolean symbolic objects. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y. (Eds.), Data Science, Classification and Related Methods. Springer-Verlag, Tokyo, Japan, pp. 370–378.

Dice, L.R., 1945. Measures of the amount of ecological association between species. Ecology 26, 297–302.

Fisher, R., 1950. Statistical Methods for Research Workers, eleventh ed. Oliver and Boyd, Hafner Publishing Company, New York.

Goodall, D., 1966. A new similarity index based on probability. Biometrics 22, 882–907.

Gowda, K., Diday, E., 1991a. Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24 (6), 567–578.

Gowda, K., Diday, E., 1991b. Unsupervised learning thought symbolic clustering. Pattern Recognition Lett. 12, 259–264.

Gowda, K., Diday, E., 1992. Symbolic clustering using a new similarity measure. IEEE Trans. Systems Man Cybernet. 22 (2), 368–378.

Gower, J., 1971. Coefficients of association and similarity, based on binary (presence–absence) data: an evaluation. Biometrics 27, 857–871.

Gower, J., Legendre, P., 1986. Metic and euclidean properties of dissimilarity coefficients. J. Classif. 3, 5–48.

Hubálek, Z., 1982. Coefficients of association and similarity, based on binary (present–absence) data: an evaluation. Biol. Rev. (57), 669–689.

Ichino, M., Yaguchi, H., 1994. Generalized minkowski metrics for mixed feature-type data analysis. IEEE Trans. Systems Man Cybernet. 24 (4), 698–708.

Jaccard, P., 1912. The distribution of the flora of the alpine zone. New Phytol. 11, 37–50.

Krantz, D., Luce, R., Suppes, P., Tversky, A., 1971. Foundations of Measurement, vol. I. Academic Press, New York.

Kullback, S., 1959. Information Theory and Statistics. John Wiley and Sons, New York.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.

Lancaster, H., 1949. The combining of probabilities arising from data in discrete distributions. Biometrika 36, 370–382.

Liebetrau, A., 1983. Measures of Association. Sage, Newbury Park, CA.

Lin, J., 1991. Divergence measures based on the shannon entropy. IEEE Trans. Inform. Theory 37 (1), 145–151.

Liu, B., Hsu, W., Ma, Y., 1998. Integrating classification and association rule mining. Knowl. Discov. Data Min., 80–86.

Rached, Z., Alajaji, F., Campbell, L., 2001. Rényis divergence and entropy rates for finite alphabet markov sources. IEEE Trans. Inform. Theory 47 (4), 1553–1561.

Russell, P.F., Rao, T.R., 1940. On habitat and association of species of anopheline larvae in southeastern, Madras. J. Malaria Inst. India 3, 153–178.

Sokal, R.R., Sneath, P.H.A., 1963. Principles of Numerical Taxonomy. W.H. Freeman, San Francisco.