Available online at www.sciencedirect.com
SCIENCE DIRECT®
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 $S_\theta$ and $T_\theta$ 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: $\mathrm{Shape} = \{\square, \diamond, \triangle\}$ and $\mathrm{Color} = \{R, G, B\}$. 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,