Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 110–118
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set

Amir Ahmad [a,*], Lipika Dey [b]

[a] MEMS Group, Solid State Physics Lab, Timarpur, Delhi 54, India
[b] Mathematics Department, Indian Institute of Technology, Hauz Khas, Delhi 16, India

Received 16 February 2005; received in revised form 10 May 2006
Available online 9 August 2006

Communicated by F. Roli

Abstract

Computation of similarity between categorical data objects in unsupervised learning is an important data mining problem. We propose a method to compute distance between two attribute values of same attribute for unsupervised learning. This approach is based on the fact that similarity of two attribute values is dependent on their relationship with other attributes. Computational cost of this method is linear with respect to number of data objects in data set. To see the effectiveness of our proposed distance measure, we use proposed distance measure with *K*-mode clustering algorithm to cluster various categorical data sets. Significant improvement in clustering accuracy is observed as compared to clustering results obtained using traditional *K*-mode clustering algorithm.

Keywords: Categorical data; Similarity; Unsupervised learning; Co-occurrences

1. Introduction

Computing similarity between two data objects is very important for classification and clustering. If vector defining data objects have numeric values, the dissimilarity between two data objects can be computed using Euclidean distance, Minkowski Matrix, Manhattan distance, Mahalanobis distance, etc. (Esposito et al., 2002). But the problem to measure similarity becomes quite difficult when we are treating data objects with categorical values. A qualitative variable Y is called categorical if its range γ bears no internal structure at all (Bock, 2002). Thus, for two categories $x, y \in \gamma$ we can only distinguish between two alternatives

$$x = y \quad \text{and} \quad x \neq y$$

If we want to specify a distance $\delta(x, y)$ between two categorical values $x, y \in \gamma$ in supervised learning, Value Distance Matrix (Stanfill and Waltz, 1986) and Modified Value Distance Matrix (Cost and Salzberg, 1993) have been defined. But in traditional methods distance $\delta(x, y)$ is defined as 0 or 1 (if $x = y$ or $x \neq y$) for unsupervised learning this scheme is called Hamming distance measure (Esposito et al., 2002). Using this distance measure various similarity measures e.g., Jaccard (*S*-coefficient) similarity measure, Sokal–Michener (*M*-coefficient) similarity measure, Grower–Legendre similarity measure, etc. (Esposito et al., 2002) have been suggested to get the similarity or dissimilarity coefficient between two categorical data objects.

Recently there has been considerable interest in defining intuitive and easily computable distance measures for categorical data objects (Gowda and Diday, 1991; Agrawal et al., 1993; Ichino and Yaguchi, 1994; Jagdish et al., 1995; Knobbe and Adiaans, 1996; Ganti et al., 1999) in unsupervised learning. These have found applications in clustering (Ganti et al., 1999; Guha et al., 1999; Bellot

* Corresponding author.
E-mail addresses: amirahmad01@yahoo.co.in (A. Ahmad), lipikadey@hotmail.com (L. Dey).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.06.006

and El-Beze, 1999), document retrieval system (Karpov and Edelman, 1998) and extraction of information in hyperlinked environments (Gibson et al., 1998).

In this paper we are concentrating our work on single valued variable data. In other words if $(X_1, X_2, \ldots, X_m)$ define a data object having m attributes then every attribute value X_i $(i = 1, \ldots, m)$ can take only one value. Now we compare how two well known distance function, Gowda and Diday's dissimilarity measure (Gowda and Diday, 1991), and Ichino and Yaguchi' dissimilarity measure (Ichino and Yaguchi, 1994) work for categorical data.

For symbolic data (Bock, 2002) (multi-valued variable) and categorical attribute, Gowda and Diday's dissimilarity measure (Gowda and Diday, 1991) is

$$\delta(A_j, B_j) = D_s(A_j, B_j) + D_c(A_j, B_j)$$

where

A_j and B_j two attribute values of jth attribute
$\delta(A_j, B_j)$ distance between A_j and B_j
l_a number of categories in $A_j = |A_j|$
l_b number of categories in $B_j = |B_j|$
inters number of categories in $A_j \cap B_j = |A_j \cap B_j|$
l_s number of categories in $A_j \cup B_j = l_a + l_b - \text{inters}$

$$D_s(A_j, B_j) = |l_a - l_b|/l_s$$
$$D_c(A_j, B_j) = (l_a + l_b - 2 * \text{inters})/l_s$$

For single valued variable $l_a = l_b = 1$ so $D_s(A_j, B_j) = |l_a - l_b|/l_s$ will always be zero.

Value of inters will be one if $A_j = B_j$ else it will be zero. So for $A_j = B_j$,

$$\delta(A_j, B_j) = D_s(A_j, B_j) + D_c(A_j, B_j) = 0 + (1 + 1 - 2 * 1)/1 = 0$$

for $A_j \neq B_j$

$$\delta(A_j, B_j) = D_s(A_j, B_j) + D_c(A_j, B_j) = 0 + (1 + 1 - 2 * 0)/1 = 1$$

That shows that this dissimilarity measure compute the distance measure between two categorical single valued variable as Hamming distance.

Ichina and Yaguchi's dissimilarity measure (Ichino and Yaguchi, 1994) for symbolic data (multi-valued variable) and categorical attribute is

$$\delta(A_j, B_j) = |A_j \cup B_j| - |A_j \cap B_j| + \gamma(2 * (A_j \cap B_j) - |A_j| - |B_j|)$$

where $0.5 \geqslant \gamma \geqslant 0$ is aprepesified parameter.

For single valued variable $|A_j| = 1$, $|B_j| = 1$, if $A_j = B_j$, $|A_j \cup B_j| = 1$, $|A_j \cap B_j| = 1$

$$\delta(A_j, B_j) = 0$$

For $A_j \neq B_j$, $|A_j \cup B_j| = 2$, $|A_j \cap B_j| = 0$

$$\delta(A_j, B_j) = 2(1 - \gamma)$$

This shows that distance between two categorical values is some constant irrespective of categorical values which is not a very good distance measure as we know from our experience that some categorical values are much similar as compared to other categorical values for example if color is one attribute of dataset then there is strong possibility that distance between red and orange colors is less than the distance between black and white colors.

In this paper we propose a method to compute the distance $\delta(x, y)$ between two categorical values (of same attribute) in unsupervised learning. In Section 2, we describe our proposed algorithm. In Section 3, it is experimented with real world categorical data sets and its results and comparison with other standard algorithm is presented. Section 4 summarizes our contributions and describes directions for future work.

2. Proposed algorithm

Most of the distance measures take distance between any two categorical attribute values equal. We propose an algorithm, which instead of taking distance between any two categorical attribute values equal, compute the distance by observing effect of attribute values on dataset.

Ganti et al. (1999) described notion of similarity for pure categorical data set (all attributes of data are categorical) to attribute pairs on the same attribute. Let $a_1, a_2 \in A_i$ (ith attribute) and $x \in A_j$ (jth attribute). If (a_1, x) and (a_2, x) are strongly connected then (a_1, a_2) are "similar" to each other with respect to A_j.

Das and Mannila (2000) proposed ICD (Iterated Contextual Distances) algorithm (for pure categorical data set) in which they suggested that various different similarity notions (attribute similarity, data objects similarity) are inter-dependent. Association rules (Agrawal and Srikant, 1994; Agarwal et al., 1993) derive patterns from grouped data attributes that co-occur with high frequency. Much of the emphasis of computing similarity for categorical data set in unsupervised learning is based on frequently co-occurring items (Ganti et al., 1999; Bellot and El-Beze, 1999; Gibson et al., 1998; Han et al., 1997).

These methods suggest that we can compute distance between two categorical values (of same attribute) with respect to other attributes.

Cost and Salzberg (1993) presented Modified Value Distance Matrix (MVDM) which compute distance $\delta(x, y)$ between two categorical values with respect to class column (supervised learning). According to Modified Value Distance Matrix (MVDM)

$$\delta(x, y) = \sum_{c=1}^{K} |N_{i,x,c}/N_{i,x} - N_{i,y,c}/N_{i,y}|$$

$$\delta(x, y) = \sum_{c=1}^{K} |p_i^x(c) - p_i^y(c)|$$

k the number of classes in dataset D
$N_{i,x,c}$ the number of data objects in D that have the value x for the ith attribute and the data object belong to the cth class