2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A New Distance Metric for Unsupervised Learning of Categorical Data

Hong Jia[a] and Yiu-ming Cheung[a,b]
[a]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[b]The United International College, BNU-HKBU, Zhuhai, China
Email: {hjia, ymc}@comp.hkbu.edu.hk

***Abstract*—Distance metric is the basis of many learning algorithms and its effectiveness usually has significant influence on the learning results. Generally, measuring distance for numerical data is a tractable task, but for categorical data sets, it could be a nontrivial problem. This paper therefore presents a new distance metric for categorical data based on the characteristics of categorical values. Specifically, the distance between two values from one attribute measured by this metric is determined by both of the frequency probabilities of these two values and the values of other attributes which have high interdependency with the calculated one. Promising experimental results on different real data sets have shown the effectiveness of proposed distance metric.**

I. Introduction

Measuring the distance between two data objects plays an important role in many data mining and machine learning tasks, such as clustering, classification, feature selection, outlier detection, and so on. Generally, distance computation is an embedded step for these learning algorithms and different metrics can be conveniently utilized. However, the effectiveness of adopted distance metric usually has significant influence on the performance of the whole learning method [4], [5]. Therefore, it becomes a key research issue to present more appropriate distance metrics for the various learning tasks.

For purely numerical data sets, the distance computation is a tractable problem as any numerical operation can be directly applied. In the literature, a number of distance metrics and metric learning methods have been proposed for numerical data. The most widely used metrics in practice should be the Manhattan distance, Euclidean distance, and Mahalanobis distance [1]. By contrast, measuring distance for categorical data is a more challenging problem as there is no explicit ordering information in categorical values and the only numerical operation that can be straightforwardly applied is the identical comparison operation [2]. Under the circumstances, a simplest way to overcome this problem is to transform the categorical values into numerical ones, e.g. the binary strings [3], [6], [7], and then the existing numerical-value based distance metrics can be utilized. Nevertheless, such a kind of method has ignored the information embedded in the categorical values and cannot faithfully reveal the relationship structure of the data sets [8], [9]. Therefore, it is desirable to solve this problem by proposing new distance metric for categorical data based on the characteristics of categorical values.

Among the existing work, the most straightforward and widely used distance metric for categorical data is the Hamming distance [1], in which the distance between different categorical values is set at 1 while a distance of 0 is assigned to identical values. Then, for a pair of categorical data objects with multiple attributes, the Hamming distance between them will be equal to the number of attributes in which they mismatch. Although the Hamming distance is easy to understand and convenient for computation, the main drawback of this metric is that all attribute values have been considered equally and the statistical properties of different values have not been distinguished [10]. For this reason, more researchers attempt to measure the distance for categorical data with the distribution characteristics of categorical values. For example, Cost and Salzberg [11] had proposed a distance metric namely Modified Value Distance Matrix (MVDM) for supervised learning task, in which the distance between two categorical values is calculated with respect to the class label of the data set. Additionally, for unsupervised distance measure of categorical data, Le and Ho [2] presented an indirect method which defines the distance between two values from one attribute as the sum of the Kullback-Leibler Divergence between conditional probability distributions of other attributes given these two values. Similar idea has also been adopted by [12], in which the distance of two values from one attribute is quantified with respect to the co-occurrence probabilities of the values from all the other attributes with these two values.

Besides of the aforementioned methods which directly propose special distance metric for categorical data sets, some similarity measures [13], [14], [15], [16], [17], [18], [19] presented for categorical or mixed data can also be utilized to quantify the relationship between different categorical data objects. For example, the Goodall similarity metric proposed in [13] assigns a greater weight to the matching of uncommon attribute values than common values in similarity computation without assuming the underlying distributions of categorical values. Subsequently, the similarities between pairs of values are integrated with Lancaster's method [20] to estimate the similarity between data objects. Moreover, Gowda and Diday [14], [15], [16] have proposed an algebraic method to measure the similarity between categorical data. In this method, the similarity between two attribute values is defined based on three components: position, span, and content. Here, the component "position" works only when the attribute type is quantitative, the "span" indicates the relative sizes of the attribute values without referring to common parts between them, and the "content" is to measure the common parts between attribute values. Finally, the summation of these three similarity components is the estimate of the similarity between given attribute values.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE