Neurocomputing 306 (2018) 160–170

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning category distance metric for data clustering

Baoguo Chen [a,*], Haitao Yin [b]

[a] Research Center for Science Technology and Society, Fuzhou University of International Studies and Trade, Fujian 350202, China
[b] Research Institute of Science Technology and Society, Fuzhou University, Fujian 350116, China

ARTICLE INFO

Article history:
Received 12 November 2016
Revised 17 December 2017
Accepted 16 March 2018
Available online 4 May 2018

Communicated by Prof. Zhou Xiuzhuang

Keywords:
Data clustering
Categorical attribute
Distance metric
Distance learning
Category weight

ABSTRACT

Unsupervised learning of adaptive distance metrics for categorical data is currently a challenge due to the difficulties in defining an inherently meaningful measure parameterizing the heterogeneity within matched or mismatched categorical symbols. In this paper, a new distance metric called category distance and a non-center-based algorithm are proposed for categorical data clustering. The new metric is formulated based on the category weights for each categorical attribute, no more depending on the common assumption that all categories on the same attribute are independent of each other. The problem of learning the category distance is therefore transformed into the new problem of learning a set of category weights, which can be jointly optimized with the clusters optimization. A case study on DNA sequences and experimental results on ten real-world data sets from different domains are given to demonstrate the performance of the proposed methods with comparisons to the existing distance measures for categorical data.

1. Introduction

Computing the dissimilarity (or similarity) between data objects is one of the key intermediate operations in many machine learning tasks, such as data clustering aimed at partitioning a set of objects into homogeneous groups based on some distance functions. Learning of an adaptive distance metric for data clustering has sparked wide interest because of inherent data dependency of the semantic dissimilarity between data objects [1,2]. A number of unsupervised metric learning methods have been proposed, including kernel linear transformation [3,4], relevant component analysis [5], automated attribute-weighting [6] and many others integrating feature extraction methods [1,7,8].

These methods have been applied to many distance-based clustering tasks and gained great popularity. However, they mainly focus on learning distance metric from numeric data, where the metric can be parameterized in a well-defined measure, for instance, the Mahalanobis distance function [9,10]. For categorical data, distance computation is not straightforward. The problem becomes difficult due to the fact that in categorical case the data can only take discrete values (categorical symbols or categories) and statistical measures such as mean, variance as well as covariance, which are common in numeric data, are undefined for categorical data [11,12]. Consequently, the learning methods that have been successfully used for numeric data, including the popular maximum-margin-based approaches [13], sparse representation and manifold learning [14], Laplacian regularized metric learning [15] and the continuous-kernel methods [16], cannot be directly applied to categorical data.

Learning a distance metric on categorical data is a fundamental problem as such data have become ubiquitous in machine learning applications [11,12,17,18]. A few intuitive metrics have been defined, such as the common overlap measure, occurrence frequency [19], information-theoretic metric [20], etc (see [21] for a survey or Section 2.2 for the typical measures). For example, the overlap measure (alternatively known as the *simple-matching* coefficient [12]) computes the similarity from the number of categories that appear in both objects, and have been widely used in the categorical data clustering algorithms including the popular *K*-modes and its numerous variants [11,22].

We remark that these metrics are not valid in many real applications because, essentially, they are defined based on the assumption that all the categories in the data are *independent* of each other [23]: samples having different categories are independent of each other while they are perfectly correlated as long as the same category is taken. The assumption is generally not true in reality. For instance, in the international trade catalogue, the categories “blazers” and “jackets” are quite heterogeneous since both make up one of the components of a suit. As an other example, take an attribute representing customers’ age: two customers may share the

* Corresponding author.
E-mail address: chenbaoguo@fzfu.edu.cn (B. Chen).

https://doi.org/10.1016/j.neucom.2018.03.048
0925-2312/

same category "Middle-aged", but their real ages can actually be different. Many examples like these pose a unique challenge to the distance definition, because there is currently no method for adaptively learning the category dissimilarity for a polytomous attribute in clustering categorical data [23,24].

In this paper, we propose to solve these problems by learning a category distance metric for categorical data clustering. The metric assigns an individual distance value to each pair of categories on the same attribute, either matched or mismatched, to distinguish their heterogeneity such that the independence assumption is relieved. The category distance metric is then parameterized by a set of category weights, allowing the learning problem to be transformed into the new problem of learning the optimized weights. We also define a new clustering algorithm based on the category distance metric, to perform non-center-based clustering on categorical data with the distance metric jointly learned from the data. A series of experiments on UCI categorical data sets are conducted to evaluate the performance of the distance metric and the clustering algorithm.

The remainder of this paper is organized as follows: Section 2 presents some preliminaries and related work. Section 3 describes our category distance metric. In Section 4, the distance learning method and the new clustering algorithm are presented. Experimental results are presented in Section 5. Section 6 gives our conclusions.

2. Preliminaries and related work

In this section, notation and definitions related to categorical data clustering are introduced, followed by a sampling of related work on the distance measures for categorical data.

2.1. Preliminaries

In the following pages, the sample set to be grouped into K clusters is denoted by X, which consists of $N = |X|$ data objects, each being a D-dimensional vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_D\rangle$ or $\mathbf{y} = \langle y_1, y_2, \ldots, y_D\rangle$. We call $\mathbf{x}$ a categorical data object if each attribute x_d for $d = 1, 2, \ldots, D$ is a categorical attribute, as defined in the following Definition 1.

Definition 1 (Categorical attribute)**.** An attribute is of categorical type if it takes values from the finite symbols (categories) set $S = \{s_1, s_2, \ldots, s_m\}$, where $m = |S|$ is the number of symbols.

Such categorical data have become ubiquitous in machine learning applications. In bioinformatics, for instance, the nucleotides in each position of DNA sequences can be viewed as a categorical attribute, where the category set is typically $S = \{'A', 'G', 'T', 'C'\}$. Clearly, the set *mean* is a undefined concept for such a categorical data set. As a consequence, the popular K-means type algorithms, which make use of the set mean to represent the cluster center, cannot be directly used for categorical data clustering. The K-modes algorithm and its variants [11,25] then resort to the mode categories on each attribute to represent the "center" for categorical clusters. However, such mode-based approaches can only capture partial information on the data objects in a cluster. To define an efficient clustering algorithm without the formulation for cluster centers, partition-based methods have been suggested [12], as shown in the following Definition 2.

Definition 2 (Partition-based categorical data clustering)**.** Partition-based clustering of the categorical data set X is the optimized partitioning $\Pi = \{\pi_k | k = 1, 2, \ldots, K\}$ that minimizes

$$J_0(\Pi) = \sum_{k=1}^{K} \frac{1}{|\pi_k|} \sum_{\mathbf{x}\in\pi_k} \sum_{\mathbf{y}\in\pi_k} Dis(\mathbf{x}, \mathbf{y})$$
$$\text{s.t.} \quad X = \bigcup_{k=1}^{K} \pi_k \quad \text{and} \quad \forall k : \pi_k \neq \varnothing, \tag{1}$$

where $Dis(\cdot, \cdot)$ measures the pairwise dissimilarity of categorical objects, and π_k denotes the kth cluster of X with $|\pi_k|$ being the number of objects in π_k.

Unlike the numeric case, where the pairwise dissimilarity can be measured using the common distance functions such as the Euclidean distance, here, $Dis(\cdot, \cdot)$ should be computed as that aggregation of the symbolic distance on each categorical attribute. Formally,

$$Dis(\mathbf{x}, \mathbf{y}) = \sum\nolimits_{d=1}^{D} [\psi(x_d, y_d)]^2 \tag{2}$$

where $\psi(\cdot, \cdot)$ is the distance metric measuring the symbolic dissimilarity between two categories. Based on the definitions, the problem of learning $\psi(\cdot, \cdot)$ can be alternatively represented as learning the real matrix Ψ for each categorical attribute of X, defined by

$$\Psi = \begin{bmatrix} \psi(s_1, s_1) & \psi(s_1, s_2) & \ldots & \psi(s_1, s_m) \\ \psi(s_2, s_1) & \psi(s_2, s_2) & \ldots & \psi(s_2, s_m) \\ \ldots & \ldots & \ldots & \ldots \\ \psi(s_m, s_1) & \psi(s_m, s_2) & \ldots & \psi(s_m, s_m) \end{bmatrix}$$

satisfying, for any three categories c, c', c'' on that attribute (therefore, c, c' and $c'' \in S$),

Condition (1): $\psi(c, c') \geq 0$ (non-negativity),

Condition (2): $\psi(c, c') = \psi(c', c)$ (symmetry), and

Condition (3): $\psi(c, c') \leq \psi(c, c'') + \psi(c'', c')$ (triangular inequality).

2.2. A sampling of related work

The similarity or distance measures defined for categorical data in the literature can be roughly divided into three groups, named Type I, Type II and Type III measures, respectively. The measures will all be in the context of dissimilarity (distance) for discussion, with the similarity converted by $1 - \psi(\cdot, \cdot)$.

In the Type I measures, the diagonal elements of Ψ are fixed at 0, while $\psi(c, c') = w$ for $c \neq c'$, where w is the attribute weight, which is a positive constant irrelevant to c or c'. Such a measure, called Overlap Measure (OM) [21], is defined based on the simple-matching method, given by

$$\psi_{OM}(c, c') = w \times \begin{cases} 0 & c = c' \\ 1 & c \neq c' \end{cases} \tag{3}$$

In the case where $w = 1$, it degenerates to the common simple-matching distance as mentioned before. Though simple, the measure has been used in the categorical data clustering algorithms, such as the well-known K-modes algorithm [25]. An effective extension to the simple-matching distance is to define a weighted measure, by learning the attribute weight $w \in [0, 1]$. For example, in [26], w is computed as being inversely proportional to the kernel bandwidth of the categorical attribute, while in [11] it is calculated based on the complement-entropy of the category distribution.

It can be seen that, in the Type I measures, distance is defined as zero between two samples sharing a common category without considering the heterogeneity of the categorical attribute. This is because such measures are generally based on the *independence assumption*, as described in Section 1. The measures of Type II fix this