Computational Statistics and Data Analysis 54 (2010) 1247–1265

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Conversion of categorical variables into numerical variables via Bayesian network classifiers for binary classifications

Namgil Lee [a], Jong-Min Kim [b,*]

[a] Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Republic of Korea

[b] Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN 56267, USA

ARTICLE INFO

Article history:
Received 10 September 2008
Received in revised form 4 November 2009
Accepted 5 November 2009
Available online 18 November 2009

ABSTRACT

Many pattern classification algorithms such as Support Vector Machines (SVMs), Multi-Layer Perceptrons (MLPs), and K-Nearest Neighbors (KNNs) require data to consist of purely numerical variables. However many real world data consist of both categorical and numerical variables. In this paper we suggest an effective method of converting the mixed data of categorical and numerical variables into data of purely numerical variables for binary classifications. Since the suggested method is based on the theory of learning Bayesian Network Classifiers (BNCs), it is computationally efficient and robust to noises and data losses. Also the suggested method is expected to extract sufficient information for estimating a minimum-error-rate (MER) classifier. Simulations on artificial data sets and real world data sets are conducted to demonstrate the competitiveness of the suggested method when the number of values in each categorical variable is large and BNCs accurately model the data.

1. Introduction

The primary goal of pattern classification is to estimate a classification function, i.e., a classifier, using labeled training patterns so that the estimated classifier will correctly assign class labels to novel test patterns. Some examples of the most widely used classification algorithms are Support Vector Machines (SVMs), Multi-Layer Perceptrons (MLPs), and K-Nearest Neighbors (KNNs). SVMs (Vapnik, 1995; Burges, 1998; Cristianini and Shawe-Taylor, 2000) build a sparsely formulated hyperplane classifier through the maximization of a margin criterion. MLPs (Bishop, 1995; Haykin, 1999) construct networks with two or more layers of nonlinear computation units, and the synaptic weights of each unit are estimated by a maximum likelihood estimation. KNNs (Duda et al., 2001) make a rule which classify a pattern by assigning it the label most common among its k nearest samples.

Many classification algorithms including the above examples assume that a pattern is represented as a vector of numerical values. For example, the common basic operations of those algorithms are the computations of dot products and Euclidean distances between patterns and other vectors. However, in many real world data sets a pattern is represented as a collection of discrete or structured objects. For example, a text is represented as a string of letters, gene as a sequence of nucleotides, image as a set of pixels, and so on.

In this paper we concentrate on the case that a pattern is represented as a collection of only two types of values: categorical values and numerical values. That is, each of the variables in a pattern is of either categorical type or numerical type, and a categorical variable takes its values in some finite set of categories. In this case the classification algorithms

* Corresponding author.
E-mail addresses: namgil@kaist.ac.kr (N. Lee), jongmink@morris.umn.edu (J.-M. Kim).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.11.003

such as SVMs, MLPs, and KNNs are not directly applicable, and one might have to either discard the categorical values or convert the categorical values into numerical values. One typical conversion method is to use a single number to represent a categorical value. But this method depends on an arbitrary ordering of values in a categorical variable. Alternatively, Hsu et al. (2003) suggest to use m binary numbers to represent a m-category variable. Hsu et al. (2003) remark that if there are not too many values in a categorical variable, the method is more stable than using a single number to represent a categorical variable.

On the other hand, there have been many researches on designing kernel functions for various structured data (Gärtner, 2003; Shawe-Taylor and Cristianini, 2004). A kernel function is a measure of meaningful similarities between a pair of patterns (Schölkopf and Smola, 2002, Ch.2), and appropriately selected kernel functions have led to improvements in classification performances (Vapnik, 1995; Joachims, 1998; Chapelle et al., 1999; Pavlidis et al., 2002). The Fisher kernel (Jaakkola and Haussler, 1999) and the marginalized kernel (Tsuda et al., 2002) are typical kernels defined from probabilistic models such as Hidden Markov Models (HMMs), and both of them have achieved remarkable improvements in biological sequence classifications. This implies that defining a kernel on a probabilistic model is a useful way of incorporating prior knowledge and manipulating structured data.

In this paper we propose a new method of converting mixed data of categorical and numerical values into data of numerical values by defining a kernel function from a probabilistic model. First we define an ideal kernel function for binary classification problems based on the definition of a minimum-error-rate (MER) classifier. Second we propose to use Bayesian Network Classifiers (BNCs) (Friedman et al., 1997) to accurately estimate the ideal kernel function. The estimation using BNCs allows an effective modeling of the categorical variables, and it is computationally efficient and robust to noises and data losses. Third we show that the ideal kernel function is decomposed into products of simpler kernel functions. This decomposition enables us to explicitly present the conversion of original mixed data into numerical data. Since the suggested method uses a small number of real numbers to represent a categorical value, there is not much increases in dimensions of patterns regardless of the number of values in a categorical variable. Moreover a simple linear classifier can approximate the MER classifier using the converted numerical values as far as the estimation by BNCs is accurate.

This paper is organized as follows. In Section 2 we describe the mixed data of categorical and numerical variables, and we introduce basic properties of a kernel function. In Section 3 we define the ideal kernel and the MER classifier. In Section 4 the estimation of probabilities for the mixed data using BNCs is explained. In Section 5 the decomposition of the ideal kernel is described, and the explicit conversion of the mixed data into numerical data is proposed. In Section 6 we present simulation results on artificial data sets and real world data sets comparing the suggested method with the other typical methods. We discuss about the results and future researches in Section 7.

2. Backgrounds

2.1. Mixed data of categorical and numerical variables

In this paper we suppose that an input pattern is a collection of categorical and numerical values. We denote the jth categorical variable of the ith input pattern by $x_j^{(i)}$, and the jth numerical variable of the ith input pattern by $r_j^{(i)}$. Then the ith input pattern $x^{(i)}$ is represented as a vector by

$$x^{(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)}, r_1^{(i)}, \ldots, r_m^{(i)}\right), \quad i = 1, \ldots, N. \tag{1}$$

Each categorical variable $x_j^{(i)}$ takes its value from a finite set of discrete categories, which is denoted by

$$x_j^{(i)} \in V_j = \left\{v_{j,1}, \ldots, v_{j,n_j}\right\} \tag{2}$$

where $v_{j,q}$, $q = 1, \ldots, n_j$, represents the qth categorical value that the jth categorical variable can take.

Categorical variables are abundant in real world data. For example, variables that take values from {true, false} or {yes, no} are called binary attributes, and variables that take values from any collection of words which describe characteristics of an instance are called nominal attributes. For example, {Bisexual, Homosexual, Heterosexual} is a set of words characterizing sexual orientation of a person, and {United States, Cambodia, England, …} is a set of words identifying the country of residence of a person.

However, when we want to apply the classification algorithms such as SVMs, MLPs, and KNNs to data with categorical variables, one has to either discard the categorical variables or convert them into numerical variables. The following are the two typical methods for converting categorical variables into numerical variables:

1. Single number representation: The discrete values of a categorical variable are converted into discrete real numbers. For instance, the discrete values {Bisexual, Homosexual, Heterosexual} are converted into {1, 2, 3} in order. But this conversion depends on an arbitrary ordering of values in a categorical variable, so it can result in unreliable performances.
2. Binary number representation: The n_j discrete values of a categorical variable are encoded into binary numbers of length n_j. For instance, {Bisexual, Homosexual, Heterosexual} is converted into {001, 010, 100}. Hsu et al. (2003) make a remark that this representation is more stable than the single number representation. However, if the number of categories for