Expert Systems With Applications 51 (2016) 207–217

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Bayesian network classifiers based on Gaussian kernel density

Shuang-cheng Wang[a,b], Rui Gao[b,c,*], Li-min Wang[d]

[a] Lixin Accounting Research Institute, Shanghai Lixin University of Commerce, Shanghai 201620, China
[b] School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 201620, China
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China
[d] College of Computer Science and Technology, Jilin University, Changchun 130012, China

ARTICLE INFO

Keywords:
Bayesian network classifiers
Continuous attributes
Gaussian kernel function
Smoothing parameters
Classification accuracy

ABSTRACT

For learning a Bayesian network classifier, continuous attributes usually need to be discretized. But the discretization of continuous attributes may bring information missing, noise and less sensitivity to the changing of the attributes towards class variables. In this paper, we use the Gaussian kernel function with smoothing parameter to estimate the density of attributes. Bayesian network classifier with continuous attributes is established by the dependency extension of Naive Bayes classifiers. We also analyze the information provided to a class for each attributes as a basis for the dependency extension of Naive Bayes classifiers. Experimental studies on UCI data sets show that Bayesian network classifiers using Gaussian kernel function provide good classification accuracy comparing to other approaches when dealing with continuous attributes.

1. Introduction

Naive Bayes classifier (NBC) (Bouckaert, 2005; Duda, Hart et al., 1973; Ramoni & Sebastiani, 2001) is an important probability classifier, well known for its simplicity, high efficiency, and good classification accuracy. It is capable of directly processing continuous attributes and has been widely used in medical diagnosis, text categorization, mail filtering, information retrieval, etc. However, this classifier is based on a rather strong assumption: the attributes are conditionally independent when the class is given. This leads to poor utilization of dependency information between attributes, while dependency information also being crucial for classification. To solve this problem, a series of studies on the dependency extension of NBCs have been conducted. These research studies can be traced back to the dependence tree of Chow and Liu (1968), based on which Friedman, Geiger, and Goldszmidt (1997) later constructed the well-known TAN (Tree Augmented Naive Bayesian) classifier. Grossman and Domingos (2004) conducted the learning of Bayesian network classifiers setting the conditional likelihood as the criteria. Jing, Pavlović, and Rehg (2008) set classification accuracy as the criteria, and conducted attribute selection and parameter ensemble on the TAN classifiers. Dependency extension of these classifiers can effectively improve classification accuracy, but these studies only focused on NBCs with discrete attributes. For NBCs with continuous attributes (Bouckaert, 2005), two processing methods can be applied: one is to discretize the continuous attributes, eventually turning them into classification issues with discrete attributes (Boullé, 2006; Fayyad & Irani, 1993; Friedman, Goldszmidt et al., 1996; Yang & Webb, 2009); the other does not discretize continuous attributes, but requires the estimation of the conditional density of attributes. Both of these two methods have their own advantages and disadvantages. In detail, the first method is suitable for big data sets with fewer classes where the conditional probability of attributes can therefore be reliably estimated. The second method is more likely to be used for comparatively smaller data sets with multi-class, because the estimation of conditional density can be done without many examples. This paper explores the second method, assuming that all the attributes are continuous. Nevertheless, the research findings here can also be applied to cases of mixed attributes. The core of processing continuous attributes is the estimation of conditional density, John and Langley (1995) studied NBCs and Flexible Bayes Classifiers (FBCs) which were achieved by estimating conditional density of attributes using Gaussian function and Gaussian kernel function. Although the accuracy of two kinds of classifiers is not very good, they lay the foundation of the research of Bayesian network classifiers with continuous attributes. Based on John and Langley's research, Pérez, Larrañaga, and Inza (2009) improved the estimation of Gaussian kernel function by introducing and optimizing a smoothing parameter. They used classical MISE (Mean Integrated Square Error) statistical standard to

* Corresponding author: School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai, 201620, China. Tel.: +86 13761265262.
E-mail addresses: wangsc@lixin.edu.cn (S.-c. Wang), gaorui@lixin.edu.cn, pplu_ren@163.com (R. Gao), wanglim@jlu.edu.cn (L.-m. Wang).

http://dx.doi.org/10.1016/j.eswa.2015.12.031
0957-4174/

optimize the smoothing parameter (Kobos, 2009), and named classifiers constructed with optimized smoothing parameter as Flexible Naive Bayes Classifiers (FNBCs), the classification accuracy of which is better than that of FBC. Bounhas, Mellouli, Prade, and Serrurier (2013), He, Wang, Kwong, and Wang (2014), Dong and Zhou (2014) and Pavani, Delgado-Gomez, and Frangi (2014) respectively studied Bayes classifiers based on Gaussian function and Gaussian kernel function to estimate attribute joint density. These classifiers cannot effectively use conditional independence relationship between attributes, which makes them have low accuracy and reliability.

This paper uses Gaussian kernel function with smoothing parameter to estimate conditional density. On the basis of optimizing smoothing parameter using classification accuracy as the criteria(classification accuracy criteria can better measure the fitting degree between Bayesian network classifier with continuous attributes and data), the greedy selection of parent nodes is achieved by using the same criteria. The Extended Naive Bayes Classifiers (ENBCs) are thus constructed. Then, based on the Bayesian network theory, the information provided to class by attributes is analyzed from the viewpoint of dependency extension. In the end, the experiments and analysis are given by using UCI data sets (Murphy & Aha, 2014) with continuous attributes to verify the necessity of dependency extension and the effectiveness of our methods.

The main contributions of this paper are as follows:

(1) We propose a new method to construct ENBC by imposing dependency extension on NBC with continuous attributes, which extracts the conditional dependency information among attributes for classification in the way of local optimum.
(2) We combine the smoothing parameter adjustment which determine the shapes of the Gaussian density curves in the Gaussian kernel function with the structure learning which determine the decomposition and the calculation of attribute joint density to control and optimize the fitting degree between classifier and data, and then improve the generalization ability.
(3) Based on the Bayesian network theory (Cooper & Herskovits, 1992; Heckerman, Geiger, & Chickering, 1995; Langley, Iba, & Thompson, 1992; Olesen, 1993; Pearl, 1988), we present that the attributes of ENBC can provide three types of information for class. They are transitive dependency information, direct induced dependency information and indirect induced dependency information. NBC just provides the first type of information for class. Through dependency extension, other two kinds of information can be effectively utilized and thus the classification accuracy can be improved.

This paper is divided into 4 sections: Section 2 presents the structure of ENBC, methods for estimating conditional density of attributes, the analysis of the constitution of the information provided to class by attributes, and the selection methods of attribute parent nodes; Section 3 presents the experiments and analysis comparing different levels of classification accuracy, and contribution levels of different attributes to classification; Section 4 lists the research finding and further work in the field.

2. Dependency extension of NBC

The dependency extension of NBC means that the attributes can also have other parent nodes besides class. Its purpose is to effectively utilize dependency information between attributes. With $X_1, \ldots, X_n, C$ to represent continuous attributes and class, $x_1, \ldots, x_n, c$ being their values, D being a data set with N samples, the data is generated randomly by probability distribution P, x_{im} and c_m to represent the value of sample no. $m(1 \leq m \leq N)$ of $X_i(1 \leq i \leq n)$ and C in data set D, respectively. Variables in the probability model and nodes in corresponding graph model are not always differentiated.

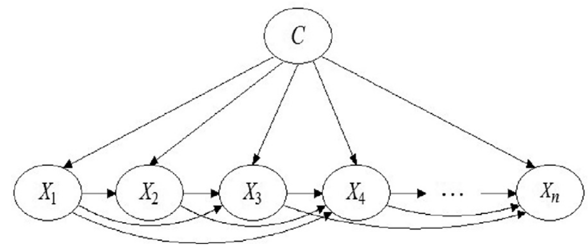

Fig. 1. The structure of ENBC.

2.1. Structure of classifiers and conditional density estimation of attributes

The classifier after dependency extension no longer possesses the star structure. Besides having class, attributes nodes can also have other parent nodes, so its structure is directed acyclic graph, with G to represent classifier structure after dependency extension, as shown in Fig. 1.

According to the Bayesian formula, Bayesian network theories (see Pearl, 1988) and conditional independency among variables in Fig. 1, we can get

$$p(c|x_1, \ldots, x_n) = \frac{p(c)\rho(x_1, \ldots, x_n|c)}{\rho(x_1, \ldots, x_n)} \propto p(c)\prod_{i=1}^{n} \rho(x_i|\pi_i, C, G) \quad (1)$$

where $p(c)$ is the prior probability of class, $\rho(x_i|\pi_i, C, G)$ is the conditional density, π_i is the configuration of parent nodes set $\prod_i$ of X_i. The dependency extension of NBC with continuous attributes requires the estimation of conditional density of multi-attributes. To ensure that the estimated conditional density fits the data well, the method of multivariate kernel function is used. With $\hat{\rho}(x_i, \ldots x_n|c, D)$ to represent the estimation of $\rho(x_i, \ldots x_n|c)$, the general form of multivariate kernel function estimation on data set D is:

$$\hat{\rho}(x_i, \ldots x_n|c, D) = \frac{1}{N(c)h_1 \ldots h_n}\sum_{m=1}^{N}\left[sign(c_m)\prod_{i=1}^{n} K_i\left(\frac{x_i - x_{im}}{h_i}\right)\right] \quad (2)$$

where $N(c)$ is the sample number of $c_m = c$ in data set D, $sign(c_m) = \begin{cases} 1 & if c_m = c \\ 0 & else \end{cases}$ ($sign(\cdot)$is an indicator function), $K_i(\cdot)$is the kernel function of X_i, $h_i(i = 1, \ldots, n)$are smoothing parameters.

Multivariate Gaussian kernel function is used to estimate the attribute conditional density emerging from the process of dependency extension. For simplicity we assume $h = h_1 = \cdots = h_n$, then $K_i(\frac{x_i - x_{im}}{h_i}) = \frac{1}{\sqrt{2\pi}h}exp[-\frac{(x_i - x_{im})^2}{2h^2}]$. We can obtain

$$\hat{\rho}(x_i, \ldots x_n|c, D) = \frac{1}{N(c)h^n}\sum_{m=1}^{N}\left[sign(c_m)\prod_{i=1}^{n} K_i\left(\frac{x_i - x_{im}}{h_i}\right)\right] \quad (3)$$

Let $\Pi_i = \{X_i^1, \ldots, X_i^{t(i)}\}$, $t(i)$ is the number of parent nodes of X_i, with $\hat{\rho}(x_i|\pi_i, c, D, G)$ to represent the estimation of $\rho(x_i|\pi_i, c, G)$, and then

$$\hat{\rho}(x_i|\pi_i, c, D, G) = \frac{\hat{\rho}(x_i, \pi_i|c, D, G)}{\hat{\rho}(\pi_i|c, D, G)}$$
$$= \frac{\frac{1}{N(c)(\sqrt{2\pi}h^2)^{t(i)+1}}\sum_{m=1}^{N}[sign(c_m)\prod_{s=1}^{t(i)+1} K_i(\frac{x_i^s - x_{im}^s}{h_i})]}{\frac{1}{N(c)(\sqrt{2\pi}h^2)^{t(i)}}\sum_{m=1}^{N}[sign(c_m)\prod_{s=1}^{t(i)} K_i(\frac{x_i^s - x_{im}^s}{h_i})]} \quad (4)$$