Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 1761–1771
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Bayesian network classification using spline-approximated kernel density estimation

Yaniv Gurwicz, Boaz Lerner *

Pattern Analysis and Machine Learning Lab, Department of Electrical and Computer Engineering, Ben-Gurion University, P.O. Box 653, 84105 Beer-Sheva, Israel

Received 7 November 2004; received in revised form 7 November 2004
Available online 8 April 2005
Communicated by E. Backer

Abstract

The likelihood for patterns of continuous features needed for probabilistic inference in a Bayesian network classifier (BNC) may be computed by kernel density estimation (KDE), letting every pattern influence the shape of the probability density. Although usually leading to accurate estimation, the KDE suffers from computational cost making it unpractical in many real-world applications. We smooth the density using a spline thus requiring for the estimation only very few coefficients rather than the whole training set allowing rapid implementation of the BNC without sacrificing classifier accuracy. Experiments conducted over a several real-world databases reveal acceleration in computational speed, sometimes in several orders of magnitude, in favor of our method making the application of KDE to BNCs practical.

Keywords: Bayesian networks; Classification; Kernel density estimation; Naïve Bayesian classifier; Spline

1. Introduction

1.1. Density estimation for Bayesian network classifiers

A Bayesian network (BN) represents the joint probability distribution (density) $p(\boldsymbol{X})$ over a set of n domain variables $X = \{X_1, \ldots, X_n\}$ graphically (Pearl, 1988; Heckerman, 1995). An arc and a lack of an arc between two nodes in the graph demonstrate, respectively, dependency and independency between variables corresponding to these nodes (Fig. 1). A connection between X_i and its parents $\boldsymbol{Pa}_i$ in the graph is quantified probabilistically using the data. A node having no parents embodies the prior probability of the corresponding variable. By ordering the variables topologically, extracting the general factorization of this ordering (using

* Corresponding author.
E-mail addresses: yanivg@ee.bgu.ac.il (Y. Gurwicz), boaz@ee.bgu.ac.il (B. Lerner).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.12.008

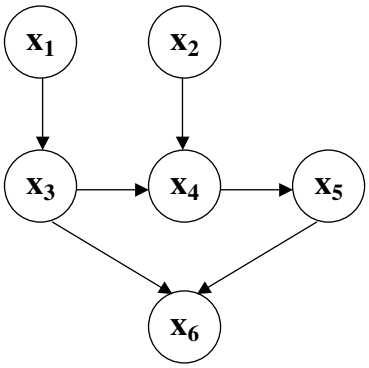

Fig. 1. A graph of an example Bayesian network. Arcs manifest dependencies between nodes representing variables.

the chain rule of probability) and applying the directed Markov property, we can decompose the joint probability distribution (density)

$$p(\boldsymbol{X}) = p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\boldsymbol{Pa}_i). \quad (1)$$

The naïve Bayesian classifier (NBC) is a BN used for classification thus belonging to the Bayesian network classifier (BNC) family (John and Langley, 1995; Heckerman, 1995; Friedman et al., 1998; Lerner, 2004). It predicts a class C for a pattern $\boldsymbol{x}$ using Bayes' theorem

$$P(C|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(\boldsymbol{X} = \boldsymbol{x}|C) \cdot P(C)}{p(\boldsymbol{X} = \boldsymbol{x})} \quad (2)$$

i.e., it infers the posterior probability that $\boldsymbol{x}$ belongs to C, $P(C|\boldsymbol{X} = \boldsymbol{x})$, by updating the prior probability for that class, $P(C)$, by the class-conditional probability density or likelihood for $\boldsymbol{x}$ to be generated from this class, $p(\boldsymbol{X} = \boldsymbol{x}|C)$, normalized by the unconditional density (evidence), $p(\boldsymbol{X} = x)$. The NBC represents a restrictive assumption of conditional independence between the variables (domain features) given the class allowing the decomposition and computation of the likelihood employing local probability densities

$$p(\boldsymbol{X}|C) = \prod_{i=1}^{n} p(X_i|C). \quad (3)$$

Estimating probability densities of variables accurately is a crucial task in many areas of machine learning (Silverman, 1986; Bishop, 1995). While estimating the probability distribution of a discrete feature is easily performed by computing the frequencies of its values in a given database, the probability density of a continuous feature taking any value in an interval cannot be estimated similarly thus requiring other, more complex methodologies. This is a major difficulty in the implementation of BNCs (John and Langley, 1995; Friedman et al., 1998; Elgammal et al., 2003; Lerner, 2004), and it requires either discretization of the variable into a collection of bins covering its range (Heckerman, 1995; Friedman et al., 1998; Yang and Webb, 2002; Malka and Lerner, 2004) or estimation, using parametric, non-parametric or semi-parametric methods (John and Langley, 1995; Lerner, 2004). Discretization is usually chosen for problems having small sample sizes that cannot guarantee accurate density estimation (Yang and Webb, 2002). Noticeably, prediction based on discretization is prone to errors due to lost of information. Generally, the accuracy discretization methods provide will peak for a specific range of bin sizes deteriorating as moving away from the center of this range (Malka and Lerner, 2004). A too small number of bins will smooth the estimated density and a too large number of bins will lead to the curse of dimensionality resulting in performance worsening in both cases. Besides, a too large number of bins will overload the calculation.

In parametric density estimation we assume a model describing the density and look for the optimal parameters for this model. For example, for a Gaussian model we ought estimating the data mean and variance. A single Gaussian estimation (SGE) is straightforward to implement and it bares almost no computational load to the NBC but its accuracy declines with the degree of deviation of the data from normality, which is expected in many real-world problems (John and Langley, 1995; Lerner, 2004). Extending parametric density estimation using Bayesian approaches (Heckerman, 1995), we update an a priori probability (e.g., Dirichlet prior) on the parameters using the likelihood for the data, thus combining prior and acquired knowledge jointly. However, when enough data is available (and the number of parameters is not too large) the likelihood in Bayesian estimation