



Learning Bayesian network classifiers by risk minimization

Roy Kelner^a, Boaz Lerner^{b,*}

^a Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

^b Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ARTICLE INFO

Article history:

Received 22 June 2011

Received in revised form 1 October 2011

Accepted 24 October 2011

Available online 29 October 2011

Keywords:

Bayesian networks

Classification

Probabilistic graphical models

Structure learning

ABSTRACT

Bayesian networks (BNs) provide a powerful graphical model for encoding the probabilistic relationships among a set of variables, and hence can naturally be used for classification. However, Bayesian network classifiers (BNCs) learned in the common way using likelihood scores usually tend to achieve only mediocre classification accuracy because these scores are less specific to classification, but rather suit a general inference problem. We propose risk minimization by cross validation (RMCV) using the 0/1 loss function, which is a classification-oriented score for unrestricted BNCs. RMCV is an extension of classification-oriented scores commonly used in learning restricted BNCs and non-BN classifiers. Using small real and synthetic problems, allowing for learning all possible graphs, we empirically demonstrate RMCV superiority to marginal and class-conditional likelihood-based scores with respect to classification accuracy. Experiments using twenty-two real-world datasets show that BNCs learned using an RMCV-based algorithm significantly outperform the naive Bayesian classifier (NBC), tree augmented NBC (TAN), and other BNCs learned using marginal or conditional likelihood scores and are on par with non-BN state of the art classifiers, such as support vector machine, neural network, and classification tree. These experiments also show that an optimized version of RMCV is faster than all unrestricted BNCs and comparable with the neural network with respect to run-time. The main conclusion from our experiments is that unrestricted BNCs, when learned properly, can be a good alternative to restricted BNCs and traditional machine-learning classifiers with respect to both accuracy and efficiency.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

One fundamental task of machine learning is classification, where instances (patterns) are assigned to their corresponding classes. This is a supervised learning task, where a training dataset of instances with labels representing instance classes is used to train a classifier. Since a Bayesian (belief) network (BN) [1–3] provides a graphical model for encoding relationships, such as dependencies and conditional independencies between variables, and for inferring probabilistically about variables, it is natural to use the BN for classification.

Indeed, along with the traditional classifiers based on the neural network (NN), the support vector machine (SVM), and the (decision) classification tree (CT), classifiers based on BN have recently been introduced and studied [4–11]. Learning a Bayesian network classifier (BNC) requires learning the structure (graph) of the graphical model and its parameters so that the learned BN will excel in inference of a specific variable, that is the class variable, and not necessarily of all variables. When focusing on structure learning, exhaustively searching the space of possible graphs is infeasible [2], and thus search and score (S&S) structure learning algorithms sub-optimally search the space and select the structure achieving the highest value of a score [2,3,12]. However, until very recently, all S&S structure learning algorithms used a generative score, and thereby led to learning a generative model that is not specific to classification, but to general inference.

* Corresponding author. Tel.: +972 8 6479375; fax: +972 8 6472958.

E-mail addresses: kelnerr@bgu.ac.il (R. Kelner), boaz@bgu.ac.il (B. Lerner).

Common approaches for learning a model are roughly partitioned into generative, discriminative, or a combination of both approaches [13–15]. Generative models (e.g., BN, density estimation) summarize data probabilistically and are more flexible, since the user can bring in conditional independence assumptions, priors, and hidden variables. Generative classifiers learn a model of the joint probability of the variables and the related class label, and use Bayes' theorem to compute the posterior probability of the class variable and make predictions. Discriminative models (e.g., NN and SVM) only learn from data to make accurate predictions by directly estimating the class posterior probability or via discriminant functions, and thus offer the user less flexibility in data representation and inference. The dilemma in the machine learning community regarding which approach of learning – generative or discriminative – is more appropriate for learning a BNC structure, has gained considerable attention in recent years. Most empirical studies demonstrate superiority of the discriminative approach with respect to the accuracy of the learned BNC [5–7, 16]. For some models, however, it is shown [13] that the choice of either of the approaches depends on the sample size; for small sample sizes, the generative approach, which relatively quickly approaches its asymptotic error, is favored, whereas the discriminative method is preferred for larger sample sizes. Classifiers combining generative and discriminative modeling use generative models, yet estimate the model structure and/or parameters to reduce the classification error.

Several studies [4–7, 10, 17] have demonstrated that BNC structures learned using generative scores do not usually contribute to high classification accuracy since there is lack of agreement between the score used for learning and the score used for evaluation, i.e., the classification accuracy. That is, classifiers based on structures having high values of the generative scores are not necessarily highly accurate. To address this issue, we propose risk minimization by cross validation (RMCV) for a classification-oriented score and S&S algorithm for learning unrestricted BNCs. Note that other uses of classification-oriented scores in learning unrestricted BNCs [7, 18] are in a somewhat different context. Moreover, RMCV is an extension to common use of classification-oriented scores in learning restricted-BNCs and non-BN classifiers. While commonly used S&S algorithms use likelihood-based scores suitable for general inference, RMCV minimizes an empirical estimation of the classification error rate, and thereby learns highly accurate BNCs. That is, RMCV performs discriminative learning of a generative (BN) model. This model does not need to estimate the true distribution, generate data from this distribution, or infer about any non-class variable. It needs to perform a discriminative classification task. RMCV learns generative models that are complicated, only to discriminate accurately among classes.

In the beginning, we suggest and compare several variants of the RMCV score and algorithm. We further show that the RMCV score is better suited for classification than any other score commonly used for learning a BNC, i.e., compared with other scores and BNCs, the accuracy of an RMCV-based classifier increases monotonically with the improvement in the value of the RMCV score. Then, we compare the classifier learned using RMCV with likelihood-based, conditional-likelihood-based, and other BNCs, as well as with non-BN classifiers, such as NN, SVM, and CT. This involves nine leading BNCs and five state of the art non-BN classifiers in a most extensive comparison of BNCs and non-BN classifiers using twenty-two real-world datasets. The comparison demonstrates that an RMCV-based classifier is faster and significantly more accurate than all BNCs and comparable (usually favorably), with respect to accuracy, with the non-BN classifiers. These are encouraging news for researchers and practitioners who appreciate the benefits of the BN model, but once they encounter a classification task, replace the BN with a traditional classifier and thereby lose the BN benefits.

We begin by reviewing BN in Section 2. In Section 3, we focus on learning a BNC using common scores. The RMCV score and algorithm are presented in Section 4, and classifiers learned using RMCV are experimentally compared to other BNCs and non-BN classifiers in Section 5. Section 6 describes recent studies in learning a BNC. Finally, we draw conclusions and summarize the study in Sections 7 and 8, respectively.

2. Bayesian networks

A BN model \mathcal{B} for a set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, each having a finite set of mutually exclusive states, consists of two main components, $\mathcal{B} = (\mathcal{G}, \Theta)$. The structure $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph (DAG). \mathbf{V} is a finite set of nodes of \mathcal{G} corresponding to \mathbf{X} , and \mathbf{E} is a finite set of directed edges of \mathcal{G} connecting \mathbf{V} . Θ is a set of parameters that quantify the structure. The parameters are local conditional probability distributions (or densities), $P(X_i = x_i | \mathbf{Pa}_i, \mathcal{G})$, for each $X_i \in \mathbf{X}$ conditioned on its parents in the graph, $\mathbf{Pa}_i \subset \mathbf{X}$. In this study, we are interested only in discrete variable BNs and complete data.

The joint probability distribution over \mathbf{X} given \mathcal{G} – assumed to encode this distribution – is the product of these local probability distributions [2, 3],

$$P(\mathbf{X} = \mathbf{x} | \mathcal{G}) = \prod_{i=1}^n P(X_i = x_i | \mathbf{Pa}_i, \mathcal{G}), \quad (1)$$

where \mathbf{x} is the assignment of states to the variables in \mathbf{X} and x_i is X_i 's state.

During inference, the conditional probability distribution of a subset of nodes in the graph (the 'hidden' nodes) given another subset of nodes (the 'observed' nodes) and the BN model is calculated. A common method for exact inference is the junction tree algorithm [19], but when there is only one hidden node (e.g., the class node in classification), direct inference based on (1) and Bayes' rule is more feasible. Note that the computation of conditional probability distributions for inference depends on the graph. Thus, a structure, either based on expert knowledge or learned from the data, must first be obtained.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات