Pattern Recognition 46 (2013) 464–471

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Stochastic margin-based structure learning of Bayesian network classifiers

Franz Pernkopf*, Michael Wohlmayr

Laboratory of Signal Processing and Speech Communication, Graz University of Technology, Austria

ARTICLE INFO

Article history:
Received 29 November 2011
Received in revised form
24 May 2012
Accepted 4 August 2012
Available online 17 August 2012

Keywords:
Bayesian network classifier
Discriminative learning
Maximum margin learning
Structure learning

ABSTRACT

The margin criterion for parameter learning in graphical models gained significant impact over the last years. We use the maximum margin score for discriminatively optimizing the structure of Bayesian network classifiers. Furthermore, greedy hill-climbing and simulated annealing search heuristics are applied to determine the classifier structures. In the experiments, we demonstrate the advantages of maximum margin optimized Bayesian network structures in terms of classification performance compared to traditionally used discriminative structure learning methods. Stochastic simulated annealing requires less score evaluations than greedy heuristics. Additionally, we compare generative and discriminative parameter learning on both generatively and discriminatively structured Bayesian network classifiers. Margin-optimized Bayesian network classifiers achieve similar classification performance as support vector machines. Moreover, missing feature values during classification can be handled by discriminatively optimized Bayesian network classifiers, a case where purely discriminative classifiers usually require mechanisms to complete unknown feature values in the data first.

1. Introduction

Generative probabilistic classifiers optimize the joint probability distribution of the features $\mathbf{X}$ and the corresponding class labels C using maximum likelihood (ML) estimation. The class label is usually predicted using the maximum a posteriori estimate of the class posteriors $P(C|\mathbf{X})$ obtained by applying Bayes rule. Discriminative probabilistic classifiers such as logistic regression model $P(C|\mathbf{X})$ directly. Discriminative classifiers may lead to better classification performance, particularly when the class conditional distributions poorly approximate the true distribution [1].

Basically, in Bayesian network classifiers both parameters and structure can be learned either generatively or discriminatively [2]. Discriminative learning requires objective functions such as classification rate (CR), conditional log-likelihood (CL), or the margin (as we propose to use in this paper), that optimize the model for a particular inference scenario, e.g. for a classification task. We are particularly interested in learning the discriminative structure[1] of a generative Bayesian network classifier that factorize as $P(C,\mathbf{X}) = P(\mathbf{X}|C)P(C)$.

Learning the graph structure of a Bayesian network classifier is hard. Optimally learning various forms of constrained Bayesian network structures is NP-hard [3] even in the *generative* sense. Recently, approaches for finding the *globally optimal* generative Bayesian network structure have been proposed. These methods are based on dynamic programming [4,5], branch-and-bound techniques [6,7], or search over various variable orderings [8]. The experiments of these *exact* methods are restricted to $\sim$50 variable nodes. Alternatively, *approximate* methods such as stochastic search or greedy heuristics are used, which can handle cases with many more variables.

Discriminative structure learning is not less difficult because of the non-decomposability[2] of the scores. Discriminative structure learning methods – relevant for learning Bayesian network classifiers – are usually approximate methods based on local search heuristics. In [9], a greedy hill-climbing heuristic is used to learn a classifier structure using the CR score. Particularly, at each iteration one edge is added to the structure which complies with the restrictions of the network topology and the acyclicity constraints of a Bayesian network. In a similar algorithm, the CL has been applied for discriminative structure learning [10]. Recently, we introduced a computationally efficient order-based greedy search heuristic for finding discriminative structures [2]. Our order-based structure learning is based on the observations in [11] and shows similarities to the K2 heuristic [12]. However, we proposed to use a discriminative scoring metric (i.e. CR) and suggest approaches for establishing the variable ordering based on conditional mutual information [13].

One of the most successful discriminative classifiers, namely the support vector machine (SVM), finds a decision boundary which maximizes the margin between samples of distinct classes

* Corresponding author. Tel.: +43 316 873 4436; fax: +43 316 873 10 4436.
E-mail addresses: pernkopf@tugraz.at (F. Pernkopf), michael.wohlmayr@tugraz.at (M. Wohlmayr).

[1] Discriminative scoring functions (e.g. classification rate, conditional log-likelihood, or the margin) are used for structure learning.

[2] Unfortunately, discriminative scores are usually not decomposable, while generative scores, e.g. log likelihood, are decomposable, i.e. they can be expressed as a sum of terms where each term depends on a variable and its conditioning variables (parents).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.08.007

resulting in good generalization properties [14] of the classifier. Recently, the margin criterion has been applied to learn the parameters of probabilistic models. Taskar et al. [15] observed that undirected graphical models can be efficiently trained to maximize the margin. More recently, Guo et al. [16] introduced the maximization of the margin for parameter learning based on convex relaxation to Bayesian networks. We proposed to use a conjugate gradient algorithm for maximum margin optimization of the parameters and show its advantages with respect to computational requirements [17]. Further generative and discriminative *parameter* learning methods for Bayesian network classifiers are summarized in [2,17] and references therein.

In this paper, we use the maximum margin (MM) criterion for discriminative *structure learning*. We use greedy hill-climbing (HC) and stochastic search heuristics such as simulated annealing (SA) [18,19] for learning discriminative classifier structures. SA is less prone to get stuck in local optima. We empirically evaluate our margin-based discriminative structure learning heuristics on two handwritten digit recognition tasks, one spam e-mail, and one remote sensing data set. We use naive Bayes (NB) as well as generatively and discriminatively optimized tree augmented naive Bayes (TAN) [20] structures. Furthermore, we experimentally compare both discriminative and generative parameter learning on both discriminative and generatively structured Bayesian network classifiers. Maximum margin *structure* learning outperforms recently proposed generative and discriminative structure learning approaches. SA heuristics mostly lead to better performing structures at a lower number of score evaluations (CR or MM) compared to HC methods. Discriminative parameter learning produces a significantly better classification performance than ML parameter learning on the same classifier structure. This is especially valid for cases where the structure of the underlying model is not optimized for classification [21]. We introduced the MM score for structure learning in [22] using the HC heuristic. The benefit of the MM score over other discriminative scores (i.e. CR) remained open in [22] since the HC heuristic might get trapped in local optimal solutions. This makes the reported performance gain of the MM score during structure learning ambiguous—either MM is useful, or the HC heuristic using CR gets stuck in low-performing local optimal solutions. For this reason we use SA which partially alleviates this problem. Recently, we also used the MM score for *exact* structure learning of Bayesian network classifiers [23]. This method is capable to find the global optimal solution. It is based on branch-and-bound techniques within a linear programming framework which offers the advantage of an *any-time* solution, i.e. premature termination of the algorithm returns the currently best solution together with a worst-case sub-optimality bound. Empirically it is shown that MM optimized structures compete with SVMs and outperform generatively learned network structures. Unfortunately, experiments are limited to rather small-scale data sets from the UCI repository [24]. To overcome these limitations, we use *approximate* methods for structure learning in this paper.

The paper is organized as follows: In Section 2, we introduce Bayesian network classifiers as well as NB and TAN structures. In Section 3, we present the non-decomposable discriminative scores CL, CR, and MM. Additionally, we discuss techniques for making the determination of these discriminative scores computationally competitive. Section 4 introduces different structure learning heuristics. Particular focus is on SA which is rarely used for discriminative learning of Bayesian network structures. In Section 5, we present experimental results. Section 6 concludes the paper.

2. Bayesian network classifiers

A Bayesian network [25] $\mathcal{B}=\langle \mathcal{G},\Theta \rangle$ is a directed acyclic graph $\mathcal{G}=(\mathbf{Z},\mathbf{E})$ consisting of a set of nodes $\mathbf{Z}$ and a set of directed edges

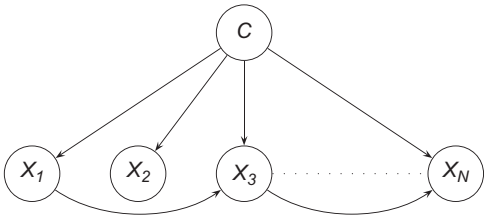

Fig. 1. An example of a TAN classifier structure.

$\mathbf{E}=\{E_{Z_i,Z_j},E_{Z_i,Z_k},\ldots\}$ connecting the nodes where E_{Z_i,Z_j} is an edge directed from Z_i to Z_j. This graph represents factorization properties of the distribution of a set of random variables $\mathbf{Z}=\{Z_1,\ldots,Z_{N+1}\}$. The variables in $\mathbf{Z}$ have values denoted by lower case letters $\mathbf{z}=\{z_1,z_2,\ldots,z_{N+1}\}$. We use boldface capital letters, e.g. $\mathbf{Z}$, to denote a set of random variables and correspondingly boldface lower case letters denote a set of instantiations (values). Without loss of generality, in Bayesian network classifiers the random variable Z_1 represents the class variable $C\in\{1,\ldots,|C|\}$, where $|C|$ represents the number of classes and $\mathbf{X}_{1:N}=\{X_1,\ldots,X_N\}=\{Z_2,\ldots,Z_{N+1}\}$ denotes the set of random variables representing the N attributes of the classifier. In a Bayesian network each node is independent of its non-descendants given its parents. The set of parameters which quantify the network are represented by Θ. Each random variable Z_j is represented as a local conditional probability distribution given its parents Z_{Π_j}. We use $\theta^j_{i|h}$ to denote a specific conditional probability table entry (assuming discrete variables); the probability that variable Z_j takes on its i^{th} value assignment given that its parents Z_{Π_j} take their h^{th} assignment, i.e. $\theta^j_{i|h}=P_\Theta(Z_j=i|Z_{\Pi_j}=h)$. The training data consists of M independent and identically distributed samples $\mathcal{S}=\{\mathbf{z}^m\}_{m=1}^M=\{(c^m,\mathbf{x}^m_{1:N})\}_{m=1}^M$ where $M=|\mathcal{S}|$. The joint probability distribution of a Bayesian network factorizes according to the graph structure and is given for a sample $\mathbf{z}^m$ as

$$P_\Theta(\mathbf{Z}=\mathbf{z}^m)=\prod_{j=1}^{N+1}P_\Theta(Z_j=z_j^m|Z_{\Pi_j}=z^m_{\Pi_j}). \tag{1}$$

The class labels are predicted according to

$$c^*=\operatorname*{argmax}_{c\in\{1,\ldots,|C|\}}P_\Theta(C=c|\mathbf{X}_{1:N}=\mathbf{x}^m_{1:N}), \tag{2}$$

$$c^*=\operatorname*{argmax}_{c\in\{1,\ldots,|C|\}}P_\Theta(C=c,\mathbf{X}_{1:N}=\mathbf{x}^m_{1:N}), \tag{3}$$

where the last equality follows from neglecting $P_\Theta(\mathbf{X}_{1:N})$ in $P_\Theta\ (C|\mathbf{X}_{1:N})=P_\Theta(C,\mathbf{X}_{1:N})/P_\Theta(\mathbf{X}_{1:N})$.

In this work, we restrict ourselves to NB and TAN structures. The NB network assumes that all the attributes are conditionally independent given the class label. As reported in [20], the performance of the NB classifier is surprisingly good even if the conditional independence assumption between attributes is unrealistic or even wrong for most of the data. Friedman et al. [20] introduced the TAN classifier which is based on structural augmentations of the NB network. In order to relax the conditional independence properties of NB, each attribute may have at most one other attribute as an additional parent. This means that the tree-width of the attribute induced sub-graph is unity, i.e. we have to learn a 1-tree over the attributes. A TAN classifier structure is shown in Fig. 1. In [2], we noticed that k-trees over the features – k indicates the tree-width[3] – often do not improve

[3] The tree-width of a graph is defined as the size of the largest clique (i.e. number of variables within the largest clique) of the moralized and triangulated directed graph minus one. Since there can be multiple triangulated graphs, the tree-width is defined by the triangulation where the largest clique contains the fewest number of variables. More details are given in [26] and references therein.