International Journal of Approximate Reasoning 51 (2010) 544–557

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Learning locally minimax optimal Bayesian networks

Tomi Silander *, Teemu Roos, Petri Myllymäki

Helsinki Institute for Information Technology (HIIT), P.O. Box 68, 00014 University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 10 December 2008
Received in revised form 15 June 2009
Accepted 25 August 2009
Available online 28 January 2010

Keywords:
Bayesian networks
Minimum description length
Model selection
Prediction
Normalized maximum likelihood,
Information theory

ABSTRACT

We consider the problem of learning Bayesian network models in a non-informative setting, where the only available information is a set of observational data, and no background knowledge is available. The problem can be divided into two different subtasks: learning the structure of the network (a set of independence relations), and learning the parameters of the model (that fix the probability distribution from the set of all distributions consistent with the chosen structure). There are not many theoretical frameworks that consistently handle both these problems together, the Bayesian framework being an exception. In this paper we propose an alternative, information-theoretic framework which sidesteps some of the technical problems facing the Bayesian approach. The framework is based on the minimax optimal normalized maximum likelihood (NML) distribution, which is motivated by the minimum description length (MDL) principle. The resulting model selection criterion is consistent, and it provides a way to construct highly predictive Bayesian network models. Our empirical tests show that the proposed method compares favorably with alternative approaches in both model selection and prediction tasks.

1. Introduction

Bayesian networks [1,2] are one of the most popular model classes for multivariate data. Learning a Bayesian network from data reveals the probabilistic structure of the domain and provides a tool for predicting future observations. Under certain restrictions and assumptions, Bayesian networks even allow principled speculations about the causal mechanisms of the domain, and provide estimates about effects of interventions [3].

Traditionally, learning of Bayesian networks has been divided in two separate tasks: learning the structure of the network that represents conditional independence relations, and learning the parameters that specify the joint probability distribution, see [4]. The methods for learning the structure are usually based on either conditional independence tests [5,6], or some scoring function such as *a posteriori* probability or description length, see [7]. These methods are not totally separate and there are also some hybrid methods [8,9].

Methods based on conditional independence tests are sensitive to choice of significance levels. Furthermore, since they are based on interpretation of Bayesian network structures as sets of independence assumptions, they do not usually offer a natural way to learn the parameters for the structure.

The popular Bayesian BDeu [10] criterion for learning Bayesian network structures has recently been reported to be very sensitive to the choice of prior hyperparameters [11,12]. On the other hand, some alternative model selection criteria, like the Akaike information criterion (AIC) [13] and the Bayesian information criterion (BIC) [14], are derived through asymptotics, and their behavior is suboptimal for finite sample sizes, nor do they suggest a particular way to learn the parameters for

* Corresponding author. Present address: A*STAR Institute of High Performance Computing, Singapore. Tel.: +65 64191301.
E-mail addresses: tomi.silander@iki.fi (T. Silander), teemu.roos@hiit.fi (T. Roos), petri.myllymaki@hiit.fi (P. Myllymäki).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.01.012

Bayesian networks. To our knowledge, apart from the methods presented in this paper, the Bayesian approach is one of the very few frameworks that offer a theoretically coherent solution to both structure and parameter learning.

For large networks, the study of different scoring criteria is hindered by the fact that learning the network structure is NP-hard for all popular scoring criteria [15], even if these criteria have a convenient characteristic of decomposability, which allows incremental scoring in heuristic local search [16]. However, owing to recent advances in exact structure learning [17,18], it is feasible to find the optimal network for decomposable scores when the number of variables is about 30 or less. This makes it possible to study the behavior of different scoring criteria for problems of realistic size without the uncertainty stemming from heuristic search.

In this paper we introduce a new decomposable scoring criterion for learning Bayesian network structures, the *factorized normalized maximum likelihood* (fNML). This score features no tunable parameters, and thus avoids the sensitivity problems of Bayesian scores. We show that the new criterion is asymptotically consistent. Unlike AIC and BIC, it is derived in closed form for finite sample sizes, and it has a probabilistic interpretation as a distribution which has a certain minimax optimality property.

We also use the predictive form of the normalized maximum likelihood (NML) model [19] to find well predicting parameters given the learned network structure. This new method for learning the parameters, which we call the *factorized-sequential normalized maximum likelihood* (fsNML), is a natural extension of the fNML model selection criterion for predictive purposes. In order to demonstrate the theoretical validity of fsNML, we give a non-asymptotic upper-bound on the logarithmic loss (or code length) of the fsNML predictions relative to the optimal parameters – for a fixed graph structure, the fsNML predictions are never (for any data set) much worse than those obtained by optimizing the parameters with hindsight. Both the fNML and fsNML methods are motivated by the Minimum Description Length (MDL) principle, see [20,7].

The rest of the paper is structured as follows. In Section 2, we first introduce Bayesian networks and the notation needed later. In Section 3, we first briefly review the most popular decomposable scores, after which we are ready to introduce the fNML criterion for structure learning. In Section 4 we turn our focus to the parameter learning and introduce our sNML-based solution. We then describe the empirical experiments and their results in Section 5 and draw conclusions in Section 6.Proofs for some less significant results can be found in appendices at the end of the paper.

2. Bayesian networks

We assume the reader to be familiar with Bayesian networks (for a tutorial, see [4]), and only introduce the notation needed later in this paper.

A Bayesian network defines a joint probability distribution for an m-dimensional multivariate data vector $X = (X_1, \ldots, X_m)$. We will only consider cases in which all the variables are discrete, so that variable X_i may have r_i different values which, without loss of generality, may be denoted $\{1, \ldots, r_i\}$.

A Bayesian network consists of a directed acyclic graph G and a set of conditional probability distributions. We specify the DAG with a vector $G = (G_1, \ldots, G_m)$ of parent sets so that $G_i \subset \{X_1, \ldots, X_m\}$ denotes the parents of variable X_i, i.e., the variables from which there is an arc to X_i. Each parent set G_i has $q_i (q_i = \prod_{X_p \in G_i} r_p)$ possible values that are the possible value combinations of the variables belonging to G_i. We assume a non-ambiguous enumeration of these values and denote the event that G_i holds the jth value combination simply by $G_i = j$.

The local Markov property for Bayesian networks states that each variable is independent of its non-descendants given its parents. Formally, this is equivalent to the following factorization of the joint distribution:

$$P(x|G) = \prod_{i=1}^{m} P(x_i|G_i). \tag{1}$$

The conditional probability distributions $P(X_i|G_i)$ are determined by a set of parameters, Θ, via the equation

$$P(X_i = k|G_i = j, \Theta) = \theta_{ijk},$$

where k is a value of X_i, and j is a value configuration of the parent set G_i. We denote the set of parameters associated with variable X_i by Θ_i and define $\Theta_{ij} = (\Theta_{ij1}, \ldots, \Theta_{ijr_i})$.

For learning Bayesian network structures we assume a data D of N complete independent and identically distributed (i.i.d.) instantiations of the vector X, i.e., an $N \times m$ data matrix without missing values. It turns out to be useful to introduce a notation for certain parts of this data matrix. We often want to select rows of the data matrix by certain criteria. We then write the selection criterion as a superscript of the data matrix D. For example, $D^{G_i=j}$ denotes those rows of D where the variables of G_i have the jth value combination. If we further want to select certain columns of these rows, we denote the columns by subscripting D with a corresponding variable set. As a shorthand, we write $D_{\{X_i\}} = D_i$. For example, $D_i^{G_i=j}$ selects the ith column of the rows $D^{G_i=j}$.

Since the rows of D are assumed to be i.i.d., the probability of a data matrix can be calculated just by taking the product of the row probabilities. Combining equal terms yields

$$P(D|G, \Theta) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}, \tag{2}$$

where N_{ijk} denotes number of rows in $D^{X_i=k, G_i=j}$. We also define a vector $\vec{N}_{ij} = (N_{ij1}, \ldots, N_{ijr_i})$ and a sum $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.