International Journal of Approximate Reasoning 54 (2013) 526–540

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Locally averaged Bayesian Dirichlet metrics for learning the structure and the parameters of Bayesian networks

Andrés Cano, Manuel Gómez-Olmedo, Andrés R. Masegosa *, Serafín Moral

Department of Computer Science and Artificial Intelligence, University of Granada, Spain

ARTICLE INFO

Article history:
Available online 16 October 2012

Keywords:
Probabilistic graphical models
Bayesian networks
Structure learning
Parameter estimation
Bayesian metrics

ABSTRACT

The marginal likelihood of the data computed using Bayesian score metrics is at the core of *score+search* methods when learning Bayesian networks from data. However, common formulations of those Bayesian score metrics rely on free parameters which are hard to assess. Recent theoretical and experimental works have also shown that the commonly employed BDe score metric is strongly biased by the particular assignments of its free parameter known as *the equivalent sample size*. This sensitivity means that poor choices of this parameter lead to inferred BN models whose structure and parameters do not properly represent the distribution generating the data even for large sample sizes. In this paper we argue that the problem is that the BDe metric is based on assumptions about the BN model parameters distribution assumed to generate the data which are too strict and do not hold in real settings. To overcome this issue we introduce here an approach that tries to marginalize the meta-parameter locally, aiming to embrace a wider set of assumptions about these parameters. It is shown experimentally that this approach offers a robust performance, as good as that of the standard BDe metric with an optimum selection of its free parameter and, in consequence, this method prevents the choice of wrong settings for this widely applied Bayesian score metric.

1. Introduction

A Bayesian network [14] is a statistical model which provides a compact representation of a multivariate joint probability distribution based on a directed acyclic graph (DAG), where the nodes represent random variables and the edges are direct probabilistic relationships. This property, along with their graphical nature, make BNs excellent models for representing the complex probabilistic relationships existing in many real problems (see [16] for a recent review).

The problem of learning the structure of a Bayesian Network from a previously given set of observational data has been subjected to a great deal of research [8,13,10]. Very different approaches have been considered for this problem: *constraint-based* learning, based on carrying out several independence tests on the learning data set, leading to a Bayesian network in agreement with test results [20]; *algebraic and geometric based learning*, based on the representation of a BN using a uniquely determined vector, called a standard itemset, which can be interpreted as the set of vertices of a certain polytope [23,24,9]; and *scoring and searching*-type learning approaches [6,8], based on the definition of a score or metric function to assess how well a given BN explains the data and an heuristic search method which explores the different BN structures (the DAG space) and retrieves the one with the highest score.

* Corresponding author.
E-mail addresses: acu@decsai.ugr.es (A. Cano), mgomez@decsai.ugr.es (M. Gomez-Olmedo), andrew@decsai.ugr.es (A.R. Masegosa), smc@decsai.ugr.es (S. Moral).

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2012.09.003

To solve this problem, many score metrics inspired in several different principles (AIC [2], BIC [17], MDL [7], etc.) have been proposed. One the most widely used for multinomial data is the so-called Bayesian Dirichlet metric [4,8,6]. These metrics, which are based on Bayesian principles, score a BN by computing the marginal likelihood of the data given the graph structure. To compute this marginal likelihood for multinomial data, it is assumed a Dirichlet prior over the parameters of the BN. But these Dirichlet distributions depend on a parameter vector that has to be assessed. In the case of the widely applied BDe metric [8], these Dirichlet priors depend on a meta-parameter known as the *equivalent sample size* (ESS).

Roughly speaking, this meta-parameter aims to capture the strength of our prior belief in the uniformity of the network parameter distribution. Although it is very hard to assess in practical settings, it was thought to have little impact in the learning process since, as a prior distribution, its effect on the metric decreases with the size of the learning data. However, later works [19] have shown that the chosen ESS value of the BDe score strongly affects the selection of the *maximum a posteriori* (MAP) model even in data sets with large sample sizes. They found that large (small) ESS values usually retrieve very dense (sparse) BN models. For some UCI data sets used to carry out these evaluations, the number of edges monotonically increased from an empty network to a fully connected BN model when increasing the ESS. This same sensitivity was also found in simple independence assessments between two binary variables [12,1]. These works showed that changing the value of the ESS parameter has an impact on Type I and Type II errors from the hypothesis tests made by Bayesian metrics.

The sensitivity of the BDe metric to this ESS parameter has also been theoretically analyzed for very large or very small values of this parameter in several works [22,21,25]. They found that BDe score has an intrinsic tendency to favor either the presence or the absence of an edge between two variables depending on the particular ESS value. Furthermore, for large ESS values, this tendency was found to be independent of the particular probabilistic dependency between the variables. Specifically, Steck [21] showed that the BDe score can predict dependency between two independent random variables if their marginals are very skewed. A major consequence of this sensitivity of Bayesian scores arises when they are employed in knowledge discovery tasks: we will find that different conditional independencies discovered in the data strongly depend on a free parameter which, although it is supposed to represent our prior beliefs, it is also extremely complex to assess in practical settings.

A possible solution, originally suggested in [19], for the sensitivity of BDe metric to this parameter is to use a Bayesian approach: assume a prior distribution on the ESS parameter and marginalize it out in the score value. As no closed form solution is known to compute the integral required to perform this marginalization, a uniform over the ESS parameter was assumed and the integral was approximated by a simple averaging operation. However, the cited work does not give any evidence about whether this integrating approach retrieves an optimal approximation either in terms of capacity for predicting unseen data or in terms of correct inference of the underlying structure that generates the data. It only shows that this integration approach retrieves the same BN model than the one inferred using a few single ESS values.

On the other hand, the assumption of a particular Dirichlet prior for the parameters of a BN also defines the very relevant task of determining how to estimate the parameters of this model once the structure is learnt. In the case of the BDe metric, the ESS parameter defines the correction strength in these estimates (i.e. very small values of ESS will produce parameter estimates close to maximum likelihood estimates). In consequence, the predictions made by an inferred BN will also be very sensitive to this parameter. Again, this may cause again serious inconveniences in knowledge discovery tasks because the elicited parameters may change significantly depending on the specific values assigned to this free meta-parameter.

This paper is devoted to analyzing the performance of the BDe metric and justifying its improvement when the ESS parameter is locally marginalized. We demonstrate that the averaging approach pursued in [19] is not an optimal strategy to eliminate the effect of this kind of parameters on these Bayesian metrics. In that way, we introduce a novel and more powerful approach to marginalizing the free parameter, able to make better inferences when the parameter space of the model generating the data is very complex. In our experimental evaluation, we show that this strategy is quite robust and able to remove the sensitivity of the Bayesian metric to this meta-parameter when inferring the structure and the parameters of a BN. Thus, when learning the structure, this approach prevents the elicitation of spurious conditional independence relationships due to wrong assessments of this meta-parameter and can help to make the *score+search* methods for learning BNs from data more robust and usable approaches for knowledge discovery tasks. When learning the parameters of the BN, this approach also prevents an over- or under-regularization of the estimates, so the parameter estimates of the inferred BNs will be much more accurate.

The paper is organized as follows. In Section 2 we introduce the formulation of a Bayesian score metric to learn the graph structure from a data set. Section 3 presents the details and motivation behind our new proposal of locally averaged Bayesian metrics. The experimental evaluation of these proposals is depicted in Section 4. And finally, Section 5 contains the main conclusions and future work.

2. Bayesian Dirichlet metrics

Let us assume we are given a vector of n random variables $\mathbf{X} = (X_1, \ldots, X_n)$ each taking values in some finite domain $Val(X_i)$. A Bayesian network $\mathcal{B}$ is defined by two components: the structure of the network, denoted by G; and the associated values of the numerical parameters, denoted by Φ. The structure G is a directed acyclic graph where the nodes represent the variables in the domain, and the edges are direct probabilistic dependencies between them. This graph encodes a set of conditional independence assumptions which are called the *Markov condition*: each node X_i is conditionally independent of