Applied Soft Computing 13 (2013) 1110–1120

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

New skeleton-based approaches for Bayesian structure learning of Bayesian networks

Andrés R. Masegosa*, Serafín Moral

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

ARTICLE INFO

Article history:
Received 7 May 2012
Received in revised form 6 September 2012
Accepted 28 September 2012
Available online 22 October 2012

Keywords:
Probabilistic graphical models
Bayesian networks
Bayesian structure learning
Markov chain Monte Carlo
Stochastic search

ABSTRACT

Automatically learning the graph structure of a single Bayesian network (BN) which accurately represents the underlying multivariate probability distribution of a collection of random variables is a challenging task. But obtaining a Bayesian solution to this problem based on computing the posterior probability of the presence of any edge or any directed path between two variables or any other structural feature is a much more involved problem, since it requires averaging over all the possible graph structures. For the former problem, recent advances have shown that *search + score approaches* find much more accurate structures if the search is constrained by a previously inferred skeleton (i.e. a relaxed structure with undirected edges which can be inferred using local search based methods). Based on similar ideas, we propose two novel skeleton-based approaches to approximate a Bayesian solution to the BN learning problem: a new stochastic search which tries to find directed acyclic graph (DAG) structures with a non-negligible score; and a new Markov chain Monte Carlo method over the DAG space. These two approaches are based on the same idea. In a first step, both employ a previously given skeleton and build a Bayesian solution constrained by this skeleton. In a second step, using the preliminary solution, they try to obtain a new Bayesian approximation but this time in an unconstrained graph space, which is the final outcome of the methods. As shown in the experimental evaluation, this new approach strongly boosts the performance of these two standard techniques proving that the idea of employing a skeleton to constrain the model space is also a successful strategy for performing Bayesian structure learning of BNs.

1. Introduction

Bayesian networks (BN) are statistical models which allow to graphically represent, by means of a directed acyclic graph (DAG), complex structures of dependencies among stochastic variables [1]. They have been widely employed in a great variety of real world problems because of their excellent properties for reasoning under uncertainty [2].

The problem of automatic learning of the structure of a BN from a database has been the subject of a great deal of research [3–5]. Traditionally, two procedures have been considered for this problem: one based on *scoring and searching* [3,6]; and other, *constraint based* (CB) learning, based on carrying out several independence tests on the learning data set to build a Bayesian network in agreement with tests results [5]. However, in the past years several methods have been proposed which combine aspects of both basic procedures. For example [7,8], employ Bayesian scores to carry out the statistical tests in a PC-like algorithm. They showed that these scores reduce the average number of structural errors. Other works [9,10] introduce some PC variants which use a greedy procedure to introduce links, similar to the K2 algorithm [6], in order to make independent tests only relative to the added links.

But the most successful strategy has been based on the induction of a BN skeleton (i.e. a graph with undirected edges) by means of CB approaches and, in a second step, run a *score + search* method to find a maximum scoring structure over the DAG space constrained by this skeleton. To the best of our knowledge, Van Dijk et al. [11] were the first ones to propose a learning method based on this strategy. But Max Min-Hill Climbing (MM-HC) [12] is probably the best-known method of this kind of approaches. The idea of finding high scoring BNs in a restricted or constrained search space has been further pursued in many works [13–16]. For example [16], presents a method which is able to find the BN model constrained to a given skeleton with the global optimal score. At the same time, many other outperforming methods have been proposed in order to elicit the skeleton of a BN by means of local structure discovery methods (see [17] for a recent review).

* Corresponding author.
E-mail addresses: andrew@decsai.ugr.es (A.R. Masegosa), smc@decsai.ugr.es (S. Moral).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.09.029

In parallel to these works, researchers have also focused on the computation of full Bayesian solutions for the problem of learning BNs from data. They were motivated by the uncertainty in the model selection which is especially notorious in problems where the size of the learning data sets is usually small compared to the super-exponential size of the space of DAGs. Rather than look for a single BN maximizing a given score, like *score + search* approaches, the problem now is computing the marginal posterior probability of the edges of a BN [18] (or any other structural feature) and making predictions using a Bayesian model averaging approach.

Two different sets of approaches have been employed to obtain these Bayesian solutions. Markov chain Monte Carlo (MCMC) methods [18–22], whose aim is to sample different DAG structures according to the stationary distribution of the defined Markov chain; and stochastic search methods [23–25] which, unlike MCMC, do not have the goal of converging to a stationary distribution (which may be unattainable, and is usually impossible to assess), but simply listing and scoring a collection of high scoring models which are visited during the search process.

In this work we aim to show that the successful Max Min-Hill Climbing decomposition idea [12] for finding maximum scoring single BNs can also be applied to develop new methods for obtaining Bayesian solutions to the BN structure learning problem. More precisely, what we show is that the computation of a Bayesian approximation in two steps, firstly over the DAGs constrained by a previously given skeleton and, in a second step, over an unconstrained DAG space but using the information of the first step, is a very successful strategy for this problem. Two different implementations of this idea are evaluated in this work. The first one is a stochastic search method which, as first step, carries out a search constrained by the skeleton performing random movements according to the Bayesian scores of the alternative local movements. In the second step, this stochastic search is run again, starting randomly at any of the high scoring models found in the first step and without constraining the random local moves. The second method introduced in this work is a new MCMC in the DAG space. Similarly to the other method, we firstly run a MCMC constrained by the skeleton of the BN which should identify high scoring constrained DAG models. After that, in a second step, a standard MCMC over the DAG space is run without any constraint. However, this MCMC is conducted by a new proposal distribution, which introduces global movements using the samples generated by the first MCMC over the constrained DAG space.

This paper is structured as follows. In Section 2, we provide background knowledge on Bayesian scores and skeletons, as well as on Bayesian structure learning methods. Subsequently, in Section 3, we present the details of our two skeleton-based approaches. The experimental evaluation is given in Section 4. Finally, in Section 5, we provide the main conclusions and future research.

2. Background knowledge

2.1. The Bayesian score of a BN

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 BN is defined by a directed acyclic graph, denoted as G, which represents the dependency structure among the network variables. More precisely, this graph G is specified by means of a vector with the parent sets, $\Pi_i \subset \mathbf{X}$, of each variable $X_i \in \mathbf{X}$: $G = (\Pi_1, \ldots, \Pi_n)$. The parent set Π_i is represented in G by those variables with an edge pointing to X_i. The definition of a BN model is complete with a numerical vector, denoted by Θ, which contains the parameters of the conditional probability distributions encoded in this graph G: θ_{ij} is a vector of length $|Val(X_i)|$ ($|\cdot|$ is the cardinality operator) associated to the conditional multinomial distribution of $P(X_i|\Pi_i = j)$, where $\Pi_i = j$ denotes the jth assignment of the variables in Π_i. We also use $|Val(\Pi_i)|$ to denote the number of all the possible combinations.

Let us also assume we are given a fully observed multinomial data set D. To compute the marginal likelihood of the data given the graph structure, $P(D|G) = \int P(D|G, \Theta)P(\Theta|G)d\Theta$, the most common settings [3] define a prior Dirichlet distribution for each parameter θ_{ij} with parameter vector α_{ij}, $\theta_{ij} \sim Dir(\alpha_{ij})$. They also assume a set of parameter independence assumptions in order to factorize the joint probabilities and make feasible the computation of the multidimensional integral.

In that way, the marginal likelihood of data given a graph structure and a set of vectors of Dirichlet parameters, α_{ij}, has the following well-known closed-form equation:

$$P(D|G) = \prod_{i}^{n} \prod_{j=1}^{|Val(\Pi_i)|} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{|Val(X_i)|} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (1)$$

where N_{ijk} are the number of data instances in D consistent with jth assignment of Π_i and $X_i = k$, while $N_{ij} = \sum_k N_{ijk}$ and $\alpha_{ij} = \sum_k \alpha_{ijk}$. In the case of the Bayesian Dirichlet equivalent metric or BDeu metric, these α_{ijk} are set to $\alpha_{ijk} = (1/(|Val(\Pi_i)||Val(X_i)|))$. The relevance of these settings relies on the following property of this Bayesian metric known as *likelihood equivalence*: if two different BN models encode the same conditional independencies, then the score metric assigns the same score value to each model.

With the definition of a prior distribution for the graph structures, $P(G)$, we fully specify the *Bayesian score metric* of a graph structure: $score(G|D) = P(G)P(D|G)$. Furthermore, if the prior distribution is locally decomposable, $P(G) = \prod_i P_i(\Pi_i)$, then this score can also be made locally decomposable: $score(G|D) = \prod_i score(X_i, \Pi_i|D)$.

The prior over graph structures, $P(G)$, is usually taken to be uniform. However, as pointed out in several works [26–28], this uniform prior is not optimal, specially because it does not account for the problem of "multiplicity correction" (i.e. if the number of candidate parents for a variable X_i grows, the probability of edge inclusion should be decreased in order to control the number of false positive edges). So, these authors propose the following prior, which is also the one employed by the approaches presented in this work: $P(G) \propto \prod_i \binom{i}{|\Pi_i|}^{-1}$.

2.2. Learning the skeleton of a BN

In this subsection we give a brief overview of the main ideas behind the methods employed in this work to build the skeleton of the BN as a previous step to constrain the subsequent search in the DAG space. We will use SK to denote this skeleton, which is defined by a set of undirected edges between a pair of variables, $SK = \{X_i \leftrightarrow X_j : X_i, X_j \in \mathbf{X}\}$. We also use $Neighbors_{SK}(X_i)$ to denote the set of variables which have an undirected edge in SK connected to X_i (i.e. the neighbors of X_i in SK). As pointed out in [16], the main property that this skeleton should satisfy in order to define a correct constrained DAG search space, denoted by $\mathcal{G}_{SK}$, is that it must be a *super-structure* of the true DAG generating the data. We say