International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Efficient learning of bounded-treewidth Bayesian networks from complete and incomplete data sets

Mauro Scanagatta [a], Giorgio Corani [a], Marco Zaffalon [a], Jaemin Yoo [b], U. Kang [b]

[a] *IDSIA, Switzerland*
[b] *Seoul National University, Republic of Korea*

ARTICLE INFO

Article history:
Received 11 December 2017
Received in revised form 12 February 2018
Accepted 13 February 2018
Available online xxxx

Keywords:
Structural learning
Bounded treewidth
Bayesian networks
Structural EM
Incomplete data sets

ABSTRACT

Learning a Bayesian networks with bounded treewidth is important for reducing the complexity of the inferences. We present a novel anytime algorithm (k-MAX) method for this task, which scales up to thousands of variables. Through extensive experiments we show that it consistently yields higher-scoring structures than its competitors on complete data sets. We then consider the problem of structure learning from incomplete data sets. This can be addressed by structural EM, which however is computationally very demanding. We thus adopt the novel k-MAX algorithm in the maximization step of structural EM, obtaining an efficient computation of the expected sufficient statistics. We test the resulting structural EM method on the task of imputing missing data, comparing it against the state-of-the-art approach based on random forests. Our approach achieves the same imputation accuracy of the competitors, but in about one tenth of the time. Furthermore we show that it has worst-case complexity linear in the input size, and that it is easily parallelizable.

1. Introduction

The size of an explicit representation of the joint distribution of n categorical random variables is exponential in n. Bayesian networks [1] compactly represent joint distributions by exploiting independence relations and encoding them into a directed acyclic graph (DAG), also referred to as *structure*. Yet, algorithms able to perform structure learning from thousands of variables have been devised only very recently for Bayesian networks [2,3] and for chordal log-linear graphical models (that can be exactly mapped on Bayesian networks) [4,5].

Given a Bayesian network, the task of computing the marginal distribution of a set of variables, possibly given evidence on another set of variables, is called *inference*. The complexity of exact inference grows exponentially in the *treewidth* [1, Chap. 7] of the DAG, under the exponential time hypothesis [6]. In order to allow tractable inference we thus need to learn Bayesian networks with a bounded-treewidth structure; this problem is NP-hard [7].

Most research on learning bounded-treewidth Bayesian networks adopts a score-based approach. The score measures the fit of the DAG to the data; the goal is hence to find the highest-scoring DAG that respects the treewidth bound. Exact methods [7–9] exist, but their applicability is restricted to small domains. Approximate approaches that scale up to some

E-mail addresses: mauro@idsia.ch (M. Scanagatta), giorgio@idsia.ch (G. Corani), zaffalon@idsia.ch (M. Zaffalon), jaeminyoo@snu.ac.kr (J. Yoo), ukang@snu.ac.kr (U. Kang).

https://doi.org/10.1016/j.ijar.2018.02.004
0888-613X/

hundreds of variables [10,11] have been more recently proposed. A recent breakthrough has been achieved by the k-greedy algorithm [3]. It consistently yields higher-scoring DAGs than its competitors and it scales to several *thousands* of variables.

In this paper we present a new algorithm called k-MAX, which improves over k-greedy. Both k-MAX and k-greedy are anytime algorithms: they can be stopped at any moment, yielding the current best solution. k-MAX adopts a set of more sophisticated heuristics compared to k-greedy; as a result it consistently yields higher-scoring DAGs than both k-greedy and other competitors, as demonstrated by our extensive experiments on complete data sets.

Structure learning algorithms commonly assume data sets to be complete; yet real data sets are often incomplete. Structure learning on incomplete data sets can be accomplished via the *structural expectation-maximization* (SEM) algorithm [12], which alternates between an estimation of the sufficient statistics given the current model (expectation step), and the search of a new model given the expected sufficient statistics (maximization step). Yet, SEM is computationally demanding: in particular the expectation step requires computing several inferences, which might become prohibitive if the model has unbounded treewidth and/or there are many missing data whose actual value has to be inferred. We adopt k-MAX as the structure learning algorithm within SEM; in this way we obtain a fast implementation of SEM, since the bounded-treewidth structures learned in the different iterations perform efficient inferences. To the best of our knowledge, this is the first implementation of SEM that is able to scale to thousands of variables.

To test our method, we use the Bayesian networks learned by SEM in order to perform data imputation. We consider as a competitor a recent method for data imputation based on random forests [13] and we compare the two approaches on data sets with different degrees of missingness. The two approaches achieve the same imputation accuracy, but our approach is faster by almost one order of magnitude. Furthermore we show that the complexity of our method scales linearly in the input size (Subsec. 7.4), and that it is easily parallelizable (Subsec. 7.5). To the best of our knowledge, it is the first approach in the literature able to do so.

In Section 2 we present the technical background of the paper. In Section 3 we detail our approach for bounded-treewidth structure learning, k-MAX. In Section 4 and 5 we evaluate its performance against existing state-of-the-art approaches. In Section 6 we present how k-MAX can be used in the SEM algorithm, obtaining the SEM-k-MAX algorithm. It is evaluated in Section 7 on the task of data imputation against the state-of-the-art approach. Section 8 concludes our paper.

The software of this paper is available from http://ipg.idsia.ch/software/blip, together with supplementary material containing the detailed results of our experiments.

2. Treewidth and k-trees

Intuitively, the treewidth k quantifies the extent to which a graph resembles a tree. Following the terminology of [14] we now provide a formal definition. Let us recall that a *clique* of an undirected graph is a subset of its nodes such that every two distinct nodes are linked by an edge. Moreover, a clique is *maximal* if it is not a subset of a larger clique.

Treewidth of an undirected graph. We denote an undirected graph by $H = (V, E)$ where V is the vertex set and E is the edge set. An undirected graph is *triangulated* when every cycle of length greater than or equal to 4 has a *chord*, that is, an edge connecting two non-consecutive nodes in the cycle [1, Def. 9.16]. Triangulated graphs are also called *chordal* graphs. The *triangulation* of a graph is the operation of adding chords until the graph is triangulated. The treewidth of a triangulated graph is the size of its largest clique minus one. The treewidth of H is the minimum treewidth among all the possible triangulations of H.

Treewidth of a Bayesian network. The moral graph of the DAG associated to a Bayesian network is an undirected graph that includes an edge $(i - j)$ for every edge $(i \to j)$ in the DAG and an edge $(p - q)$ for every pair of edges $(p \to i)$, $(q \to i)$ in the DAG. The treewidth of the DAG is the treewidth of its moral graph.

2.1. k-trees

A k-tree is an undirected *edge-maximal* graph of treewidth k, that is, the addition of any edge to the k-tree increases its treewidth. It is defined inductively as follows [15]. *Base case*: a clique with $(k + 1)$ nodes is a k-tree. *Inductive step*: given a k-tree H_n on n nodes, a k-tree H_{n+1} on $(n + 1)$ nodes is obtained by connecting the $(n + 1)$-th node to a k-clique of H_n (a k-clique is a clique over k nodes). See Fig. 1 for an example. As a final remark, a sub-graph of a k-tree is called *partial k-tree*; its treewidth is at most k.

3. Structure learning of Bayesian networks

We consider the problem of learning the structure of a Bayesian network from a complete data set. The set of n categorical random variables is $\mathcal{X} = \{X_1, \ldots, X_n\}$. The goal is to find the highest-scoring bounded-treewidth DAG $\mathcal{G} = (V, E)$, where V is the collection of nodes and E is the collection of arcs. E can be represented by the set of parents $\Pi_1, \ldots, \Pi_n$ of all variables.

Structure learning is usually accomplished in two steps. First, *parent set identification* is the identification of a list (*cache*) L_i of candidate parent sets independently for each variable X_i. Second, *structure optimization* is the assignment of a parent set to each node in order to maximize the score of the resulting DAG.

Please cite this article in press as: M. Scanagatta et al., Efficient learning of bounded-treewidth Bayesian networks from complete and incomplete data sets, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2018.02.004