Causal Learning of Mixtures of Trees

François Schnitzler, Louis Wehenkel

Department of Electrical Engineering and Computer Science & GIGA-Research,
University of Liège, Belgium
fschnitzler@ulg.ac.be, L.Wehenkel@ulg.ac.be

Abstract: *Mixtures of trees can be used to model any multivariate distributions. In this work the possibility to learn these models from data by causal learning is explored. The algorithm developed aims at approximating all first order relationships between pairs of variables by a mixture of a given size. This approach is evaluated based on synthetic data, and seems promising.*

Keywords: Mixture, trees, causal learning

1 Introduction

A bayesian network (BN) represents the probability density of a set of random variables by exploiting independence relations among them. Such a model allows to perform prediction or optimization in uncertain problems.

However, the complexity of those operations scales poorly with the number of variables [1], whereas practical problems keep growing in size. Using a mixture of simple models has the potential to greatly improve the scalability of the algorithms, without necessarily leading to a decrease in accuracy. Indeed, a mixture of trees (MT) can perfectly model any probability density [2].

In this work we explore the possiblity to learn a MT by causal learning. This class of methods build BNs based on the independence relationships between variables derived from data. To assess the interest of this approach, we devised a first basic algorithm, and applied it to synthetic problems.

We will first briefly discuss scaling and modeling properties of BNs and MTs. In the second part, a short introduction to causal learning will be provided, followed by a description of our method. Finally, our testing methodology and first results will be presented.

2 Bayesian Networks

A BN over a set $\mathcal{X}$ of n discrete variables is a directed acyclic graph (i.e. a graph without directed cycle) associating to each variable one node of the graph and a conditional probability table over the parents of that node. Therefore, a BN encodes a joint distribution over $\mathcal{X}$, and it is theoretically possible to perform any inference over that distribution. However, it has been shown that this operation is NP-hard should the underlying undirected graph (skeleton) have cycles [1]. Stochastic algorithms have also been developed for that task, but their behavior is not guaranteed for large problems.

Trees are a subclass of BNs, for which each node of the graph has only one parent, and hence its skeleton is acyclic. For that last reason, standard operations are much more efficient over trees : learning the optimal tree (MWST) scales proportionally to n^2, and inference, to n. However, due to more drastic limitations on their structure, the modeling power of trees is inferior to that of a BN.

2.1 Mixtures of Trees

A mixture model of size m consists of a weighted set of different models, each of them defined over $\mathcal{X}$. The probability of an event is equal to the weighted sum of its probability encoded by each term of the mixture. We consider here mixtures of tree structured BN models.

In that case, it has been shown that it is possible to model any density, provided m is large enough. Moreover, the MT retains the scaling properties of trees: inference is still linear in the number of variables [2], making MT an attractive candidate for scaling graphical models to large problems.

Several approaches have already been developed to build MT, such as maximizing data likelihood [2], random drawing [3], or using a Dirichlet Process [4].

The main drawback of MTs is their reduced interpretability: it is no longer possible to graphically infer conditional independences. However, MTs can still be used to analize the structure of a problem, by counting edges between pairs of variables over all terms of the mixture [2].

3 Causal Learning of Mixtures of Trees

Two widespread frameworks exists for learning BNs from data: optimization and causal learning. The main idea behind causal learning is to build a model by exploiting independence relations between variables. These methods usually start by detecting independence relationships that seem to be satisfied by the data set (for example by using a χ^2 hypothesis test), and to derive from them the absence (or the presence) of an edge in the graph, and its orientation.

The idea behind the algorithm presented in this work is to build a MT that closely represents first-order independence relations between variables, in the sense that the MT can be interpreted as proposed by [2] to recover them: the influence between two variables is deemed proportional to the number of edges linking these two variables in the whole MT.

We decided to use the mutual information (MI) to measure the strength of the relations between any two variables i, j. Thus, the number of edges between i and j ($N_{i,j}$) over the MT should be proportional to the MI between i and j. To any edge we therefore associate a fictious weight (denoted by I_{edge}, and identical for all edges) corresponding to the portion of MI it represents. Since the goal of the model is to represent perfectly all the pair-wise interactions between variables, the sum of all edge-weights provided by the MT should be equal (to a factor of proportionality, here chosen equal to 1) to the sum of all pair-wise mutual informations between variables. Since a MT of size m over n variables has exactly $m * (n - 1)$ edges, we have that:

$$I_{\text{edge}} = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} I_{i,j}}{m * (n - 1)}, \tag{1}$$

and knowing I_{edge}, it is possible to compute the target value of $N_{i,j}$ by:

$$N_{i,j} * I_{\text{edge}} = I_{i,j}. \tag{2}$$

However, this number is rarely an integer, and the MT can only be an approximation of the relations between variables.

The last challenge is to distribute the edges between the terms of the MT, ensuring the formation of trees. To do so, we decided to build one tree at a time, starting by an empty tree, and to recursively add the edge with the highest corresponding $N_{i,j}$ (then decreased by 1) while retaining the tree structure.

4 Results

To assess the interest of this approach, we tested our algorithm according to the methodology proposed in [3]. We randomly generated 10 BNs over 16 binary variables and of bounded in-degree 5. For each of these BNs, 10 data sets of 50 observations were generated. Results given below are averaged over these data sets and BNs.

Our algorithm was applied to these data sets for growing values of m, and outputs were evaluated by computing the Kullback-Leibler divergence (KLD) to the data generating distribution, while using Laplace estimates of the conditional probability tables for each tree in each MT model. The KLD is a non-symetric measure of the difference between two probability distributions. These values were compared against two references for every data set: the data generating BN whose parameters had been reestimated from the data set, and the MT of size one, which is, by construction, the MWST.

The main result of our simulations, summarized in Tab. 1, is that our approach outperforms both reference models. Another important observation is that our method does not seem to overfit the data.

	BN rel.	MT1	MT10	MT100	MT200
KLD	6.08	1.33	1.11	1.1	1.1

Table 1. The Kullback-Leibler divergence (KLD) to target BN of different MTm and of 2 references : the relearned structure (BN rel.) and the MWST (MT1). Results are averaged on 10 BN times 10 sets of 50 samples.

5 Conclusion

The algorithm proposed is a first attempt to learn a MT by a causal approach, and the results obtained are promising, although further work is needed to test our approach on real applications, and to compare it to other algorithms.

Based on those results, several research directions can be proposed. A first one would be to seek to approximate more than first-order independence relationships. A second line of research would be to investigate different approaches to the distribution of the edges among the different trees.

6 Acknowledgments

This work was funded by the Belgian Fund for Research in Industry and Agriculture (FRIA).

References

[1] G. F. Cooper, The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

[2] M. Meila and M. I. Jordan Learning with mixtures of trees. *J. Mach. Learn. Res.*, 1:1–48, 2001.

[3] S. Ammar, P. Leray, B. Defourny and L. Wehenkel, High-dimensional probability density estimation with randomized ensembles of tree structured bayesian networks, in Z. Ghahramani (eds.), *Proceedings of the fourth European Workshop on Probabilistic Graphical Models (PGM08)*, Hirtshals, Denmark, pp. 9-16, 2008.

[4] S. Kirshner and P. Smyth, Infinite mixtures of trees, in M. Jaeger and T. D. Nielsen (eds.), *Proceedings of the 24th international conference on Machine learning (ICML '07)*, Corvalis, Oregon, pp. 417-423, 2007.