International Journal of Approximate Reasoning 88 (2017) 259–281

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Towards using the chordal graph polytope in learning decomposable models ☆

Milan Studený [a,*], James Cussens [b]

[a] Institute of Information Theory and Automation of the CAS, Pod Vodárenskou věží 4, Prague, 18208, Czech Republic
[b] Dept of Computer Science & York Centre for Complex Systems Analysis, University of York, Deramore Lane, York, YO10 5GH, United Kingdom

ARTICLE INFO

Article history:
Received 22 November 2016
Received in revised form 30 May 2017
Accepted 1 June 2017
Available online 8 June 2017

Keywords:
Learning decomposable models
Integer linear programming
Characteristic imset
Chordal graph polytope
Clutter inequalities
Separation problem

ABSTRACT

The motivation for this paper is the *integer linear programming* approach to learning the structure of a *decomposable graphical model*. We have chosen to represent decomposable models by means of special zero–one vectors, named *characteristic imsets*. Our approach leads to the study of a special polytope, defined as the convex hull of all characteristic imsets for chordal graphs, named the *chordal graph polytope*. In this theoretical paper, we introduce a class of *clutter inequalities* (valid for the vectors in the polytope) and show that all of them are facet-defining for the polytope. We dare to conjecture that they lead to a complete polyhedral description of the polytope. Finally, we propose a linear programming method to solve the *separation problem* with these inequalities for the use in a cutting plane approach.

1. Introduction: explaining the motivation

Decomposable models are fundamental probabilistic graphical models [16]. A well-known fact is that elegant mathematical properties of these structural models form the theoretical basis of the famous method of local computation [6]. Decomposable models, which are described by *chordal undirected graphs*, can be viewed as special cases of Bayesian network models [19], which are described by directed acyclic graphs.

Two traditionally separate disciplines in probabilistic graphical models are learning and inference. *Structure learning* is determining the graphical model, represented by a graph, on the basis of observed statistical data. *Inference* in Bayesian network models has two phases. The first one is transformation of the (learned) directed acyclic graph into a *junction tree*, which can be viewed as a representative of a decomposable model. The second phase in inference is proper local computation (of conditional probabilities) in a junction tree. The motivation for the present paper is the idea to merge structural learning with the junction tree construction in one step, which basically means direct learning the *structure of a decomposable model* on basis of data.

There are various methods for learning decomposable model structure, most of them being specializations of the methods for learning Bayesian network structure [18]. There are methods based on statistical conditional independence tests like the PC algorithm [23] or MCMC simulations [11]. This particular paper deals with a *score-based approach*, where the task is to maximize some additively decomposable score, like the BIC score [21] or the BDeu score [12]. There are some arguments in

☆ This paper is part of the Virtual special issue on the Eighth International Conference on Probabilistic Graphical Models, Edited by Giorgio Corani, Alessandro Antonucci, Cassio De Campos.

* Corresponding author.
E-mail addresses: studeny@utia.cas.cz (M. Studený), james.cussens@york.ac.uk (J. Cussens).

http://dx.doi.org/10.1016/j.ijar.2017.06.001
0888-613X/

favour of this approach in comparison with the methods based on statistical tests. Specifically, the study in [29] indicates that some classes of domain models cannot be learned by procedures that modify the graph structure by one edge at a time.

We are interested in the *integer linear programming* (ILP) approach to structural learning of decomposable models. The idea behind this approach is to encode graphical models by certain vectors with integer components in such a way that the usual scores become linear or affine functions of the vector representatives. There are several ways to encode Bayesian network models; the most successful one seems to be to encode them by *family-variable* vectors as used in [14,7,1]. On the other hand, since the present paper deals with learning decomposable models we have intentionally chosen to encode them by different vector representatives, called *characteristic imsets*; these vectors have also been applied in the context of learning Bayesian network models in [13] and [26]. This mode of representation leads to an elegant and unique way of encoding decomposable models which we believe is particularly suitable for structure learning of these models.

Let us note that two recent conference papers have also been devoted to ILP-based learning of decomposable models, but they use different binary encodings of the models. More specifically, Sesh Kumar and Bach [22] used special codes for junction trees of the graphs, while Pérez et al. [20] encoded certain special coarsenings of maximal hyper-trees. Moreover, the goal in both these papers was learning a specifically restricted class of decomposable models (namely, all cliques have the same prescribed size and the same holds for separators) unlike in this theoretical paper, which we hope to be the first step towards a general ILP method for learning decomposable models.

Two other recent papers devoted to structure learning of decomposable models also used encodings of junction trees. Corander et al. [5] expressed the search space in terms of logical constraints and used constraint satisfaction solvers. Even better running times have been achieved by Kangas et al. [15], who applied the idea of decomposing junction trees into subtrees, which allowed them to use the method of dynamic programming. We note that the junction tree representation is closely related to the (superset) Möbius inversion of the characteristic imset we mention in Section 6.1.

Our approach leads to the study of the geometry of a polytope defined as the convex hull of all characteristic imsets for chordal graphs (over a fixed set of nodes N), with the possible modification that a clique size limit is given. This polytope has already been dealt with by Lindner [17] in her thesis, where she derived some basic observations on the polytope. For example, she mentioned that a complete facet description of the polytope with cliques size limit two, which corresponds to learning *undirected forests*, can be derived. She also identified some non-trivial inequalities for the polytope with no clique size limit. Being inspired by Lindner we name this polytope the "chordal graph characteristic imset polytope", but abbreviate this to the *chordal graph polytope*.

In this paper, which is an extended version of a proceedings paper [25], we assume that the reader is familiar with basic concepts of polyhedral geometry, as presented in numerous textbooks on this topic; for example in [2] or [28]. We present a complete facet description of the polytope where $|N| \leq 4$ and mention the case $|N| = 5$, where the facet description is also available. We have succeeded in classifying all facet-defining inequalities for this polytope in these cases. What we found out is that, with the exception of a natural *lower bound inequality*, there is a one-to-one correspondence between the facet-defining inequalities and the *clutters* (alternatively named *antichains* or *Sperner families*) of subsets of the variable set N (i.e. of the set of nodes) containing at least one singleton; so we call these *clutter inequalities*.

This establishes a sensible *conjecture* about a complete polyhedral description of the polytope (with no clique size limit). We prove that every clutter inequality is both valid and facet-defining for the polytope. We also tackle an important *separation problem*: that is, given a non-integer solution to a linear programming (LP) relaxation problem, find a clutter inequality which (most) violates the current solution.

The structure of the paper

Basic concepts are recalled in Section 2, where the concept of a chordal graph polytope is introduced. Section 3 reports on the facet description of this polytope in the case of at most five nodes, which was found computationally. The clutter inequalities are defined and illustrated by a couple of simple examples in Section 4. Our completeness conjecture is then formulated in Section 5; various other versions of clutter inequalities are given in Section 6. In Section 7 we present the idea of the proof of their validity for any vector in the chordal graph polytope. The main result of the paper saying that every clutter inequality is facet-defining for the polytope is presented in Section 8. Section 9 is devoted to a sub-problem of finding a suitable clutter inequality in the context of the cutting plane method. A brief report on a small preliminary empirical study is given in Section 10. Important open tasks are recalled in Conclusions, which is Section 11. The proofs of most observations have been put in the Appendix to make the paper smoothly readable.

2. Basic concepts

Let N be a finite set of *variables*; assume that $n := |N| \geq 2$ to avoid the trivial case. In the statistical context, the elements of N correspond to *random variables*, while in the graphical context they correspond to *nodes* of graphs.

2.1. Some conventional notation and terminology

The symbol $\subseteq$ will be used to denote non-strict set inclusion of unlike $\subset$, which will serve to denote strict inclusion: $S \subset T$ means $S \subseteq T$ and $S \neq T$. The *power set* of N will be denoted by $\mathcal{P}(N) := \{S : S \subseteq N\}$.