Integer linear programming approach to learning Bayesian network structure: towards the essential graph

Milan Studený

Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

The Sixth European Workshop on *Probabilistic Graphical Models*
Granada, Spain, September 19, 2012, 14:50-15:10

Summary of the talk

1. Introduction: learning Bayesian network structure
2. Preliminaries: essential graph
3. Linear-algebraic approach to learning: characteristic imset
4. Other (integer) linear programming approaches
5. An extended characteristic imset
6. Conclusion: invitation to the poster

Introduction: learning BN structure by score-maximization

Bayesian networks (BN) are special graphical models widely used both in artificial intelligence and in statistics. They are described by *acyclic directed graphs*, whose nodes correspond to (random) variables.

The motivation for the research reported here is learning a BN structure from data by maximizing a quality criterion.

By a *quality criterion*, also called a *score*, is meant a real function of the BN structure (= of a graph G, usually) and of the database D.

The value $\mathcal{Q}(G, D)$ should say how much the BN structure given by G is good to explain the occurrence of the database D.

The aim is to maximize $G \mapsto \mathcal{Q}(G, D)$ given the observed database D.

Examples of such criteria are *maximized log-likelihood (MLL)* criterion, Schwarz's *BIC criterion* and *Bayesian Dirichlet Equivalence (BDE)* score.

Preliminaries: assumptions on quality criteria

There are two important technical assumptions on a quality criterion $\mathcal{Q}$ brought in connection with the maximization problem.

The first assumption is that $\mathcal{Q}$ is *score equivalent*, which means it ascribes the same value to *Markov equivalent graphs* (= graphs defining the same BN structure).

R. R. Bouckaert (1995). Bayesian belief networks: from construction to evidence. PhD thesis, University of Utrecht.

The other assumption is that $\mathcal{Q}$ is (additively) *decomposable*, which means $\mathcal{Q}(G, D)$ is the sum of contributions that correspond to the factors in the factorization according to the graph G.

D. M. Chickering (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3**:507-554.

Preliminaries: graphical representative

A classic graphical characterization of equivalent graphs states that they are Markov equivalent iff they have the same *adjacencies* and *immoralities*, which are special induced subgraphs.

T. Verma and J. Pearl (1991). Equivalence and synthesis of causal models. In *6th Conference on Uncertainty in Artificial Intelligence*, pages 220-227.

Researchers calling for methodological simplification proposed to use a unique representative for each individual BN structure. The classic unique graphical representative is the *essential graph*.

S. A. Andersson, D. Madigan and M.D. Perlman (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* **25**:505-541.

Preliminaries: essential graph

Definition

Let $\mathcal{G}$ be a Markov equivalence class of acyclic directed graphs over N. The *essential graph* G^* of $\mathcal{G}$ is defined as follows:

- $a \to b$ in G^* if $a \to b$ in every G from $\mathcal{G}$,
- $a - b$ in G^* if there are graphs G_1 and G_2 in $\mathcal{G}$ with $a \to b$ in G_1 and $a \leftarrow b$ in G_2.

M. Studený (2004). Characterization of essential graphs by means of the operation of legal merging of components. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **12**:43-62.

Lemma

Let $\mathcal{G}$ be an equivalence class of acyclic directed graphs over N and $\mathcal{H}$ an equivalence class of chain graphs without flags such that $\mathcal{G} \subseteq \mathcal{H}$. Then G^* is the largest graph in $\mathcal{H}$.

Linear-algebraic approach: characteristic imset

The basic idea of a linear-algebraic approach is to represent the BN structure given by an acyclic directed graph G by a certain vector.

In last PGM, we have proposed a special zero-one vector to represent uniquely BN structures.

M. Studený, R. Hemmecke and S. Lindner (2010). Characteristic imset: a simple algebraic representative of a Bayesian network structure. In the *5th European Workshop on Probabilistic Graphical Models*, pages 257–264.

Definition (equivalent, not the original one)

Assume $|N| \geq 2$. Given an acyclic directed graph G over N, the *characteristic imset* for G is a zero-one vector with components indexed by subsets S of N with $|S| \geq 2$ such that

$$c_G(S) = 1 \qquad \text{there exists } i \in S \text{ with } S \setminus \{i\} \subseteq pa_G(i).$$

Characteristic imset: properties

Observation

Two acyclic directed graphs G and H over N are Markov equivalent if and only if $c_G = c_H$. Moreover, every score equivalent and decomposable criterion $\mathcal{Q}$ has the form

$$\mathcal{Q}(G, D) = \mathcal{Q}(G^{\emptyset}, D) + \sum_{S \subseteq N, |S| \geq 2} r_D^{\mathcal{Q}}(S) \cdot c_G(S)\,,$$

where $G^{\emptyset}$ is the empty graph over N (= without adjacencies) and $r_D^{\mathcal{Q}}$ uniquely determined real vector, depending on the database D only, called the *revised data vector* (relative to $\mathcal{Q}$).

R. Hemmecke, S. Lindner, M. Studený (2012). Characteristic imsets for learning Bayesian network structure. To appear in *International Journal of Approximate Reasoning*, see doi:10.1016/j.ijar.2012.04.001.

Characteristic imset and graphical description

The characteristic imset is close to the graphical description:

Corollary

Let G be an acyclic directed graph over N and a, b (and c) are distinct nodes. Then

(i) *a and b are adjacent in G iff $c_G(\{a,b\}) = 1$.*

(ii) *$a \to c \leftarrow b$ is an immorality in G iff $c_G(\{a,b,c\}) = 1$ and $c_G(\{a,b\}) = 0$.*

In particular, one can observe that the characteristic imset c_G is uniquely determined by its values $c_G(S)$ for $S \subseteq N$, $2 \leq |S| \leq 3$.

WARNING: However, the remaining values are **do not depend linearly** on the values $c_G(S)$ for $S \subseteq N$, $2 \leq |S| \leq 3$.

The characteristic imset and the essential graph

In fact, there is a direct formula for the characteristic imset on the basis of the essential graph.

Theorem

Let H be a chain graph without flags equivalent to an acyclic directed graph G. For any $S \subseteq N$, $|S| \geq 2$ one has $c_G(S) = 1$ iff

$$\exists \emptyset \neq K \subseteq S \text{ line-complete in } H, \text{ with } j \to i \text{ for any } j \in S \setminus K,\ i \in K.$$

The proof can also be found in (Hemmecke, Lindner, Studený 2012).

Other LP approaches: straightforward codes of graphs

Definition

An acyclic directed graph G over N can be also encoded by a vector η_G, whose components indexed by pairs $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$:

$$\eta_G(i|B) = \begin{cases} 1 & B = pa_G(i), \\ 0 & \text{otherwise.} \end{cases}$$

T. Jaakkola, D. Sontag, A. Globerson, M. Meila (2010). Learning Bayesian network structure using LP relaxations. In *JMLR Workshop and Conference Proceedings, volume 9: AISTATS*, pages 358–365.

They characterized the η_G-codes by means of a finite list of *linear inequalities* and, thus, turned the learning task into an ILP problem:

> *to optimize a linear function over vectors with integer components within a polyhedron.*

They even made computational experiments based on that approach.

The idea of extending the BN vector representative

J. Cussens (2010). Maximum likelihood pedigree reconstruction using integer programming. In *Workshop on Constraint Based Methods for Bioinformatics*, pages 9–19.

J. Cussens (2011). Bayesian network learning with cutting planes. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 153–160.

He was interested in pedigree learning, in which case the parent set cardinality is bounded by 2. However, to ensure the acyclicity of the encoded graph G he used another trick: the idea of *extending* the vector BN representatives.

In the other paper, Cussens (2011) was inspired by Jaakkola et al. (2010). Unrestricted BN structure learning was the goal and to overcome the problem with the exponential number of these inequalities Cussens used the *cutting plane* approach.

The idea of pruning

To overcome the technical problem with the exponential length (in $|N|$) of vectors η_G the idea of *pruning* of their components was applied.

C.P. de Campos, Z. Zeng, Q. J. (2009). Structure learning Bayesian networks using constraints. In *26th International Conference on Machine Learning*, pages 113–120.

C.P. de Campos, Q. J. (2011). Efficient structure learning Bayesian networks using constraints. *Journal of Machine Learning Research*, **12**:663–689.

A particular form of scores used in practice allows one to conclude (on the basis of an observed database D) that the optimal graph G has no node $i \in N$ with large $|pa_G(i)|$.

This pruning procedure is time demanding, but useful: in practical cases it typically results in the reduction of the parent set cardinality to at most 5, only in a few cases the maximal cardinality was 7 or 8.

An extended characteristic imset

S. Lindner (2012). Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Definition

Let H be a hybrid graph over N. We ascribe to H a zero-one vector (a_H, c_H) with components given as follows: for distinct $i, j \in N$

$$a_H(i \to j) = 1 \iff i \to j \text{ in } H,$$

and, for $S \subseteq N$, $|S| \geq 2$,

$c_H(S) = 1 \Leftrightarrow \exists \emptyset \neq K \subseteq S$ complete, with $j \to i$ for any $j \in S \setminus K$, $i \in K$.

If H is a chain graph without flags equivalent to an acyclic directed graph G, then c_H is nothing but the characteristic imset for G.

Conclusion: an invitation to the poster

The extension does not kill the comparative advantage relative to the straightforward graph codes because the extended characteristic imsets are still more than $\frac{|N|}{3}$-times shorter than the codes.

In the present paper, a list of linear inequalities is presented which characterize the (a_H, c_H)-codes for chain graphs without flags equivalent to acyclic directed graphs.

This allows one to re-formulate the BN learning task as an ILP problem, whose solution is a chain graph without flags equivalent of the optimal acyclic directed graph G.

The advantage of the presented approach is that one to get directly the *essential graph* as a result of solving a subsequent ILP problem.

If you are interested in further (technical) details,
you are welcome to discuss the topic near my poster!