ON LEARNING BAYESIAN NETWORK MODELS BY MAXIMIZATION OF A QUALITY CRITERION

MILAN STUDENÝ
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

Praha, February 8, 2005
seminar of the group "Decision Making and Control under Uncertainty"

Summary

1. Introduction - Bayesian network models
2. Structural learning
3. Quality criteria
4. Problem of representative choice
5. Local search methods
6. Inclusion neighbourhood
7. Algebraic approach: standard imsets
8. Discussion - transition between different model representatives

Introduction - Bayesian network models I

Bayesian network is a graphical model of probabilistic relationships among variables. It is described by an acyclic directed graph G which has the set of variables N as the set of nodes.

From a mathematical point of view, a statistical model is a class of probability distributions. In the considered case, it is the class of (discrete) distributions which *recursively factorize* according to G:

$$P(x) = \prod_{i \in N} P_{i|pa_G(i)}(x_i | x_{pa_G(i)}) \qquad \text{for } x \in \mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i \,,$$

where $pa_G(i) = \{ j \in N \,;\, j \to i \;\text{ in } G \}$, X_i are fixed (finite) sample spaces, and x_A, $A \subseteq N$ denotes the restriction of a configuration of values $x \in \mathsf{X}_N$.

Introduction - Bayesian network models II

A graphical model can be viewed as a special case of a statistical model of conditional independence structure.

That means, it can equivalently be introduced as the class of distributions on X_N which satisfy some conditional independence restrictions:

$A \perp\!\!\!\perp B \,|\, C \; [P]$ whenever triplet $\langle A, B|C\rangle$, $A, B, C \subseteq N$ pairwise disjoint is represented in G according to the d-separation criterion.

These distributions are called Markovian distributions with respect to G.

The definition of the graphical d-separation criterion is omitted here.

graph $\longrightarrow$ conditional independence restrictions $\longrightarrow$ statistical model

Structural learning

Learning of a statistical model is the procedure of determining the most suitable model from a considered universum of available statistical models, typically on basis of data.

The methods for learning (graphical) models of conditional independence structure can be divided into two groups:

- the methods based on *significance tests*,
- the methods based on the maximization of a suitable *quality criterion*.

Note that some methods can be classified in both groups – (Whittaker 1990). Moreover, there is a simulation method, namely MCMC, applicable to learning graphical models which does not belong to either of these two groups.

Most of the methods were developed for learning Bayesian network models.

Quality criteria

This talk deals with methods for *learning Bayesian network models* based on the *maximization of a quality criterion*.

N	non-empty finite set of variables
$\mathsf{DAGS}(N)$	the class of acyclic directed graphs over the set of nodes N
$\mathsf{DATA}(N, d)$	the set of all databases over N of the length d, $d \geq 1$ (some finite sample spaces are fixed)

Quality criterion (for learning Bayesian networks) is a real function $\mathcal{Q}$ on $\mathsf{DAGS}(N) \times \mathsf{DATA}(N, d)$.

A quality criterion should be *consistent*, but there are other reasonable mathematical requirements, which will be discussed later in the talk.

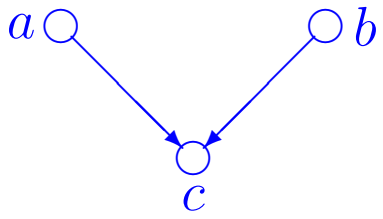
Equivalence of Bayesian networks

One statistical model can be described by different graphs. Two graphs $G, H \in \mathsf{DAGS}(N)$ are *Markov equivalent* if they define the same class of Markovian distributions.

Two graphs $G, H \in \mathsf{DAGS}(N)$ are *independence equivalent* if they define the same collection of conditional independence restrictions. Let us write $G \approx H$ then.

There exists a graphical characterization of independence equivalence (Verma and Pearl 1990): $G \approx H$ iff they have the same underlying graph and immoralities.

An *immorality* in a graph is its induced subgraph of this form:



Chickering (1995) gave an alternative transformational characterization of equivalent acyclic directed graphs.

Score-equivalent criteria

A quality criterion $\mathcal{Q}$ is *score-equivalent* iff $\forall\, G, H \in \mathsf{DAGS}\,(N), \forall\, D \in \mathsf{DATA}\,(N, d)$

$$\text{if } G \approx H \quad \text{then} \quad \mathcal{Q}(G, D) = \mathcal{Q}(H, D)\,.$$

Most of the criteria used in practice are score-equivalent:

- MLL (maximized log-likelihood criterion),
- AIC (Akaike's information criterion),
- BIC (Jeffrey-Schwarz criterion),
- some Bayesian criteria (this depends on the choice of 'priors').

Decomposable criteria

A quality criterion $\mathcal{Q}$ is *decomposable* if there exists a collection of functions $q_{i|B} : \mathsf{DATA}\,(B \cup \{i\}, d) \to \mathbf{R}$, $i \in N$, $B \subseteq N \setminus \{i\}$ such that $\forall\, G \in \mathsf{DAGS}\,(N)$, $\forall\, D \in \mathsf{DATA}\,(N, d)$

$$\mathcal{Q}(G, D) = \sum_{i \in N} q_{i|pa_G(i)}(D_{\{i\} \cup pa_G(i)})$$

where D_A denotes the restriction of a database D for $A \subseteq N$.

This technical requirement was brought by researchers in computer science in connection with the method of local search. (Heckerman 1995) (Chickering 2002)

Actually, one can distinguish two concepts of decomposability, depending on what is meant by a database. *Strong decomposability* corresponds to data in the form of a contingency table.

All criteria used in practice are (strongly) decomposable.

Problem of representative choice

How to represent a Bayesian network model in the memory of a computer?

There are two basic methods:

- represent it by arbitrary $G \in \mathsf{DAGS}\,(N)$ in the respective independence equivalence class,
- represent it by a special unique representative.

The most popular graphical representative is the essential graph. It is a chain graph G^* obtained from the equivalence class $\mathcal{G}$ as follows:

- $a \to b$ in G^* if $a \to b$ in G for every $G \in \mathcal{G}$,
- $a - b$ in G^* if there are $G_1, G_2 \in \mathcal{G}$ such that $a \to b$ in G_1 and $a \leftarrow b$ in G_2.

Later, another (algebraic) representative, called the *standard imset*, will be introduced.

Local search methods

Direct maximization of a quality criterion is typically infeasible. To avoid this problem various heuristic *local search methods* were developed.

The basic idea is that one introduces a *neighbourhood structure* in the set $\mathsf{DAGS}(N)$, respectively in the set $\mathsf{DAGS}(N)/_{\approx}$. Instead of the global maximum of $\mathcal{Q}$ one is trying to find a *local maximum* with respect to that neighbourhood structure.

Every graph G (respectively every equivalence class $\mathcal{G}$) is assigned a relatively small set of neighbouring graphs (respectively equivalence classes) $nei(G)$. They typically differ in the presence of one edge.

The point is that for a decomposable criterion $\mathcal{Q}$ the difference $\mathcal{Q}(G, D) - \mathcal{Q}(H, D)$ for neighbouring graphs $G, H \in \mathsf{DAGS}(N)$ is easy to compute.

Inclusion neighbourhood

Is there a natural neighbourhood for $\mathsf{DAGS}\,(N)/_{\approx}$?

Let $\mathcal{I}(G)$ be the set of conditonal independence restrictions given by $G \in \mathsf{DAGS}\,(N)$. Given $K, L \in \mathsf{DAGS}\,(N)$, $\mathcal{I}(K) \subset \mathcal{I}(L)$ means $\mathcal{I}(K) \subseteq \mathcal{I}(L)$ but $\mathcal{I}(K) \neq \mathcal{I}(L)$.

If $\mathcal{I}(K) \subset \mathcal{I}(L)$ and there is no $G \in \mathsf{DAGS}\,(N)$ such that $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$ then we will say that $\mathcal{I}(L)$ is an *upper inclusion neighbour* of $\mathcal{I}(K)$, respectively $\mathcal{I}(K)$ is a *lower inclusion neighbour* of $\mathcal{I}(L)$. We will then write $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$.

The inclusion neighbourhood was characterized in graphical terms.
This is a consequence of the validity of Meek's (1995) conjecture, confirmed by Chickering (2002). One has $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ iff there exists $K', L' \in \mathsf{DAGS}\,(N)$, $K' \approx K$, $L' \approx L$ such that L' is obtained from K' by the removal of one arrow.

There are some arguments why general neighbourhood structure in a local search method should involve the inclusion neighbourhood. (Castelo 2002)

Algebraic approach

The basic idea is to describe a Bayesian network model by a certain integral vector. This is motivated by a more general method for describing probabilistic conditional independence structures (Studený 2005).

By an *imset over N*, an integer-valued function on the power set of N is understood.

Given $A \subseteq N$, δ_A is the identifier of the set A.
Given conditional independence statement $a \perp\!\!\!\perp b \,|\, C$, where $a, b \in N$, $a \neq b$, $C \subseteq N \setminus \{a, b\}$ the respective *elementary imset* has the form

$$u_{\langle a,b|C\rangle} = \delta_{\{a,b\}\cup C} + \delta_C - \delta_{\{a\}\cup C} - \delta_{\{b\}\cup C} \,.$$

By a *structural imset*, an imset which is a combination of elementary imsets with non-negative rational coefficients is meant.

Standard Imsets I

Let $G \in \mathsf{DAGS}\,(N)$. Then the respective *standard imset* has the form

$$u_G = \delta_N - \delta_\emptyset + \sum_{i \in N} \{\delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)}\}\,.$$

As it has many 'zeros' it can be easily kept in the memory of a computer.

Proposition Given $G, H \in \mathsf{DAGS}\,(N)$ one has $G \approx H$ iff $u_G = u_H$.

Thus, the standard imset can serve as a unique representative of the respective Bayesian network model.

Proposition Given $K, L \in \mathsf{DAGS}\,(N)$ one has $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ iff $u_L - u_K$ is an elementary imset.

Standard Imsets II

Proposition Let $\mathcal{Q}$ be a score-equivalent decomposable criterion. Then it has the following form:
$\forall\, G \in \mathsf{DAGS}\,(N)$, $\forall\, D \in \mathsf{DATA}\,(N, d)$

$$\begin{aligned}\mathcal{Q}(G, D) &= k_{\mathcal{Q}}(D) + \sum_{S \subseteq N} u_G(S) \cdot t_D^{\mathcal{Q}}(S) \\ &= k_{\mathcal{Q}}(D) + \langle u_G, t_D^{\mathcal{Q}} \rangle\,.\end{aligned}$$

where $t_D^{\mathcal{Q}}$ is a real vector representing the database D (relative to $\mathcal{Q}$) and $k_{\mathcal{Q}}(D)$ is a constant (depending on data).

In particular, $\mathcal{Q}(L, D) - \mathcal{Q}(K, D) = \langle u_L - u_K, t_D^{\mathcal{Q}} \rangle$.

Thus, from purely mathematical point of view, it leads to the problem of maximization of a (shifted) linear function.

Transition between graphical and algebraic representatives

To utilize fully the algebraic approach to the local search methods one has to be able to describe the inclusion neighbourhood in terms of the standard imset.

There exists characterization of inclusion neighbourhood of a given $\mathcal{G} \in \mathsf{DAGS}\,(N)/_{\approx}$ in terms of the respective essential graph (Studený 2004).

For this reason, it is desirable to translate graphical representatives into algebraic ones and conversely. There exists a formula which gives the standard imset on the basis of the essential graph and a two-stage inverse algorithm for getting the essential graph on basis of the standard imset (Studený Vomlel 2004).

References

(Castelo 2002) The discrete acyclic digraph Markov model in data mining, PhD thesis, University of Utrecht.

(Chickering 1995) A transformational characterization of equivalent Bayesian network structures, in Uncertainty in Artificial Intelligence 11 (P.Besnard, S.Hanks eds.), Morgan Kaufmann, 87-98.

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

(Heckerman 1995) A tutorial on learning Bayesian networks, technical report MSR-TR-95-06, Microsoft Research, Redmond.

(Meek 1995) Graphical models, selecting causal and statistical models, PhD thesis, Carnegie Melon University.

(Studený 2004) Characterization of inclusion neighbourhood in terms of the essential graph, *International Journal of Approximate Reasoning 38*, 283-309.

(Studený Vomlel 2004) Transition between graphical and algebraic representatives of Bayesian network models, in Proceedings of the 2nd European Workshop on Probabilistic Graphical Models (P.Lucas ed.), University of Nijmegen, 193-200.

(Studený 2005) *Probabilistic Conditional Independence Structures*, Springer Verlag.

(Verma Pearl 1990) Causal networks, semantics and expressiveness, in Uncertainty in Artificial Intelligence 4 (R.D.Shachter, T.S.Lewitt, L.N.Kanal, J.F.Lemmer eds.), North Holland, 69-76.

(Whittaker 1990) *Graphical Models in Applied Multivariate Statistics*, John Wiley.

Derivation of quality criteria I

Let $\mathbf{M} = \{P_\theta\,;\, \theta \in \Theta\}$ be a statistical model; each P_θ is a distribution on X_N.

Likelihood function (likelihood = the probability of the occurence of a database)

$$(\theta, D) \longrightarrow L(\theta, D) \equiv \prod_{\ell=1}^{d} P_\theta(x^\ell), \quad \theta \in \Theta\,,\ D = (x^1, \ldots, x^d) \in (\mathsf{X}_N)^d$$

Maximized log-likelihood is then

$$\mathsf{MLL}(\mathbf{M}, D) \;=\; \max\{\ln L(\theta, D);\ \theta \in \Theta\}\,.$$

The complexity of a model is measured by its effective dimension $\mathsf{DIM}(\mathbf{M})$, which is the affine dimension of Θ (= the number of free parameters).

Derivation of quality criteria II

Let $\mathbf{M}$ be the class of Markovian distributions with respect to a graph G over N:

$$\mathbf{M}_G = \{P_\theta\,;\,\theta \in \Theta_G\}\,.$$

$$\begin{aligned}
\mathsf{MLL}(G,D) &= \mathsf{MLL}(\mathbf{M}_G, D)\\
\mathsf{AIC}(G,D) &= \mathsf{MLL}(\mathbf{M}_G, D) - \mathsf{DIM}(\mathbf{M}_G)\\
\mathsf{BIC}(G,D) &= \mathsf{MLL}(\mathbf{M}_G, D) - \tfrac{1}{2}\ln(d)\cdot\mathsf{DIM}(\mathbf{M}_G)
\end{aligned}$$

Bayesian criteria: One has a prior probability measure π_G on Θ_G. The respective criterion is the logarithm of the marginal likelihood

$$\mathsf{LML}(G,D) \;=\; \ln \int_{\Theta_G} L(\theta, D)\; d\pi_G(\theta)\,.$$

Of course, the priors have to satisfy additional technical assumptions; sometimes a prior on the space of graphs is also considered.