A Probabilistic Knowledge Base Using Annotated Bayesian Network Features

Péter Antal, András Millinghoffer

Department of Measurement and Information Systems
Budapest University of Technology and Economics
antal@mit.bme.hu, milli@mit.bme.hu

Abstract: The probabilistic modeling of a high dimensional domain includes the modeling of the joint distribution over the domain variables on numeric, qualitative and possibly causal levels. Additionally, it includes the combination of statistical data with domain knowledge acquired from experts and the usage of the result in a decision theoretic framework. We overview the Bayesian network representation and the Bayesian statistical framework, which are successfully applied tools for these challenges. We consider a special learning problem related to Bayesian networks, the Monte Carlo, particularly the Markov Chain Monte Carlo (MCMC) estimation of the posteriors of structural properties i.e. features. We introduce a special classification oriented feature called Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph and propose a specialized order-MCMC method for its estimation. Finally, we discuss the composition of and Bayesian inference about complex statements including structural properties of Bayesian networks and textual annotations corresponding to the network elements, which extension is proposed in a wider context as a step towards probabilistic first-order logic. The proposed methods are demonstrated in a biomedical field related to the diagnosis of ovarian cancer.

Keywords: statistical inference, Bayesian networks, feature learning, data mining

1 Introduction

In the paper we overview the Bayesian network representation and its frequent corresponding statistical framework, the Bayesian statistics. In Section 4 we discuss the Bayesian inference over structural properties and introduce a classification oriented feature called Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph. In Section 5 we discuss the composition of complex statements including structural properties of Bayesian networks and textual annotations corresponding to the network elements. Section 6 contains examples of the applications of the proposed methods in a biomedical field related to the diagnosis of ovarian cancer.

2 Bayesian Statistics and Models

The basic goal of Bayesian statistics is to provide the axiomatic foundations of inference based on observations and background knowledge, the uncertainty of which may result from e.g. the method of knowledge acquisition and data collecting, or the lack or ignorance of knowledge.

For handling uncertainty, the probabilistic framework is used, accepting the subjectivist interpretation, regarding probabilities as measures of our prior beliefs in the happening of events. It can be proven [2], that in a decision problem we can order a positive real number (probability) and a utility value to every event, so representing our preferences exactly, so that rational decisions can be met.

Representational theorems [2] state that any sequence of random variables can be generated by a proper sampling model class and a distribution over it, if the sequence satisfies the exchangeability property.

Based on the subjectivist interpretation and the existence of the above model classes, the framework of Bayesian statistics can be proposed. Here observation data is considered to be generated by ensembles of models parameterized by random variables. In practice, this parameterization is organized hierarchically: the model space consists of discrete elements (structures), to which numeric parameters belong.

2.1 Inference

The task of inference is to estimate the probability of an event (predictive inference) or of a model (parametric inference), conditioned on observation data and prior knowledge. From Bayes' theorem, we can derive the following equation for parametric inference (in the followings 'D', 'G' and 'θ' denote data, structure and parameterization, respectively):

$$P(G,\theta \mid D) = \frac{P(D \mid G,\theta)P(G,\theta)}{P(D)} \tag{1}$$

The posterior of a given structure can be calculated by integrating out θ from equation (1). In predictive inference the probability in question is calculated for every possible structure and parameterization, then these quantities are averaged using the posterior probabilities of models as weights:

$$p(x \mid D) = \sum_k p(G_k \mid D) \int p(x \mid \theta_k) p(\theta_k \mid G_k, D) d\theta_k \tag{2}$$

As the exact calculation of the above integrals and summations are usually intractable in practice, often some approximation is used, typically one of the Monte Carlo methods. We discuss these methods in detail in Section 4.

2.2 Advantages of Bayesian Approach

Opposing to its computational complexity, Bayesian methods have many advantages over classical statistics. First, by the exclusive use of probabilities for representing uncertainty of parameters, we get an efficient tool, supporting automatized methods, like learning or knowledge base building. Second, prior distributions can represent our prior knowledge (or even the total lack of it), and subsequently gained priors can be regarded as the phases of the knowledge acquisition process. Third, Bayes' theorem can combine normatively the information of prior knowledge and data, and formalizes exactly that inversion, which we need for parameter estimation. Fourth, using posterior distributions instead of point estimations considers not only the most likely configuration but the less likely ones as well. For a more complete overview, see [15].

3 Bayesian Networks

Nowadays, Bayesian networks are used primarily for the probabilistic modeling of domains without systematic structure (as against e.g. picture or sound processing). Bayesian networks store variables and their relations in a directed acyclic graph (DAG): each node symbolizes a variable, and a local conditional dependence model belonging to every node describes the connections.

As a representational tool, a Bayesian network can be interpreted in three different ways, having stronger and stronger semantics in the order of the enumeration. It can be regarded as (1) an effective tool for representing joint distributions, (2) a map of probabilistic independences and dependences of the domain, or (3) a causal domain model in which edges are direct cause-effect connections.

3.1 Probabilistic Definition: Syntax and Semantics

The connection between a structure of a Bayesian network and the represented distribution can be based on the following, equivalent conditions [4]:

(1) The distribution '$P(X_1,..X_n)$' is Markov relative to DAG 'G' or factorizes w.r.t. 'G', if $P(X_1,..X_n) = \prod P(X_i \mid Pa(X_i))$, where '$Pa(X_i)$' denotes the parents of 'X_i'.

(2) The distribution '$P(X_1,..X_n)$' obeys the ordered Markov condition w.r.t. DAG 'G', if $\forall i = 1..n : I(X_{\pi(i)} \mid Pa(X_{\pi(i)}) \mid \{X_{\pi(j)} \mid j < i\} \setminus Pa(X_{\pi(i)}))_P$, where '$\pi$' is an ancestral ordering w.r.t. 'G'.

(3) The distribution '$P(X_1,..X_n)$' obeys the local (parental) Markov condition w.r.t. DAG 'G', if $\forall i = 1..n : I(X_i \mid Pa(X_i) \mid Nondescendants(X_i))_P$, where 'Nondescendants($X_i$)' denotes the nondescendants of 'X_i' in 'G'.

(4) The distribution 'P' obeys the global Markov condition w.r.t. DAG 'G', if $\forall x, y, z \subseteq \{X_i\} : I(x \mid z \mid y)_G \Rightarrow I(x \mid z \mid y)_P$, where 'I(X|Y|Z)' denotes that 'X' and 'Y' are d-separated[1] by 'Z'.

A definition based on the properties of the dependence system given by the Markov conditions could be [14]: The 'G' DAG is the Bayesian network of the 'P(U)' distribution if (1) every variable '$u \in U$' is represented by a node in the graph, and (2) the graph is minimal.

However, while this definition regards the network as the representation of the dependence system, there is another, more practical definition. The '(G, θ)' pair is the Bayesian network of the 'P(U)' distribution if (1) G is a DAG in which nodes represent the elements of 'U' , (2) 'θ' is the whole of the numeric parameters describing the 'P(X|Pa(X))' conditional distributions belonging to the nodes.

Although Bayesian networks may contain continuous variables as well, in the followings we consider only ones with discrete finite variables, supposing that the local dependence models are multinomial distributions (i.e. they can be represented by a conditional probability table – CPT).

The structure of a given Bayesian network determines, what dependences it can describe, however the same dependence model may belong to different structures as well. If two structures imply the same dependence system, they are called observationally equivalent. By the aid of observational equivalence, the structures can be ordered into disjunctive classes. Each of these equivalence classes can be represented by a partially directed acyclic graph (PDAG), the so-called essential graph. The undirected skeleton of the essential graph is the same as the ones' belonging to the given class, and only those edges are directed which has the same direction in every member (these are the so-called compelled edges).

3.2 Causal Definition

It is formally easy define causal Bayesian networks on the basis of the previous definitions: the pair '(G, θ)' is the causal Bayesian network of the distribution 'P(U)', if (1) it is the probabilistic model of the domain according to the previous interpretations, (2) and every edge represents a direct cause-effect connection. The "only" difference is that edges represent here direct causal connections. Although causal Bayesian networks have very strong semantics, confounding factors may prevent us from regarding all of the probabilistic dependences as causal connections. Such can be the presence of common ancestors, the inadequacies of data acquisition method, the DAG representation, or the granularity of the domain.

---

[1] The sets X and Y are d-separated in the DAG G by Z, if Z blocks every non-directed path between them, i.e. (1) the path contains a node with non-converging edges that is in Z, or (2) the path contains a node with converging edges so that neither the node nor any of its children is in Z.

3.3 Inference and Learning

In a given Bayesian network the task of inference is to compute the 'P(X=x|Y=y,G,θ)' quantity, i.e. a structure, its parameterization, and the instantiation of some so-called evidence variables (Y) are give, question is the probability of an instantiation of the query variables. In order to calculate the quantity 'P(X=x|Y=y)', i.e. to make real Bayesian prediction, the summations and integrals of Eq. 2 must be computed. Inference in a given network is NP-complete [8]: because of this, either Monte Carlo estimations or so-called junction tree algorithms are used, which perform well in practice.

Learning of Bayesian networks can be regarded as a special case of parametric inference: we seek the structure and, or parameterization with the highest posterior probability. Learning is useful when we cannot make full Bayesian inference (i.e. we cannot consider multiple models), but have a considerable amount of observation data. Learning can be used instead of or beside manual construction, and the score function (the posterior probability) is often altered, usually by some penalties in order to represent our prior expectations. For an overview, see [9].

4 Feature Learning

The problem of learning structural properties (features) of Bayesian networks was proposed and investigated in a series of papers, using the bootstrap methodology and the Bayesian approach [5] as well. In the Bayesian approach the posterior of a given structural feature F with finite, discrete values $f_0,\ldots,f_R$ is defined by the following equation

$$P(F(G) = \mathrm{f}_i \mid D) = \sum_{G:F(G)=\mathrm{f}_i} P(G \mid D) = \sum_{G} I\{F(G) = \mathrm{f}_i\} P(G \mid D) \tag{3}$$

This expectation involves a "model averaging" in the space of DAG models, which computation frequently occurs in Bayesian learning of Bayesian networks or in the full-scale Bayesian predictive inference

However, this summation is usually not tractable due to the super-exponential cardinality of the space of DAGs [2], which is still the case even if the maximum size of parental sets is bounded [5].

A standard approach is to use MCMC methods over the space of DAGs, particularly to use the Metropolis algorithm to specify an irreducible and aperiodic Markov chain [13]. In finite spaces stochastic accessibility and acyclicity ensure the existence of an equilibrium (limiting) distribution, and the applicability of the Markov chain analogues of the laws of large numbers and the central limit theorem. To ensure faster convergence to the limiting distribution and better properties for the estimations of the target expectations this paper suggested the

use of orderings over the variables in the proposal distribution over DAGs. This idea was further developed by noting that a polynomial-time complexity analytic formula can be derived for the posterior probability $p(<|D)$ of an ordering $\triangleright\rangle$ and for the order conditional probability of certain features, if the number of parents are bounded [5].

The availability of the posterior for the orderings and for the conditional posteriors for features allows hybrid approaches to estimate the unconditional posteriors for features by averaging over the space of ordering using various Monte Carlo methods. In case of the so called order-MCMC methods this results in better MCMC properties, such as faster convergence and Monte Carlo variance [5].

We extend this method by introducing a new structural Bayesian network feature called Markov Blanket Graph or Mechanism Boundary Graph, which represents the relevant variables and the dependences corresponding to a given variable Y (MBG(Y)). A subgraph of G is called the Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph MBG(Y,G) of variable Y if it includes the nodes in the Markov blanket of Y (i.e. its parents, children and the other parents of its children) and the corresponding edges.

It is easy to show that the classification performance of a Bayesian network in case of complete data is fully determined by the Markov blanket graph and its parameterization. Consequently, the Markov blanket graph is a necessary and sufficient representation to identify not only the relevant subset of the variables, but their interactions as well (in case of complete data). Another interpretation of the MBG feature is that it encompasses all the causal mechanisms directly related to a given variable. These properties ensure the central importance of this complex feature particularly in classification context, because it is located on a practically relevant middle-layer between simple features and complete domain models.

The cardinality of the MBG(Y) space in case of n variables is still super-exponential (even if the number of parents are bounded above with k). Consider an ordering of the variables such that Y is the first and all the other variables are children of it, then the parental sets can be selected independently. However, at the other extreme, if Y is the last in the ordering, then the number of alternatives (i.e. parental sets) is in the order of 2^{n-1} or $(n-1)^k$). In case of a given MBG(Y,G), the types of the other variables X_i can be 1, non-occurring in the MBG, 2, parent of Y,3, children of Y and 4, (pure) other parent with a common child. So the number of types of the variables is in the order of 4^n. These types corresponds to the three categories used in the so called Feature Selection Problem, such as irrelevant (1), strongly relevant (2 and 3), and weakly relevant (4), as can be seen directly from the definitions of relevance [10].

The corresponding order-conditional posterior for the MBG feature given the compatible ordering is as follows:

$$p(\mathrm{MBG}(X_i)=G^{MB}|\succ,D)=$$

$$=p(Pa(X_i,G^{MB})|\succ,D)\prod_{\substack{X_i\prec X_j,\\ X_i\in Pa(X_j,G^{MB})}} p(Pa(X_j,G^{MB})|\succ,D)\prod_{\substack{X_i\prec X_j,\\ X_i\notin Pa(X_j,G^{MB})}} p(X_i\notin Pa(X_j,G^{MB})|\succ,D) \quad (4)$$

The order conditional posteriors for the parental sets $p(Pa(X_j,G^{MB})|\succ,D)$ are well known quantities in Bayesian network learning with analytic formulas, whereas the order conditional posterior that a variable X_j after X_i does not include X_i as a parent can be computed by summing over all the order compatible parental sets not containing X_i as follows

$$p(X_i\notin Pa(X_j,G^{MB})|\succ,D)=\sum_{X_i\notin Pa(X_j)} p(Pa(X_j)|\succ,D) \quad (5)$$

This summation similarly has polynomial time complexity if the number of parents is bounded by k., because the fixed ordering in the condition ensures the conditional independence of the parental sets of the variables. Note, that the order-conditional posterior for the value G^{MB} of the $\mathrm{MBG}(X_j)$ feature a given an ordering $\prec$ is a summation over all the DAG models compatible with the subgraph G^{MB} and with the specified ordering $\prec$ as well. The cardinality of such models can be readily read off from Eq. 4 as follows: the contribution of the variables X_j before X_i (i.e. $X_j\prec X_i$) without any constraint and the contribution of the variables X_j after X_i (i.e. $X_i\prec X_j$) that are not children of X_i. Let denote the number of such variables with $N_{B,\prec}$ and $N_{A,\prec}$ respectively, then assuming that the maximal number of parents is k, the number of the compatible DAGs is $O(n^{k(N_{B,\prec}+N_{A,\prec})})$.

However, the order-free posterior $p(\mathrm{MBG(Y)}=G^{MB}|D)$ summing all the DAG models compatible with the subgraph G^{MB} cannot be computed using the same trick of dealing with parental constraints independently, because of the general dependence of parental sets. Possible alternatives for its Monte Carlo estimation can be gained by rewriting it as an expectation in the space of DAGs or in the space of orderings

$$p(\mathrm{MBG(Y)}=G^{MB}|D)=E_{p(\succ|D)}[p(\mathrm{MBG(Y)}=G^{MB}|\succ,D)]$$
$$=E_{p(\succ|D)}[E_{p(MBG(Y)|D,\succ)}[1(\mathrm{MBG(Y)}=G^{MB})] \quad (6)$$

$$p(\mathrm{MBG(Y)}=G^{MB}|D)=E_{p(G|D)}[1(\mathrm{MBG(Y)},G)=G^{MB})] \quad (7)$$

These suggest the following methods for the Monte Carlo estimation of this expectation:

1 Sample the space of DAGs using the efficiently computable closed-form for the unnormalized posterior p(G|D) and compute the average based on the indicator function 1(.) as $\approx \frac{1}{N}\sum_{i=1}^{N} 1(\mathrm{MBG}(Y, G_i)) \cdot$

2 Sample the space of orderings using the efficiently computable closed-form for the unnormalized posterior p(<|D), for each ordering; sample the order compatible MBGs using the efficiently computable closed-form for the order conditional posterior p(MBG(Y)|D,<) and compute the average based on the indicator function as $\approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} 1(\mathrm{MBG}(Y)_{i,j} = G^{MB}) \cdot$

3 Sample the space of orderings using the efficiently computable closed-form for the unnormalized posterior p(<|D) and compute the average based on the order conditional posterior $\approx \frac{1}{N}\sum_{i=1}^{N} p(\mathrm{MBG}(Y) = G^{MB} \mid \succ, D) \cdot$

These methods form a hierarchy with their requirements and advantages. The first method does not necessitate the posterior p(<|D) (i.e. the assumption of a relatively low bound on the number of parents to ensure tractability). The second method does not require an order conditional posterior for the complex feature, it relies only on the existence of the order conditional posterior for parental sets [5]. On the contrary, the third method utilizes both quantity and offers an analytic computation partially within a simplified Monte Carlo cycle.

Another reason for the second method is the lack of a target MBG G^{MB} or if we would like to estimate the distribution p(MBG(Y)|D) or if we would like to construct a special probabilistic knowledge base containing MBGs. Because of our interest in the classification problem related to variable Y, we investigated this research problem in depth and discuss here an important special case. Under reasonable assumptions it can be shown that an ideal probabilistic knowledge base containing a limited number of MBGs consists of a high probability density region, i.e. the set of MBGs with high posteriors. For each MBG in this knowledge base, the posterior probability and the smoothed (averaged) parameterization of the MBG model can be stored, ensuring that any classification oriented structural, parametric or predictive Bayesian inference can be answered by averaging over the knowledge base. The discussion of this topic exceeds the scope of this paper, so we summarize only the main idea of the algorithm applied in the paper: we apply the order-MCMC method to sample the orderings and for each ordering we sort the compatible MBGs by sorting the posteriors for the parental sets and store and update the estimates of the most probable MBGs using their order conditional posterior.

5 A BN-Feature Based Probabilistic Knowledge Base

The Bayesian network representation is an essentially propositional representation, though it extends the propositional (Boolean) logic by providing scalar values (beliefs) to the propositions instead of binary values. Following the model-theoretic semantics, a propositional knowledge base can be conceived of a representation of a set of worlds (described by the value configurations of all the variables) that are realistic, i.e. it assigns true values to the realistic worlds called models. The probabilistic extension of this model model-theoretic semantics dictates that a probabilistic propositional knowledge base assigns a scalar value (belief) to each world. In fact, a Bayesian network is a representation of a distribution over the atomic events, which are the conjunction of the values for all the variables. Following this idea, a Bayesian network (BN) can be conceived of a representation of probabilistic truth-values (beliefs) over any Boolean proposition ϕ (W), because the probability of the proposition is given by the standard summation over the compatible worlds (models) is as follows:

$$P(\phi) = \sum_{M:\Phi \ igaz \ M-ben} P(M \mid BN) \tag{8}$$

However, the propositional nature of the Bayesian network representation is a serious obstacle towards its application representing statements over objects and their relations and generalized statements over possibly infinite objects. So the extension of this so-called monolithic Bayesian network representation is an active research topic, including for example the investigation of the object-oriented Bayesian networks [12] or the probabilistic relational models [11].

We discuss here another approach to its extension based on the concept of Annotated Bayesian Networks, proposing a language, which can incorporate information about the free-text descriptions of the domain variables.

Consider the following example: "*$A(X_i)$ contains STR_1 and $A(X_j)$ contains STR_2 and X_i=X_j*" (where X_i and X_j denote variables, STR_1 and STR_2 given strings, and $A(X_i)$ the annotation of X_i), representing the event that the values of the variables, the annotations of which contain the specified strings STR_1 and STR_2, are equal. Hence, the proposed language must consist of the following elements: (1) the finite set of domain variables with finite range and their possible instantiation values (domain objects) and the string objects (2) the so-called annotation function, mapping domain objects to free-text descriptions (strings), and (3) a standard set of string functions (e.g. containing). Obviously, this language extends the language of standard probabilistic propositions by certain properties of first-order logic (i.e. domain variables and their values do not have to be instantiated in the statements) and the use of annotations.

In this language, accepting the restriction that only domain objects can be subject of quantification, the probability of any statement is well-defined through Eq. 3,

because of the existence of the distribution over the finite number of domain objects.

Another application of this textual enrichment of an essentially probabilistic propositional knowledge base is the following: consider a first-order language, which includes (1) the DAG models as domain objects and the string objects, (2) the annotation function, and (3) a standard set of string functions as before. Hence, using the probabilistic model-based semantics in Eq. 3, statements containing structural features has well-defined probability, like in the above case. E.g.: "*there is a directed path between any two variables X_1 and X_2, the annotations of which contain the strings STR_1 and STR_2, respectively*".

Note, that the posterior of such statements can be computed analogous to the feature learning problem, i.e. by summing the probabilities of the compatible structures, and that an annotated Bayesian network can be regarded similarly as a first-ordered logical knowledge base.

6 Results

The experiments were performed in the ovarian cancer domain using thirty-five clinical variables selected from a larger study [1].

In the followings, we show examples for the feature learning algorithms discussed in Section 4. The figure below shows the learning curves of basic structural features: directed edge and Markov blanket membership relations between the central variable of the domain (Pathology) and the others. In both cases posterior probabilities were computed using noninformative priors, using a causal ordering of the variables given by an expert of the field.

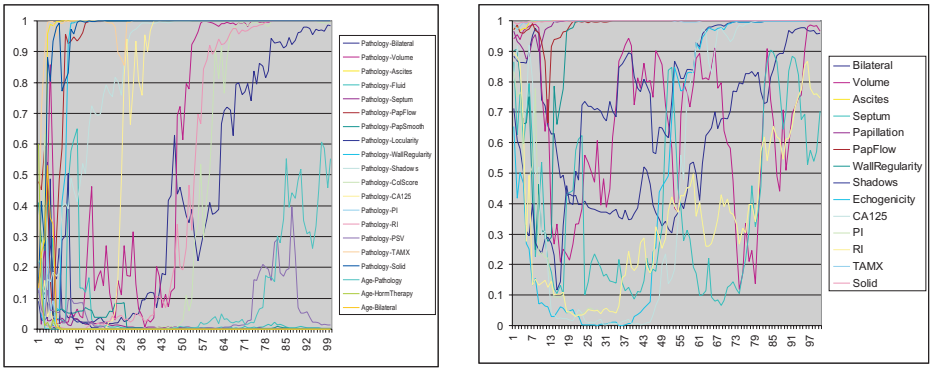

Figure 1

Learning curves of posterior probabilities of structural features for the variable 'Pathology', directed edge on the left, Markov blanket membership on the right.

The next figure shows the learning curves of posterior probabilities of the most probable Markov blanket sets of the variable ‘Pathology’ and the most probable Markov blanket spanning subgraph.

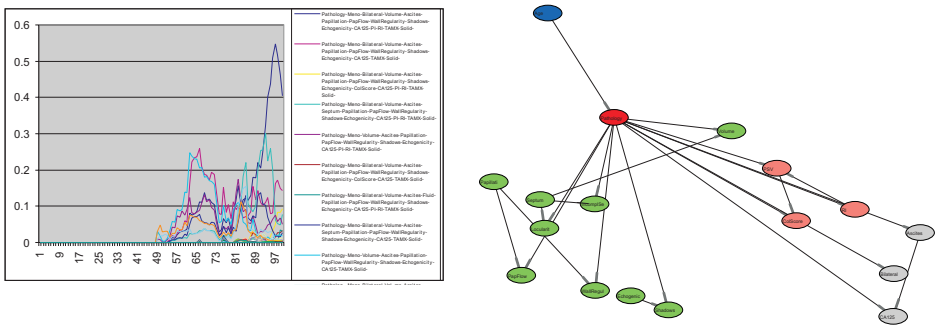

Figure 2

Learning curves of posterior probabilities of the most probable Markov blanket sets, and the most probable Markov blanket spanning subgraph of the variable ‘Pathology’

Fig. 3 shows the learning curves of the averaged posterior probabilities of edges classified by the expert as of high, medium, low and negligible relevance. This figure can be regarded as a comparison of the expert’s prior knowledge and results of the feature learning algorithm: we expect the edges with higher relevance to have higher posterior probabilities.

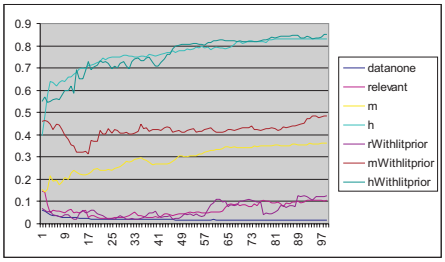

Figure 3

Learning curves of averaged posterior probabilities of edges, corresponding to different relevance classes, given by the expert

Conclusions

The Bayesian statistical application related to Bayesian networks follows the general trend that the growth of computational capacity allows the application of Bayesian methods for more and more complex models, primarily by the usage of general Monte Carlo methods. This is particularly relevant for complex Bayesian network features such as the Markov blanket subGraph feature. The other contribution of the paper, the introduction of the composition of complex statements including structural properties of Bayesian networks and textual annotations also fits in the extension of Bayesian network representation towards a probabilistic first-order logic. Our goal is the integrated development of these two lines of research.

References

[1] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, B. De Moor: Using Literature and Data to Learn Bayesian Networks as Clinical Models of Ovarian Tumors, Artificial Intelligence in Medicine, 2004, Vol. 30, pp. 257-281

[2] J. M. Bernardo: Bayesian Theory, Wiley & Sons, Chichester, 1995

[3] G. F. Cooper, E. Herskovits: A Bayesian Method for the Induction of Probabilistic Networks from Data, Machine Learning, Vol. 9, pp. 309-347, 1992

[4] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer Verlag, 1999

[5] N. Friedman and D. Koller: Being Bayesian about Network Structure, Journal of Machine Learning Research, Vol. 2, pp. 1-30, Kluwer Academic Publ., Dordrecht, Netherlands, 2002

[6] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin: Bayesian Data Analysis, Chapman & Hall, London, 1995

[7] W. R. Gilks, S. Richardson, D. J. Spiegelhalter (editors): Markov Chain Monte Carlo in Practice, Chapman & Hall, 1995

[8] C. Glymour, G. F. Cooper: Computation, Causation, and Discovery, AAAI Press, 1999

[9] D. Heckerman, D. Geiger, D. Chickering: Learning Bayesian networks, Machine Learning, Vol. 20, pp. 197-243, 1995

[10] R. Kohavi, G. H. John: Wrappers for feature subset selection, Artificial Intelligence, Vol. 97, pp. 273-324, 1997

[11] D. Koller, A. Pfeffer, Probabilistic Frame-Based Systems, In Proc. of the 15th Nat. Conf. on A.I. (AAAI), Madison, Wisconsin, 1998, pp. 580-587

[12] K. Laskey, S. Mahoney: Network Fragments, In Proc. of the 13th Conf. on Uncertainty in A.I. (UAI-1997), Morgan Kaufmann, 1997, pp. 334-341

[13] D. Madigan, S. A. Andersson, M. Perlman, C. T. Volinsky: Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs, Comm. Statist. Theory Methods, Vol. 25, pp. 2493–2520, 1996

[14] J. Pearl: Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Francisco, 1988

[15] C. P. Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Springer-Verlag 2001