ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Tomasz Kozdraj**

REMARKS ON BAYESIAN NETWORKS AND THEIR APPLICATIONS

Abstract

Bayesian networks are directed acyclic graphs that represent dependencies between variables in a probabilistic model. They are becoming an increasingly important area for research and applications in the entire field of Artificial Intelligence. This paper explores the nature of implications for Bayesian networks beginning with an overview and comparison of inferential statistics and Bayes' Theorem. It presents the possibilities of applications of Bayesian networks in a field of economic problems and also focuses on the problem of learning.

Key words: Bayesian networks, probabilistic networks, learning Bayesian networks.

I. INTRODUCTION

Bayesian networks are becoming more and more popular in the field of research and applications of artificial intelligence. They play a significant role in decision processes and knowledge representation in expert and decision support systems.

Considering expert system or decision support system with reference to their structure (Fig. 1) one can notice that the role of concluding procedures and knowledge base, where Bayesian networks can be applied, is extremely important.

Returning to Bayesian methods particularly to Bayesian networks it is common knowledge that they are classified to non-classical statistical methods, because the inference is based not only on a sample but also takes advantage of information outside the sample. The information outside the sample is

* Ph.D., Chair of Statistical Methods, University of Łódź.

called *prior* information and presents the most controversial point in the Bayesian theory.

The basis of the whole Bayesian statistics theory is conditional probability theorem published by Thomas Bayes in the year 1763.

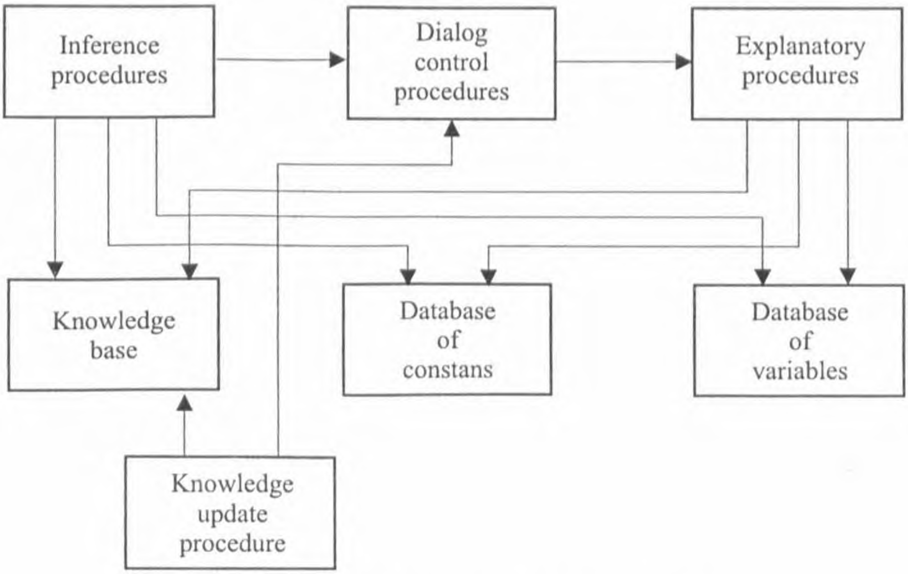

Figure 1. The main elements of an expert system
Source: Mulawka J.J. (1996), *Systemy ekspertowe*, WNT, Warszawa

The mathematical notation of this theorem is the following (for discrete random variable):

$$\mathrm{P}(A_i/B) = \frac{\mathrm{P}(B \setminus A_i)\mathrm{P}(A_i)}{\sum_{j=1}^{N} \mathrm{P}(B \setminus A_j)\mathrm{P}(A_j)} \tag{1}$$

where events $A_1, A_2, ..., A_N$ are independent events called hypotheses and $\mathrm{P}(A_1), \mathrm{P}(A_2), ..., \mathrm{P}(A_N)$ are *prior* probabilities or subjective probabilities. The probability of event A_i conditional on the occurrence of event B (for 1) is known as *posterior* probability.

It should be noticed that most of the probabilities in equation (1) are conditional probabilities. They express the trust for some proposals based on the assumption that other proposals are true.

II. GENERAL NETWORK STRUCTURE AND INFERENCE

The conception of Bayesian networks is directly connected with conditional probability theory. It is easy to notice that in the real world there are many situations in which occurrence of one event is strictly dependent on the occurrence of another event.

Applying Bayesian networks allows more precise modeling of uncertainty and to predict the possibility of occurrence of some situations through using additional information. Since the knowledge about considered problem has probabilistic character and the methods are based on probabilistic concepts it is used to call Bayesian networks as probability networks or belief networks.

Formally one can say that Bayesian networks are graphical structures which represent dependencies between variables. They are directed acyclic graphs which encode the structure of system, its uncertainty and comprehension. This sort of graphs are composed of nodes and edges, where nodes correspond to all random variables.

Therefore, one can say about node X_i corresponding to random variable X_i for $i = 1, 2, ..., n$. The edge (path) directed from node X_i to node X_j can be intuitively interpreted as representation of direct dependence of variable X_j from variable X_i.

The occurrence of such edge is usually symbolically denoted by expression $X_i \rightarrow X_j$. For all nodes we can introduce aftermath relation denoted as $\rightarrow$ that node X_j is successor of X_i or node X_i is predecessor of node X_j in Bayesian network, what can be described as $X_i \rightarrow X_i$ if one of the following conditions comes true :

– there is a directed path from node X_i to node X_j, that is $X_i \rightarrow X_j$;

– there is a directed path from node X_i to some node X_k and node X_j is successor of X_k, that is $X_i \rightarrow X_k$ and $X_k \rightarrow X_j$.

According to this definition node X_j is a successor of X_i, if there exists a path made up of directed edges from node X_i to node X_j. If $X_i \rightarrow X_j$, it means that X_i is a direct predecessor (parent) of node X_j or node X_j is a direct successor (child) of node X_i.

The introduction of aftermath relation is important because it makes possible to define more precisely and formally the way of semantics of Bayesian Network, that is the interpretation of all edges. The network is interpreted as the assertion of conditional independence of each node from all nodes which aren't its successors and with given values of its predecessors. The main idea of this approach is decomposition of the system to simpler parts, showing its modularity (graph theory) and assuring cohesion (probability theory).

The usefulness of Bayesian networks with correct structure consists in the ability to represent in an efficient way the joint probability distribution for all random variables of the model.

If we denote by symbol U_X the set (domain) of nodes which are parents of node X then the effective rule to compute the joint probability distribution requires defining conditional probability distribution $\mathrm{P}(X_i | U_{X_i})$ for each node X_i, that means the probability of variable X_i for the sake of possible outcomes of its parents. Obviously, for a node which does not have parents the conditional distribution is equal to marginal distribution $\mathrm{P}(X_i)$. The basic assumption in graph-based models is the assumption that joint probability distribution $\mathrm{P}(\mathbf{X})$ is equal to the product of marginal and conditional distributions for all random variables. Therefore such a distribution can be denoted in the following way (chain rule):

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i | U_{X_i}) \tag{2}$$

It should be easy to notice that knowledge of the joint probability makes possible the inference about values of any chosen variables when values of other variables are known. If thus it is possible to represent the joint probability distribution of variables using Bayesian networks, it is also theoretically possible to use it for probabilistic inference to get the answer for any question of interest assuming that the structure (topology) is correct.

For this reason one can distinguish two types of inference (Murphy, 2001), Niedermayer (1998)). The first one going through from effect to cause and called bottom up inference and the second one from cause to effect, that is top down inference. In some cases we use approximated inference (Settimi, Smith, Gargoum, 1999).

III. THE TYPES OF BAYESIAN NETWORKS AND LEARNING PROCEDURES

Generally Bayesian networks can be divided into two groups i.e. dynamic Bayesian networks or static Bayesian networks. Dynamic Bayesian networks are used in time series modeling, for example in signal recognition processes. In this case series and network are usually represented by first-order Markov process (Ghabramani, 1997).

If $Y_1, \ldots, Y_T$ are random variables representing time series of first-order Markov process then the joint probability distribution will be equal to:

$$\mathrm{P}(Y_1, Y_2, \ldots, Y_T) = \mathrm{P}(Y_1) \cdot \mathrm{P}(Y_2 | Y_1) \cdot \ldots \cdot \mathrm{P}(Y_T | Y_{T-1}) \tag{3}$$

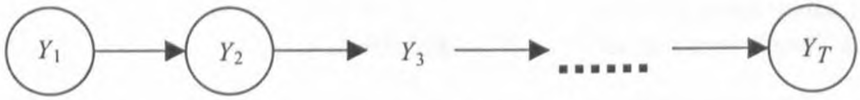

Figure 2. A Bayesian Network representing a first-order Markov process

These models do not directly represent dependencies between observations over more than one time step, therefore it is common to allow higher order interactions between variables i.e. τ^{th}-order Markov models.

Another way to extend Markov models is to posit that the observations are dependent on hidden variables which we can call the states and that the sequence of states is a Markov process. A classical model of this kind is the Kalman filter.

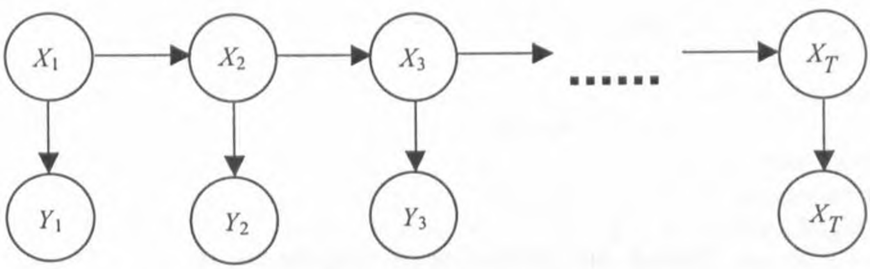

Figure 3. A Bayesian network specifying conditional independence relations for a Kalman filter model

Using the short notation the joint probability distribution for this case for sequence from $t = 1$ to $t = T$ is:

$$P(X_t, Y_t) = P(X_1) \cdot P(Y_1|X_1) \prod_{t=2}^{T} P(X_t|X_{t-1}) \cdot P(Y_t|X_t) \tag{4}$$

The state transition probability $P(X_t|X_{t-1})$ can be decomposed into deterministic and stochastic component

$$X_t = f_t(X_{t-1}) + \upsilon_t \tag{5}$$

where f_t is the deterministic transition function, and υ_t is zero-mean random noise vector.

Similarly, the observation probability $P(Y_t|X_t)$, can be decomposed:

$$Y_t = g_t(X_t) + \xi_t \tag{6}$$

There are other ways of representation commonly used in dynamic Bayesian networks i.e. Hidden Markov Models, Factorial Markov Models and Switching State Models (Bilmes (2000),Ghabramani (1997), Murphy (2002)).

Static Bayesian networks are usually used in medical diagnoses or can be applied as decision support tools in classification problems (for example in communication insurances). The structure of static network does not differ from general scheme (acyclic graph) apart from lack of dynamic variables.

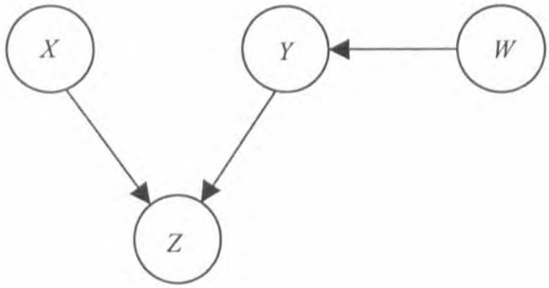

Figure 4. An example of simple static Bayesian network

If the network had for example four random variables X, Y, Z and W (Fig. 4) the joint probability distribution would be as follows:

$$P(W, X, Y, Z) = P(W) \cdot P(X) \cdot P(Y|W) \cdot P(Z|X, Y) \tag{7}$$

Certainly, regardless of the structure, the network can be subject to learning process. Such a process can concern both parameters of the network or the structure, and can be associated with variable selection and edge specification. There are adequate algorithms for parameter learning which allow to obtain the best estimations (for example gradient methods or methods based on maximum likelihood function). Generally parameter learning is simply updating of conditional probability tables for each node of the network. However, more complicated problem is learning the correct structure. The structure learning procedures are based on searching between all possible and acceptable networks of interest to find one or several optimal networks.

In such cases the solution is based on complicated mathematical algorithms based on special metrics e.g. K2 metric and in some cases (for example when some variables are hidden) the proper solution hasn't been found so far.

The searching of all possible networks can be limited by taking into consideration the *prior* knowledge about the problem of interest (expert knowledge) or by imposing additional conditions limiting the structure of Bayesian network. Usually the limitation is connected with interactions order, that is it concerns the maximum number of edges which can be directed to one node.

IV. THE POSSIBILITIES OF APPLICATION. CHANCES AND PROBLEMS

As it was noticed the Bayesian network encodes in a compressed way the joint probability distribution of random variables and this kind of distribution is sufficient for the inference. The answer to any question can be obtained by computing the joint probability distribution on the basis of the network and using it for appropriate calculations.

Unfortunately, such an approach means resignation from one of the best advantages which can be obtained by graphical representation of the joint probability distribution, lying in its efficiency. Of course, Bayesian networks give other advantages, particularly legible and intuitively comprehensible graphical knowledge representation about direct causalities, but effectiveness reasons make it impossible to use this distribution in practice, exception for cases with few number of variables.

Thus, there is a need for other inference algorithms in Bayesian networks. Unfortunately, in general case such a problem is NP-hard. This problem is becoming easier for a special type of networks called single-connection networks. In such networks ant two nodes can be linked only with one path (maximum) composed of freely directed edges. There are known effective algorithms of approximate inference for this kind of networks based on Monte-Carlo methods e.g. logical or weighted sampling. Therefore, there are some practical limitations of use caused by relatively hard obtaining of efficiency.

The field in which Bayesian networks are dynamically developed is medicine. The task of the network in such cases is usually to find the most probable cause of ailment of a patient. Therefore networks have to answer the question: what does the patient suffer from if some symptoms occur (which cannot be classified clearly-out). Hence we often direct to Bayesian networks in classification problems.

Although there is a big interest in Bayesian networks in medicine, their use in social and economical field is not so popular. The actual problem isn't only the mentioned problem of efficiency but also the problem of structure and learning, particularly when the economical variables have dynamic character.

Let us consider hypothetical network with binary nodes, represented in Figure 5. The aim is to estimate the probability of share (KGHM) price fall in future time period.
It is a problem of classification, if the share should be classified to a falling group (or neutrally) or to an increasing group. The network was made with use of Gene Software developed by Decision Systems Laboratory in Pittsburgh University.

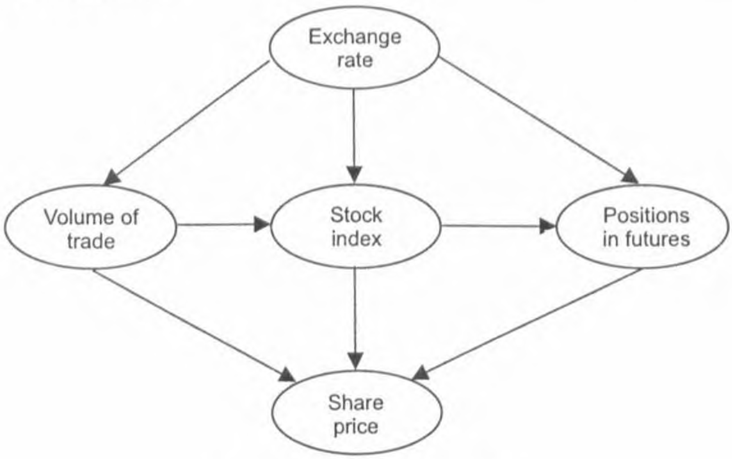

Figure 5. Network structure

Let $X_1, \ldots, Y_5$ respectively represent nodes of the net. The meaning and possible states of nodes is presented in Table 1.

Table 1. List of variables, symbols and categories

Variable	Symbol	Categories
Exchange rate	X_1	increase, fall
Volume of trade	X_2	increase, fall
Stock Exchange index	X_3	increase, fall
Positions in futures	X_4	increase, fall
Share price	X_5	increase, fall

If we denote category fall by 0 and category increase by 1 then adequate marginal and conditional distributions will be following (for learning sample of 47 elements – weekly data from 10.09.2001 to 09.09.2002):

Table 2. Marginal probability distribution $P(X_1 = Y)$ of variable X_1

$Y = 0$	$Y = 1$
0.5116	0.4884

Table 3. Conditional probability distribution $P(X_2 = Y \mid X_1 = Z)$ of variable X_2

	$Y = 0$	$Y = 1$
$Z = 0$	0.7619	0.2381
$Z = 1$	0.7273	0.2727

Table 4. Conditional probability distribution $P(X_3 = Y \mid X_2 = Z, X_1 = Q)$ of variable X_3

	$Y = 0$	$Y = 1$
$Z = 0 \quad Q = 0$	0.3750	0.6250
$Z = 0 \quad Q = 1$	0.4375	0.5625
$Z = 1 \quad Q = 0$	0.4000	0.6000
$Z = 1 \quad Q = 1$	0.5000	0.5000

Table 5. Conditional probability distribution $P(X_4 = Y \mid X_3 = Z)$ of variable X_4

	$Y = 0$	$Y = 1$
$Z = 0$	0.5417	0.4583
$Z = 1$	0.5263	0.4737

Table 6. Conditional probability distribution $P(X_5 = Y \mid X_4 = Z, X_3 = Q, X_2 = R)$ of variable X_5

	$Y = 0$	$Y = 1$
$Z = 0 \quad Q = 0 \quad R = 0$	0.5556	0.4444
$Z = 1 \quad Q = 0 \quad R = 0$	0.5000	0.5000
$Z = 0 \quad Q = 1 \quad R = 0$	0.2000	0.8000
$Z = 0 \quad Q = 0 \quad R = 1$	0.7500	0.2500
$Z = 1 \quad Q = 1 \quad R = 0$	0.5000	0.5000
$Z = 0 \quad Q = 1 \quad R = 1$	0.6000	0.4000
$Z = 1 \quad Q = 0 \quad R = 1$	0.0001	0.9999
$Z = 1 \quad Q = 1 \quad R = 1$	0.0001	0.9999

The rightness of decision making was verified in time period of 23 weeks. It turned out that in 48% of cases the net were classified correctly (in 26% of cases it could not make a decision and in 26% it made a wrong one). Obviously, it cannot be perceived as a good result. The reason for this result could be connected with the structure (topology) of the network which could be improper, and with a selection of variables.

V. SUMMARY AND CONCLUSIONS

Bayesian networks can be an effective tool for statistical inference in computational expert systems. However, there are some barriers connected with the structure of a network, variables dynamization and learning or update processes. These problems are mathematically complicated and this can explain the reason why the neural networks are still more popular than probabilistic networks. Especially the problem of learning of the structure is the matter of research. Besides, the software concerning the neural networks is more accesible.

It should be noticed that graphical knowlegde representation has a great advantage over rule-based knowlegde used in expert and decision support systems. Each rule in rule-based systems are treated independently among others and that's why it may be inconsistent and redundant. In graph-based expert systems these problems do not exist.

REFERENCES

Bilmes J.A. (2000), Dynamic Bayesian multinets, [in:] *Proceedings of the 16th conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, Stanford, California.

Chickering D.M., Geiger D., Heckerman D. (1994), *Learning Bayesian Network is NP-hard*, Microsoft Research MSR-TR-94-17, Redmond.

Diez F.J., Mira J. (1994), *Distributed Reasoning and Learning in Bayesian Expert System*, Dpto. Informatica y Automatica, UNED, Madrid.

Domański Cz., Pruska K., Wagner W. (1998), *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wyd. UŁ, Łódź.

Geiger D., Heckerman D. (1994), *Learning Gaussian Networks*, Microsoft Research MSR-TR-94-10, Redmond.

Ghabramani Z. (1997), *Learning Dynamic Bayesian Networks*, University of Toronto, Toronto.

Heckerman D., Geiger D. (1995), *Learning Bayesian Networks*, Microsoft Research MSR-TR-95-02, Redmond.

Krause P. (1998), *Learning Probabilistic Retworks*, technical report, Philips Research Labs, Redhill.

Mulawka J.J. (1996), *Systemy ekspertowe*, WNT, Warszawa.

Murphy K. (2001), *An introduction to graphical models*, http://www.cs.berkeley.edu/murphyk/Bayes/bayes_tutorial.pdf

Murphy K. (2002), *Dynamic Bayesian Networks: Representation, Inference and Learning*, PhD Thesis, UC Berkeley, Computer Science Division.

Niedermayer D. (1998), *An introduction to Bayesian networks and their contemporary applications*, http://www.gpfn.sk.ca/~daryle/papers/bayesian_networks/bayes.html

Normand S.L., Tritchler D. (1992), Parameter updating in a Bayes network, *Journal of American Statistical Association*, **82**, 420.

Russel S., Binder J., Koller D. (1994), *Adaptive Probabilistic Networks*, technical report UCB/CSD-94-824, University of California, Berkley.

Settimi R., Smith J.Q., Gargoum A.S. (1999), *Approximate Learning in Complex Dynamic Bayesian Networks*, Engineering and Physical Sciences Research Council GR/K72254.

J. Pearl, web page http://bayes.cs.ucla.edu/jp_home.html

Tomasz Kozdraj

UWAGI O SIECIACH BAYESOWSKICH I ICH ZASTOSOWANIACH

Streszczenie

Sieci Bayesa są strukturami graficznymi będącymi skierowanymi grafami acyklicznymi prezentującymi zależności pomiędzy zmiennymi losowymi. Znajdują one zastosowanie w dziedzinie tzw. oprogramowania inteligentnego, a zwłaszcza w systemach ekspertowych. Artykuł ten porusza problemy samych sieci bayesowskich, uczenia oraz ich zastosowania. Podjęto też próbę ich aplikacji na polu zagadnień ekonomicznych związanych z rynkiem kapitałowym.