STUDIA INFORMATICA
Nr 1/2(7) Systemy i technologie informacyjne 2006

Cyclic Bayesian Network – Markov Process Approach

Mieczysław A. Kłopotek [1, 2]
[1] Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw Poland
[2] Institute of Computer Science, University of Podlasie,
ul. Sienkiewicza 51, 08-110 Siedlce, Poland

Abstract. The paper proposes a new interpretation of the concept of cyclic Bayesian Networks, based on stationary Markov processes over feature vector state transitions.

Keywords. Bayesian Networks, Markov process

1 Introduction

Bayesian networks (BN), used at least since the important publications of Pearl [Pearl:88], have been considered as a representation of the joint probability distribution in multiple variables. They consist of two essential parts: the directed acyclic graph (DAG), representing (or intended to represent) causal relationships among the variables (being nodes of the graph), and conditional probability tables representing conditional distribution of the variable node given its parents. The joint probability distribution is expressed by the formula:

$$P(x_1',...,x_n'|x_1,...,x_n) = \prod_{k=1\ldots n} P(x_k' \,|(x_1,...,x_n) \downarrow pa(X_k)),$$

where $\downarrow$ is the projection operator, and $pa(X_k)$ is the set of parents of X_k in the DAG. (see Fig.1)

This point of view proved to be very fruitful resulting in development of many algorithms for knowledge acquisition from data as well as numerous practical applications of reasoning algorithms for BNs, in medical and technical diagnosis, assistant programs for complex editor programs etc.

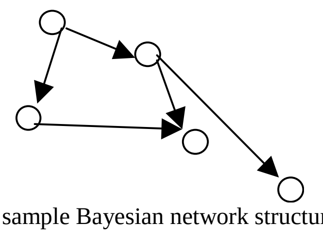

Figure 1. A sample Bayesian network structure (DAG)

However, also a different point of view is possible, related to so-called dynamic Bayesian networks [Ghahramani:98], [Zou:05], which represent evolution of a system over time.

The cyclicity of a DAG generated by an automated knowledge extraction algorithm was considered either to be a fault of the algorithm, or a deficiency of the data collection process – it was assumed that some important variables remained uncovered. Models of hidden variables were constructed, which, however, cannot explain some phenomena, that may occur in the data.

For this reason, we attempt here to introduce cyclic Bayesian networks in a systematic way, grounding them in an interpretation of a Markov process.

In section 2 we will recall the Markov process and embed BNs into it. In section 3 we will define the cyclic Bayesian networks. Section 4 contains some remarks on reasoning in cyclic BNs. Section 5 makes an outline of a possible approach to learning such networks from data. The paper ends with some remarks of general nature in section 6.

2 Bayesian Network As State Transition

In probability theory, a stochastic process has the Markov property if the conditional probability distribution of future states of the process, given the present state, depends only upon the current state, i.e. it is conditionally independent of the past states (the path of the process) given the present state. A process with the Markov property is usually called a Markov process[1] or Markov chain.

The transition matrix P of a Markov chain is a square matrix, where P_{ij} is equal to the probability of making the transition from a state i to a state j, in one step.

A square matrix with real, non-negative entries, where each row sums up to 1, is called a row-stochastic matrix. By definition, each transition matrix of a Markov chain is row-stochastic and vice versa.

Let GM(S,T) denote a directed graph corresponding to a Markov chain defined in the following way. The set S represents the set of all the states of Markov Chain M and for i,j in S there is a directed edge (i,j) in T if and only if there is a non-zero transition probability of changing the state from i to j in a single step.

We say that a Markov chain M is irreducible if its corresponding graph GM is strongly connected (for every two nodes i and j there is a directed path from i to j in GM).

We say that a Markov chain M is aperiodic if its graph GM has the following property: the greatest common divisor of lengths of all cycles in this graph is equal to 1.

A stationary distribution d of a Markov chain M with the transition matrix P is a probability vector over the set of states S which satisfies the condition d=P*d }.

If Markov chain over a finite set of states is irreducible and aperiodic, the stationary distribution exists and is unique [Motwani:1995].

[1] en.wikipedia.org/wiki/Markov_process

Now let us consider a set of states S such that each state s from S is an n-dimensional vector over a set of variables $X=\{X_1,X_2,\ldots,X_n\}$. Now pa state would have the form $(x_1,x_2,\ldots,x_n)$ x_k is a value of the variable X_k.

Now let the state transition from state $(x_1,x_2,\ldots,x_n)$ to state $(x_1',x_2',\ldots,x_n')$ have the form

$$P(x_1',\ldots,x_n'|x_1,\ldots,x_n) = \prod_{k=1\ldots n} P(x_k' |(x_1,\ldots,x_n) \downarrow pa(X_k)),$$

where $\downarrow$ is a projection operator and $pa(X_k)$ is the set of parents of variable X_k in a dag D A symbolic transformation of the dag from Fig. 1 to a Markov chain model is visible in Fig. 2. In Fig. 3 we see that by chaining multiple steps of a Markov chain we obtain a network where we can recognize the structure of the original dag. Notice further, that if we extend the Markov chain representation to "infinity", then there will still exist only finite length directed paths can trace infinite directed paths (with head-to-tail meeting points only).

Obviously, if the conditional probabilities are all non-zero, then the stationary distribution d will be given by the probability distribution defined by the Bayesian network with dag D, and conditional probabilities $P(x_k |(x_1,\ldots,x_n) \downarrow pa(X_k))$, that is

$$d=\prod_{k=1\ldots n} P(x_k|(x_1,\ldots,x_n) \downarrow pa(X_k)) .$$

In this way, we have defined a Bayesian network distribution in terms of a stationary distribution of a Markov chain.

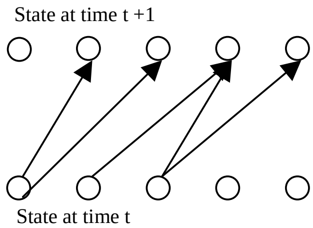

Figure 2. Markov chain view of a BN

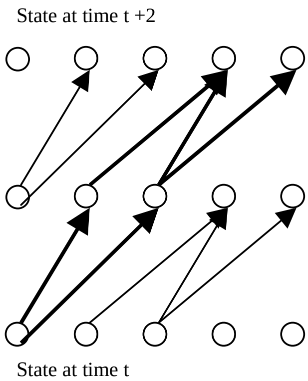

Figure 3. A BN becomes visible in a longer Markov chain

We could also define a Markov chain connecting state $(x_1,x_2, \ldots,x_k,\ldots,x_n)$ with state differing by one variable only $(x_1,x_2',\ldots,x_k',\ldots,x_n')$ with state transition probability:

$$P(x_1,\ldots,x_k',\ldots,x_n|x_1, \ldots,x_k,\ldots,x_n) = P(x_k' |(x_1,\ldots,x_n) \downarrow pa(X_k)),$$

(A proper definition of transition to the same state is needed).

Again the same stationary distribution is achieved, with the same condition for irreducibility.

Note that both above-mentioned concepts of Markov chain BN are basis for some well-known sample generation algorithms.

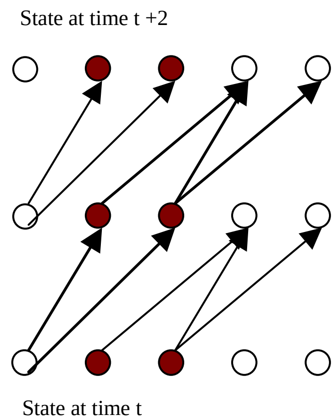

Figure 4. d-separation idea in a Markov chain

An important concept in BNs is the concept of d-separation. If two variables are d-sparated by a set of other variables, then these two are conditionally independent given the others. A d-separation means that there is no path between the nodes such that no tail-to-head or tail-to-tail edge meeting node belongs to the other set and each head-to-head meeting node either belongs to the other set itself or has a (direct or indirect) successor therein.

Returning to the first concept of a Markov chain representation of s sequence (Fig. 3) we can easily translate this concept to the Markov chain representation as in Fig. 4. If we extend the graph to infinity, the d-separation property in the "infinite" graph will still hold.

3 Cyclic Bayesian Networks

In the traditional view, the philosophy of Dags was straightforwardly grounded on our basic epistemological assumptions: nothing can be the cause of itself, neither directly nor indirectly.

But in the context of the interpretation of BN in terms of a Markov chain, it makes sense to drop the assumption of acyclicity of BNs. The state of a variable is not influenced by itself, but rather the future state is influenced by the past one.

Cyclic Bayesian networks are usually considered as faulty results of BN learning from data [Guyader:01]. Either structural limitations of the learning algorithm (e.g. Pearl's algorithm learning only poly-trees even if the real distribution is more complex), or presence of hidden variables (e.g. in case of PC algorithm of Spirtes, Glymour and Scheines) may lead to emerging cycles in the network structure (Fig. 5).

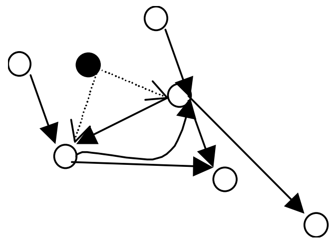

Figure 5. A simple effect of a hidden variable when recovering BN structure from data. The dark node is a hidden variable. The cycle would be absent, if the structure discovery algorithm were conscious of it

Procedures, such as the CI and FCI algorithms of Spirtes, Glymour and Scienes have been developed to detect potential hidden variables.

However, there are probability distributions that cannot be described in this way even with the concept of hidden (latent) variables. If we drop the acyclicity condition, the Markov chain interpretation of a cyclic BN may be the same as for the classical case, that is:

$$P(x_1',...,x_n'|x_1,...,x_n) = \prod_{k=1...n} P(x_k' \,|(x_1,...,x_n) \downarrow pa(X_k)),$$

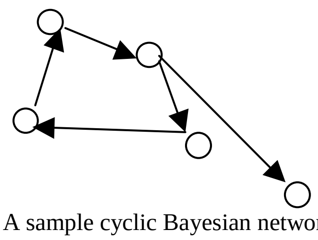

Figure 6. A sample cyclic Bayesian network structure

Obviously, the joint probability distribution will be considered as the respective stationary distribution. The interesting property would be the conditional independence of non-neighbours in the cycle visible in the picture (Fig. 6, 7) given the other two elements of the cycle. This phenomenon cannot be explained by hidden variable hypotheses.

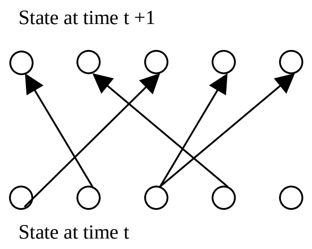

Figure 7. Markov chain view of a cyclic BN

Note that if we create an “infinite” Markov chain representation of a cyclic BN (Fig. 8), we do not see the cyclicity directly (in fact we obtain a dag). But the cyclicity becomes visible because we can trace infinite directed paths (with head-to-tail meeting points only).

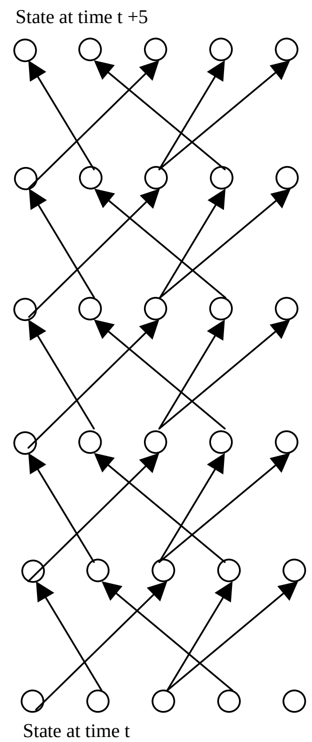

Figure 8. A longer Markov chain of a cyclic BN

4 Reasoning in Cyclic Bayesian Networks

The fundamental problem with reasoning in cyclic Bayesian networks is that the joint probability distribution is not given explicitly, but is rather an asymptotic limit of a multi-stage process.

Therefore, when reasoning with cyclic BN, we would rely on the Markov chain interpretation. In fact, one can exploit many existing reasoning techniques based e.g. on Gibbs' sampling.

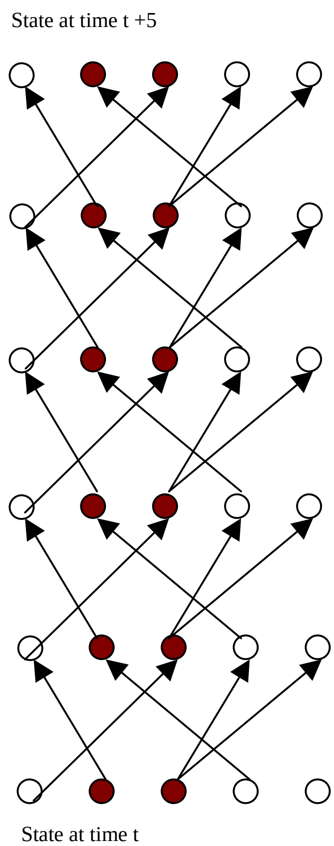

Figure 9. d-separation in a Markov chain of a cyclic BN

An important issue when reasoning about BNs is detection if there holds the relationship of d-separation.

In Fig. 9 we presented a longer Markov chain time sequence for the cyclic BN in Fig. 7. We can now extend the concept of d-separation for a cyclic BN as being defined as d-separation in the infinite dag of a Markov chain time sequence. For practical purposes, the actual sequence may be limited to twice the number of edges in the cyclic BN because no (undirected) path can be longer than the number of edges.

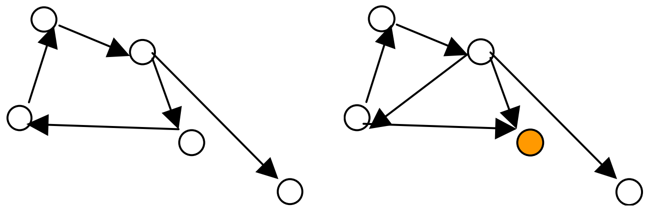

Figure 10. Reducing cycle length in a Bayesian network structure

Note also, that by the technique of the edge reversal we can theoretically shorten any cycle in the cyclic BN until it consists of only one node, for which the transformation to a non-cycle is apparently trivial. (Fig. 10) This is, however, an illusive tactic except for very simple networks, because all the parents of a node moved out of the cycle remain parents of some node in the cycle, so that occasionally a very large number of variables may condition the left nodes of the cycle.

5 Issues When Learning Cyclic Bayesian Networks

Basic problems with learning of cyclic BN from data lies in the difficulty of finding a proper counterpart of the well-known concept of d-separation, on which many approaches rely.It is, however, very helpful to keep in mind, that if proper structure is found, then the conditional probabilities of child on parents will be directly reflected in the stationary distribution.So, in fact, when the actual distribution reflects the stationary one, we can use statistical independence tests as indicators of d-separation.

It seems that no complexities would arise for maximum based on Bayesian methods of structure recovery.

6 Final remarks

In this investigation, we have pointed to possibilities of handling the concept of cyclic Bayesian networks within the framework of Markovian processes. The potential and application areas behind this new kind of BNs are still to be carefully explored. But we can state for sure that this concept seems to be mathematically sound.

Acknowledgement

Research has been partially supported under KBN research grant 4 T11C 026 25 "Maps and intelligent navigation in WWW using Bayesian networks and artificial immune systems"
http://www.ipipan.waw.pl/~klopotek/mak/current_research/KBN2003/KBN2003Translation.htm

References

[Cooper:92] G.F. Cooper, E. Herskovits: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 (1992), 309-347.

[Geiger:90] D. Geiger, T. erma, J. Pearl: d-Separation: From theorems to algorithms. W: M.Henrion, R.D.Shachter, L.N.Kamal, J.F.Lemmer (eds): *Uncertainty in Artificial Intelligence* 5, Elsevier Science Publishers B.V. (North-Holland), 1990, 139-148.

[Ghahramani:98] Z. Ghahramani: Learning dynamic Bayesian networks. In C.L. Giles and M. Gori, editors, Adaptive Processing of Sequences and Data Structures, volume 1387 of Lecture Notes in Computer Science, pages 168--197. Springer, 1998.

[Guyader:01] Guyader A. Fabre, E. Guillemot, C. Robert, M.: Joint source-channel turbo decoding of entropy-coded sources, IEEE Journal on Selected Areas in Communications, Sep 2001, Volume: 19, Issue: 9, pp. 1680-1696 URL: www.uhb.fr/sc_sociales/ labstats/AGUYADER/doc/ieee-jsac.pdf

[Motwani:95] R. Motwani, P. Raghavan: Randomized Algorithms. Cambridge University Press, 1995.

[Pearl:88] J. Pearl: Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference, Morgan Kaufmann, San Mateo CA, 1988.

[Spirtes:93] P. Spirtes, C. Glymour, R. Scheines: *Causation, Prediction and Search*, Lecture Notes in Statistics 81, Springer-Verlag, 1993.

[Zou:05] Min Zou, Suzanne D. Conzen: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data Bioinformatics 2005 21(1): 71-79.