Strategies for Improving the Interpretability of Bayesian Networks Using Markovian Time Models and Genetic Algorithms

Ádamo L. de Santana*[a], Cláudio A. Rocha[b], Carlos R. Francês[a],
Solon V. Carvalho[c], Nandamudi L. Vijaykumar[c], João C. W. A. Costa[a]
[a]Federal University of Pará, R. Augusto Côrrea, 01, 66075-110, Belém, PA, Brazil;
[b]University of the Amazon, Av. Alcindo Cacela, 287, 66060-902, Belém, PA, Brazil;
[c]National Institute for Space Research, Av. dos Astronautas 1758, Jd. Granja
12227-010, São José dos Campos, SP, Brazil

ABSTRACT

One of the main factors for the success of the knowledge discovery process is related to the comprehensibility of the patterns discovered by the data mining techniques used. Among the many data mining techniques found in the literature, we can point the Bayesian networks as one of most prominent when considering the easiness of knowledge interpretation achieved in a domain with uncertainty. However, the static Bayesian networks present two basic disadvantages: the incapacity to correlate the variables, considering its behavior throughout the time; and the difficulty of establishing the optimum combination of states for the variables, which would generate and/or achieve a given requirement. This paper presents an extension for the improvement of Bayesian networks, treating the mentioned problems by incorporating a temporal model, using Markov chains, and for intermediary of the combination of genetic algorithms with the networks obtained from the data.

Keywords: Bayesian networks, Markov chains, genetic algorithms, time models, optimization.

1. INTRODUCTION

One of the main factors for the data mining process success is due to the comprehensibility of the discovered patterns [1]. In this way, the use of mining techniques that provide mechanisms of presentation and visualization to simplify the analysis of the knowledge obtained can strongly contribute for the users to measure the quality of this knowledge. This quality is often given by the utility, novelty and interestingness of the patterns obtained [2].

Among the data mining techniques available, it is possible to point the Bayesian networks (BNs) as one of most prominent ones when it comes of interpretability of the knowledge obtained in a domain with uncertainty, once that they are capable to represent the complete causal model from the date [2] [3].

However, the BNs presents some “restrictions”, amongst which we can point: the difficulty to correlate the variables considering the factor of time and the difficulty to establish which is the optimum combination of states for given variables that would achieve a certain requirement (desirable state for a certain variable of the domain).

This paper presents methods to treat the mentioned problems by incorporating a stochastic model, associating Markov chains to the BNs; and by combining genetic algorithms (GAs) with the networks obtained from the data.

The paper is organized as follow: in section 2, the basic concepts of BNs are shown. Section 3 presents the methods for improving the interpretability of the BNs, using the concepts of Markov chains. In section 4, the use of GAs is presented for the discovery of the optimum combination of values for the variables of a BN, as a way to maximize a target variable. The final remarks of the paper are presented on section 5.

*adamo@ufpa.br; phone +55 91 3201-73021; lprad.ufpa.br

2. BAYESIAN NETWORKS

The BNs can be seen as models that codify the probabilistic relationships between the variables that represent a given domain [4]. These models possess as components a qualitative structure, representing the dependences between the nodes, and a quantitative, evaluating, in probabilistic terms, these dependences [2]. Together, these components propitiate an efficient representation of the distribution of joint probability for the variables X of a given domain [3]. The joint distribution is given by (1):

$$P(x_i \mid c_1, c_2, ..., c_n) = P(x_i)\prod_{k=1}^{n} P(c_k \mid x_i) \tag{1}$$

Where $c_1, c_2, ..., c_n$ are possible evidenced events and x_i is the event we want to observe.

One of the great advantages of the BNs is its semantics, which facilitates, given the inherent causal representation of these networks, the understanding and the decision making process, by the part of the users of these models [2]. This if due, basically, to the fact that the relations between the variables of the domain are able to be visualized graphically, besides providing an inference mechanism that allows to quantify, in probabilistic terms, the effect of these relations.

However, these inferences do not provide time analyses in the data, neither accomplish optimization processes, aiming to obtain the values that would maximize a given variable.

The majority of the works proposed in the literature that establishes new strategies of inferences in models that use probabilistic reasoning are based on hybrid solutions, such as the union of solutions that involve Bayesian techniques, Markov models, GAs and Fuzzy Logic [4] [5] [6].

In sections 3 and 4, the mechanisms that implement the inferences of forecast and optimization, using, respectively, Markov chains and GAs will be presented.

3. MARKOV CHAINS INCORPORATION MODEL

Despite allowing, through inferences, to verify the future behavior of its attributes, the BNs does not present means that would allow us to discover how close or distant these events would be from happening, that is, it does not allow us to quantify and point in the time how long it would take for the impact of these inferences to occur.

A classic initial problem when working with the BNs in the time would be the existing necessity to mount conditional probability tables for each discrete unit of time analyzed. This way, it is assumed here, as well as described in literature [4], to be working with a stationary random process.

In this paper it is presented the time analysis from the modeling of the data and characteristics proceeding from a BN into a Markov chain. The idea is to establish an isomorphism of a BN in the time as being a discrete time Markov chain.

The model used seeks to analyze the forecast, differently as it would be if a hidden Markov model would be used or even a dynamic Bayesian network, modeling in a simplified way the Markovian time transition according to a 1st order process, but also intrinsically considering in its transitions the other variables of the domain that also might also influence in the behavior of this attribute. That is, just as a Markov chain, a BN can be seen as a matrix of attributes that are correlated and that also has an influence over each other throughout the time.

To exemplify the model, a simple example of a BN can be considered (Figure 1), composed by only two variables: *Grade* and *Study*; where the grade obtained on a given test depends on the amount of study applied. It is also assumed that the tests are taken on a monthly time scale.

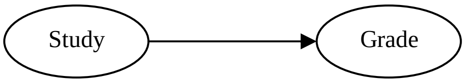

Fig. 1. Bayesian network mounted with the variables *Grade* and *Study*.

It is considered as possible values for the attributes the following: Study (Hard, Medium, Little); and Grade (Excellent, Good, Regular). In this way, the BN would also present the values of initial and conditional (for Grade only, once that it is the only attribute that possess a parent attribute, that is, a dependence relation of the Grade given the Study) probabilities.

The dependences model and the probability tables would represent all the data the BN could offer us. Following the Markovian modeling, what we are seeking to obtain is the time instant n that, given an inference, a determined probability configuration of an attribute would happen (e.g. considering our example, when we would have, given that we study *Hard*, that the grade obtained would be Excellent with probability of 70%, Good with 25% and Regular with 5%).

Given that what we seek is in fact the new configuration of a determined attribute, what we end up needing is to mount the Markovian transition matrix of this attribute. This is done by mapping the transition probabilities for the states of the attribute onto the matrix, based on the conditional probabilities that it possess given its dependencies with the other attributes (e.g. also considering the example, we must map the transition probabilities of Grade for: *Excellent* and pass to *Good*, *Excellent* to *Regular*, *Excellent* and achieving *Excellent* again etc). That is, we would have to mount a Markov transition matrix, according the model on Table 1.

Table. 1. Model of the Markov transition matrix to be mounted.

$$\begin{array}{c|ccc|}
\textbf{Grade\textbackslash Grade} & \textit{Excellent} & \textit{Good} & \textit{Regular} \\
\textit{Excellent} & A_{E\to E} & A_{E\to G} & A_{E\to R} \\
\textit{Good} & A_{G\to E} & A_{G\to G} & A_{G\to B} \\
\textit{Regular} & A_{R\to E} & A_{R\to G} & A_{R\to R}
\end{array}$$

However, considering only the factor of study in relation to the grade is not enough to verify the relation of the variable *Grade* with itself and to make the transition between its states, once that this way the Markov transition matrix would immediately converge to the stationary state. This way, we must also consider the value of the attribute *Grade* in a previous point of time, acting together with the variable *Study* and thus obtaining the transition relations for the variable *Grade*.

For such, the first record in the existing historical database is ignored so that we can insert in the analysis, analogously to a 1st order Markovian process, the *Previous Grade* obtained. Tables 2 and 3 present the initial and conditional (*Study*, *Grade* and the *Grade* in the previous period) probabilities of the *current Grade* considering the *Study* and the *Previous Grade* (Grade-1).

Table. 2. Initial probabilities of the Bayesian network.

Study	
Hard (Ha)	0.133
Medium (Me)	0.534
Little (Li)	0.333

Grade	*Grade*	*Grade-1*
Excellent (E)	0.210	0.333
Good (G)	0.467	0.333
Regular (R)	0.323	0.333

Table. 3. Conditional probabilities of the Bayesian network – P(Grade | Study $\cap$ Grade-1).

Study $\cap$ G-1\Grade	*E*	*G*	*R*
$Ha \cap E$	0.934	0.033	0.033
$Ha \cap G$	0.333	0.333	0.333
$Ha \cap R$	0.333	0.333	0.333
$Me \cap E$	0.491	0.491	0.018
$Me \cap G$	0.033	0.934	0.033
$Me \cap R$	0.018	0.491	0.491
$Li \cap E$	0.333	0.333	0.333
$Li \cap G$	0.018	0.491	0.491
$Li \cap R$	0.033	0.033	0.934

The calculus for the Markov transition matrix would follow according (2):

$$P(E \rightarrow G) = P(E) \times \left[P(G|Ha \cap E)P(Ha) + P(G|Me \cap E)P(Me) + P(G|Li \cap E)P(Li) \right] \tag{2}$$

Generalizing we would have (3):

$$P(A_{x \rightarrow y}) = \frac{\sum_{i=1}^{n} P(A_y \mid A_x \cap Pa_i) \times P(Pa_i)}{\sum_{j=1}^{m} \sum_{k=1}^{n} P(A_j \mid A_x \cap Pa_k) \times P(Pa_k)} \tag{3}$$

Where:

A is the observed variable;
Pa is the variable that represents the attributes of which variable A presents a dependency;
n is the number of possible states and/or combinations that the parents of this attribute can assume;
m is the number of states the attribute can assume.

Calculating from (3), we obtained the Markov transition matrix presented below (Table 4).

Table. 4. Markov transition matrix obtained.

Grade\Grade	*Excellent*	*Good*	*Regular*
Excellent	0.497	0.378	0.125
Good	0.068	0.707	0.225
Regular	0.065	0.318	0.618

The matrix obtained presents the transition probabilities values for the states of a given variable analyzed. If we apply a solution of the chain to find the probabilities vector in a given time n, we will then have to calculate the nth power of the random probabilities matrix. As described by the Equations of Chapman - Kolmogorov [7], presented below:

In matricial notation, the expression is:

$$P^{(n)} = P^{(m)} \times P^{(m-n)} \tag{4}$$

Where $P^{(n)}$ is the transition matrix in the step n. From (4) it can be concluded, therefore, that:

$$P^{(n)} = P^{n} \tag{5}$$

Demonstrating that the matrix in step n corresponds to the nth power of this matrix. Thus, for example, if the unit of time is discretized in months and if we wanted to obtain the probabilities values for the grades occurrence three months from now, we would have that to find the power P^3 of the matrix (Table 5).

Table. 5. States transition matrix in the step *n* = 3.

Grade\Grade	*Excellent*	*Good*	*Regular*
Excellent	0.1878	0.5274	0.2851
Good	0.1085	0.5561	0.3359
Regular	0.1071	0.4976	0.3974

4. MAXIMIZATION MODEL

The objective of the maximization model is to identify the best configuration, among the possible values of the existing variables in the domain, that maximize a given attribute, identifying initially the other variables that present a dependency from it. It is also worth mentioning that, just it is possible accomplish the same process using not the maximum value, but any other value that may be relevant.

In contrast with the way the GAs are used in the majority of the hybrid systems proposed in literature, where it is adopted to optimize the process of learning of the structure of BNs [5] [8] [9] [10], here, this technique is used for the discovery of the most probable values of the variables of a BN, given the value of an key attribute.

The analyses described here were originated from the demands of the research project “PREDICT - Support Decision Tool for Load Prediction of Electrical Systems”, financed by the “National Agency of Electric Energy of Brazil - ANEEL” in course since September of 2004. This project, made together with the Government of the state of Pará and the Power Supplier the State of Pará, aims at projecting and implementing a decision support system, using mathematical and computational intelligence methods, to foresee the necessity of energy purchase in the futures market and to make inferences on the power system situation, from the consumption historical data of its correlations with socio-economic and climatic data.

The case study example, proposed by the domain specialists of the power system market, and used for the optimization model was to discover under which circumstances the power consumption would be maximized. For such, the optimization model was based on a few steps.

Firstly, identify which, among the attributes of the database, influences directly in the power consumption; also this way building the BN structure.

For the learning of the BN graphical model, that is, for the learning if existing correlations between the variables, it was used the search and scoring algorithm K2 [11], which allows us to find the most probable belief network structure B_S given a dataset D. The K2 algorithm applies a Bayesian scoring method, according (6).

$$P(B_S \mid D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(r_i)}{\Gamma(r_i + N_{ij})} \prod_{k=1}^{r_i} \Gamma(N_{ijk} + 1) \qquad (6)$$

Where:

n is the number of nodes;

q_i is the number of configurations of the parents of the variable X_i;

r_i is the number of possible values of X_i;

N_{ijk} is the number of cases in D where the attribute X_i is evidenced with its value k, and the configuration of the parents of X_i is evidenced with value j;

N_{ij} is the number of observations in which the configuration of the parents of X_i is evidenced with the value j, being $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

Analyzing it, it was verified, and confirmed by the domain experts, the correlation of the power consumption with some of the remaining variables, especially with the following: the number of employments in the sectors of the transformation industries and agropecuary, and the values of the turnover total and of the dolar; which were shown to be more representative than the others..

Given the knowledge that the variables of number of employments in the transformation industries (*emp_ind*), employments in the agropecuary (*emp_agro*), value of the turnover total (*val_turn*) and the value of the dollar (*val_dol*) are the main influencers in the variation of the power consumption, they were used for the next step, which consisted in the mounting of a Naive Bayesian Network (Figure 2), in which all of the remaining attributes (*emp_ind, emp_agro, val_turn* and *val_dol*) were dependent of only one, in our case, the power consumption.

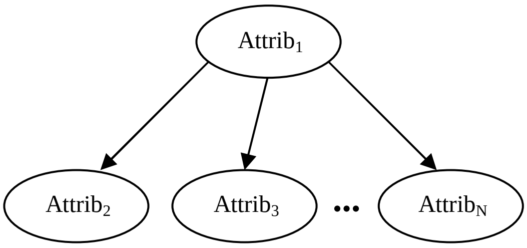

Fig. 2. Example of a Naive Bayes.

All the attributes were discretized on ten states, according to the frequency of their values, allowing us to verify the probability associated to each one of them, as well as the conditional probabilities existing between the variables.
Once mounted the network, the next step is, making use of the data given by the BN, search the network attributes for the states that would maximize the power consumption. In this step we use an altered GA.

Here, instead of a cost function to validate the individuals of the population, it is implemented a Bayesian inference algorithm (equation 1); each of the individuals of the genetic algorithm represents an inference configuration of the BN generated randomly (e.g. evidencing the variables *emp_ind* with the state 7, *emp_agro* with state 1, *val_turn* with 7 and *val_dol* with 4 generates the individual 2-1-7-4). Each individual is then, for its classification, submitted to the Bayesian inference module in order to verify the probability in which the attribute power consumption would be maximized, obtaining, at the end of the iterations, the best possible configuration of inferences on the BN for the maximization of the power consumption.

However, we would have at the end of this step (after the GA analysis) only the respective states (i.e. band of values) for this maximization, instead of a single value (for each attribute), which is what we seek. For such, we make use again of a GA; but this time a traditional GA, whose aptitude function we obtain from the data.
The function used for the GA is obtained from a regression of multiple variables [12] [13] [14] made over the attributes of the network. The multivariate analysis is however made over the consumption data, but considering only the data instances located within the bands found in the previous step. This way, we obtain an equation with a good representativity (R^2 of approximately 0.9039) over the domain. The equation obtained is presented below.

$$Y = 258{,}598{,}510.5 + 3{,}675.6834 \times X_1 + 4{,}430.9036 \times X_2 + 0.4701 \times X_3 - 12{,}182{,}208.61 \times X_4 \tag{7}$$

Where Y represents the power consumption and X_1, X_2, X_3 and X_4 represent the values of the attributes *emp_ind*, *emp_agro*, *val_turn* and *val_dol*, respectively.

Based on function (7), the GA is then used, thus obtaining the values, for each of the attributes, that would maximize the power consumption. Mentioning again that the individuals evaluated by the aptitude function (7) are only those within the bands of values that maximize the value of consumption. Thus, in order to achieve the occurrence of the maximum consumption, equivalent the 305,760,544 MWh, it is necessary that the values on Table 6 are observed, for the attributes *emp_ind*, *emp_agro*, *val_turn* and *val_dol*.

Table. 6. Values of the attributes for the maximization of the consumption value.

Attribute	**Value**
emp_ind	5,380
emp_agro	3,357
val_turn	R$ 100,752,576.00
val_dol	R$ 2.861

It is worth mentioning that the optimization model used is not restricted only to the discovery of the maximum values of consumption, but can also be used to identify the scenarios that cause a minimum, average or another value to be reached by the power supplier, given the variation of the considered economic aspects.

Among the main results obtained in the "Predict" Project with the use of this model, it is possible to highlight: the extension of the interpretability of the generated BNs to measure the causal relationship for the consumption and the socio-economic variables, from the discovery of the values that compose an optimum combination given a certain target, for example, the consumption; and the interest of those involved in the Project in using the functionalities of the model for many other scenarios, not only relative to the power consumption, but also to government actions (e.g. discovered of the variables, that would maximize the employment and income), has encouraged and certified the use of the proposed model.

5. FINAL REMARKS

The possibility to represent graphically the structure of the patterns obtained from the data, as well as the exploratory character of the analyses allowed by the BNs, makes possible to indicate more deeply the relationship between the variables of a domain, what favors the increase of the comprehensibility of the discovered patterns, as well as the identification of the usefulness and relevance of these patters.

Statistical models of correlation and regression can associate the idea of time to many variables, however, possess restrictions regarding the studied scale, besides omitting information such as dependences between the variables. Thus, to keep the characteristics of interpretability, adding a time approach is a sufficiently interesting contribution for applications that represent problems of the real world, in which the time connotation is almost always obligatory. For such, this work proposes a Markovian approach to represent correlations in the time. This approach introduces innumerable advantages, amongst which we can point that, the Markovian models possess relatively simple solutions compared to its "computational effort" and to the mathematical complexity involved, what stimulates and facilitates its use.

Moreover, there is another present necessity for studies of problems of the real world, that, usually, require some kind of interpretation relating aspects of optimization, of cost functions and maximization and minimization of values of the involved variables.

Again, traditional methods of optimization (simplex, for example) or models of evolutionary computing can be used. This work proposes a hybrid model based on the association of the interpretation given by the BNs with GAs, in order to obtain, given the value of a parameter-target, the Bayesian combination necessary to achieve it. Thus, the cost function becomes a BN.

With these strategies it is possible to extend the interpretability of the BNs and adjust them even further for applications of the real world, providing the decision support systems with innumerable other possibilities of interpretation and inferences.

ACKNOWLEDGEMENTS

The authors would like to thank the financial support given by: CAPES, CNPq and ANEEL/Rede Celpa.

REFERENCES

1. S. O. Rezende, *Sistemas Inteligentes - Fundamentos e Aplicações*, Manole, 2003.
2. Z. Chen, *Data Mining and Uncertain Reasoning - an Integrated Approach*, John Wiley Professional, 2001.
3. J. Pearl, *Probabilistic Reasoning in Intelligent System*, Morgan Kaufmann Publishers, 1988.
4. S. Russel, P. Norvig, *Artificial Intelligence*, Prentice Hall, 2003.
5. C. C. Yang, "Fuzzy Bayesian Inference". Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Orlando, Florida (1997).
6. Li Xiao-Lin, He Xiang-Dong, Yuan Sen-Miao, "Learning Bayesian networks structures from incomplete data based on extending evolutionary programming". Machine Learning and Cybernetics, Proceedings of 2005 International Conference on, Volume 4, 2039-2043, (2005).

7. G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi, *Queuing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, John Wiley & Sons, Inc, New York, USA, 1998.
8. H. Handa, O. Katai, "Estimation of Bayesian network algorithm with GA searching for better network structure", Neural Networks and Signal Processing, Proceedings of the 2003 International Conference on, Volume 1, 436–439, (2003).
9. L.M. de Campos, J.A. Gamez, S. Moral, "Partial abductive inference in Bayesian belief networks - an evolutionary computation approach by using problem-specific genetic operators", Evolutionary Computation, IEEE Transactions on, Volume 6, Issue 2, 105-131, (2002).
10. S. Shetty, M. Song, "Structure learning of Bayesian networks using a semantic genetic algorithm-based approach", Information Technology: Research and Education, 2005. ITRE 2005. 3rd International Conference on, 454-458, (2005).
11. G. Cooper, E. Herskovitz, "A Bayesian Method for the Induction of Probabilistic Networks from Data", Machine Learning, 9, 309–347, (1992).
12. R. S. Pindyck, , D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, Irwin/McGraw-Hill, 1998.
13. J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.
14. J. F. Hair Jr., R. E. Anderson, R. L. Tatham, W. C. Black, *Multivariate data analysis*, Prentice-Hall, 1998.