Algorithm for Graphical Bayesian Modeling Based on Multiple Regressions

Ádamo L. de Santana, Carlos Renato L. Francês, João C. Weyl Costa

Laboratory of High Performance Networks Planning, Federal University of Para
R. Augusto Correa, 01, 66075-110, Belem, PA, Brazil
{adamo,rfrances,jweyl}@ufpa.br

Abstract. One of the main factors for the knowledge discovery success is related to the comprehensibility of the patterns discovered by applying data mining techniques. Amongst which we can point out the Bayesian networks as one of the most prominent when considering the easiness of knowledge interpretation achieved. Bayesian networks, however, present limitations and disadvantages regarding their use and applicability. This paper presents an extension for the improvement of Bayesian networks, incorporating models of multiple regression for structure learning.

1 Introduction

Bayesian networks stand as one of the best computational intelligence techniques among the existing paradigms, being nowadays one of the best methods for treating uncertainty in the field of artificial intelligence [1]. Particularly due to their exceptional analytical properties to represent domains, correlate and study the dependences among its variables, allowing a more easily visualization and understanding of the relations among the variables, consisting on a decisive factor and of great value for the representation and analysis of the domain by the users.

However, just like any other computational algorithm, the Bayesian networks also present limitations and disadvantages, regarding their use as well as their applicability. Among these restrictions, we can point out: the difficulty to correlate variables considering the time factor, or yet, by the absence of models that would allow a deeper the use of its information and results, such as establishing the optimal combination of parameters to achieve a certain requirement. Such studies are of utter importance, given that they constitute on problems and needs observed on real world domains, and which are, eventually, crucial for the decision making process.

This paper presents a new and optimized method for learning the graphical representation of Bayesian networks, in which the creation of the network and the correlation analysis among the variables are made applying multiple regression models.

The paper is organized as follows: in section 2, some related works to the structure learning of Bayesian networks are shown. In section 3, the algorithm for structure

learning based on multiple regression is presented. The final remarks of the paper are presented in section 4.

2 Previous and related works

This section we will present some of the works found in literature and that also served as basis as well as comparison for the studies presented in this paper. The works are also divided here according to the fundaments of their approaches: whether it is based on the graphical structure learning of the network; on the search for the best configuration, that is, the set of actions or inferences and furthermore its singular values to achieve a specific state; or the temporal analysis for Bayesian networks.

The graphical learning, the construction of a Bayesian network involves the learning of the network structure and the definition of the probabilities associated with its variables. This process can be done directly with the help of experts in the studied domain or automatically, with learning algorithms, which we will focus here. The learning algorithms can be classified as being constraint based, where the structure is obtained by identifying the dependencies among the variables; or through a search and score of the best network structure.

Here, the search and score approach will be used for the learning of the network topology. The search and score works searching through the space of possible existing structures, starting from a graph with no arcs and adding new ones, calculating a score for the given structure until no new arc can be added.

In [2] a search and score method to induce Bayesian networks is proposed, using both fuzzy systems and genetic algorithms. It is proposed a scoring metric based on the evaluation of different quality criteria, which is computed by the fuzzy system; using the genetic algorithm as means to search through the space of possible structures, which has also been applied to the learning of Bayesian networks [3].

The fuzzy system uses as input metrics the Bayesian measure, the minimum description length principle [4], Akaike information criteria [5], and the estimated classification accuracy of the network; thus providing the quality of the network as output. The genetic algorithm is used to search the possible network structures.

Comparatives as to the algorithm performance with well-known algorithms (BayesN [6], Bayes9 [7], Tetrad [8] and K2 [9]), which will also be presented as comparative in section 3, are also shown.

The use of this approach brings however some limitations such as the fact of it being sensitive to the selection of the initial population (for the genetic algorithm) as well as for the different membership functions (for the fuzzy system).

Other recent methods implemented for the learning of the Bayesian graphical structure, usually based on hybrid models can also be seen in [10], [11] and [12], each with its own metric of scoring and evaluation: use of (semantic) crossover and mutation operators to help the evolution process, penalty measure, and Minimum Description Length metric, respectively; [12] proposal however does not involve a need for a complete ordering of the variables as input. Further use of genetic algorithms can also be seen in [13] and [14].

In [15], the use of a previous ordering of the variables is also studied, proposing a multi-phase approach for the graphical learning based on the use of distinct but easy to implement algorithms, which involves a search method for optimal parents to build the structure, followed by a method to eliminate existing cycles in the graph and an finally an evaluation of the network using structural perturbation.

Aside from the ordering of nodes, the dataset (here we will treat only with fully observed cases) size is also an important aspect when considering the network quality and convergence speed of the algorithm. Especially since it is NP-hard [16], exponentially increasing the searching space with the number of variables.

A more thorough overview on the techniques and algorithms for the learning of Bayesian networks can be seen in [17].

3 Structure learning based on multiple regressions

The algorithm presented here searches for the best configuration, amongst the space of possible structures, for the construction of a Bayesian network from the analysis of existing dependences and independences between the variables. The algorithm uses the search and score method, analyzing all the possible graphical combinations that can be set from the variables of the domain. It will be assumed here, at first, the need for an ordinance of the variables; which, though some recent algorithms work without this need, they are not, usually, very efficient [17].

The structure learning is an important problem to be studied, motivated by the fact that the search space of possible structures increases exponentially with the number of variables of the model. This exponential growth can be calculated as follow [18]:

$$G(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} G(n-i) \qquad (\mathbf{1})$$

Equation (2) calculates the number of possible directed acyclic graphs G, that can be formed with a number of n variable. Table 1 [19] presents the number of possible graphs (values of G) as n increases.

Table 1. Values for G obtained as n increases

n	$G(n)$
1	1.0×10^{0}
2	3.0×10^{0}
5	2.9×10^{4}
10	4.2×10^{18}
20	2.3×10^{72}
50	7.2×10^{424}
100	1.1×10^{1631}

Here, the analysis for the search of the best Bayesian network that represents the domain is made with the use of multiple regressions [20] [21]. The technique of multiple regression denotes a specific model of multivariate analysis.

The models of multivariate analysis are used to adequately study the multiple relations existent among the variables of a domain, in order to obtain a more complete and realistic understanding in the decision making process [20]. With the use of multiple regressions the changes in the dependent variable can be predicted, in response to the changes in the independent variables.

The search method of the algorithm follows from the ordinance of the variables, where for each attribute X_i the possible dependencies of the variable with its precedents are examined (variables parents - Pa_i), adding arcs between them and verifying the quality of the network created according to its score; continuing, as follows, with the search of another attribute, that added to the previous one(s) would increase the score of the network.

The validation of the network, created by each new added arc, is made through regressions, that can be single (when analyzing the relation with only one variable) or with multiple variables, as it is usually applied.

This algorithm of Multiple Regressions for Structure Learning (MRSL) attributes the score of each network through the value found by the adjusted coefficient of each regression ($\overline{R}^2$); which is obtained as described next.

3.1 Modeling and structure of the algorithm

Assuming a database D with n records and i number of variables, we are searching for the best Bayesian network structure B_S for it. We denote the target variable that we are analyzing as Y_i, the k variables candidates for parents as X_{iA}, the A_k parameters to be estimated and the random errors as u_i, the generalized formula of the multiple regression model can be specified as follow:

$$Y_i = A_0 + A_1 X_{1i} + A_2 X_{2i} + \cdots + A_k X_{ki} + u_i \qquad (2)$$

The general system of the multiple regression can then be seen as a matricial system and represented according (3).

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} \cdots X_{1k} \\ 1 & X_{21} & X_{22} \cdots X_{2k} \\ \vdots & \vdots & \vdots \quad \vdots \\ 1 & X_{n1} & X_{n2} \cdots X_{nk} \end{bmatrix} \times \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_k \end{bmatrix} + \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_k \end{bmatrix} \qquad (3)$$

This way, $Y = XA + u$, where: Y is a column vector, with dimension $n \times 1$; X is a matrix of size $n \times k$, that is, with *n* observations and *k* variables; with the first column representing the intercept A_0; A is a vector with $k \times 1$ unknown parameters; and u is a vector with $n \times 1$ disturbances.

This specification intends to generate the parameters of vector *A*; which can be estimated as follow:

$$A = \left(X^t X\right)^{-1} \times X^t Y \quad (4)$$

With the values of A, the value of the regression coefficient (R^2) can then be calculated according to:

$$R^2 = \frac{A^t \left(X^t X\right) A - n\overline{Y}^2}{y^t y} \quad (5)$$

Where $\overline{Y}$ is the mean value of variable Y, and y is obtained by the subtraction of Y by $\overline{Y}$. And thus calculating its adjusted value by:

$$\overline{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k}\right) \quad (6)$$

In the same way, the absence of dependencies ($Pa_i = \phi$) for the variable in question is assigned when obtained, for each possible relation of dependence, a value close to or below zero for $\overline{R}^2$.

Another important aspect is regarding the relevance of the inclusion of new variables in the model. This analysis is made in order to verify whether or not the inclusion of one or more arcs for the variable is indeed relevant for the model, even though with this inclusion a higher $\overline{R}^2$ might be obtained. The analysis of this aspect is due, in particular, to the fact of being verified during the evaluation tests of the algorithm, that the $\overline{R}^2$ obtained for the best Pa_i configuration for the variable X_i with a number of arcs x was very close to the one achieved by the best Pa_i configuration with a number of arcs $x+1$. The same behavior was also observed when comparing the latter with the value obtained with a number of arcs $x+2$, and so on.

In order to provide the analysis relevance to be generally applicable for datasets disregarding their sizes, the F test, whose formula is presented below, was used to evaluate the contribution of the new variables added to the model.

$$F = \frac{\left(R_U^2 - R_R^2\right)/m}{\left(1 - R_U^2\right)/(n-k)} \quad (7)$$

Where R_U^2 and R_R^2 are the $\overline{R}^2$ values obtained for the unrestricted (with the inclusion of the new variables) and restricted (without the inclusion of the variables) regressions, respectively, and m is the number of variables added to the model. The statistic distribution F follows with m and $n-k$ degrees of freedom. If the F statistics presents a value different from zero, the added variable is accepted as a possible *parent* variable.

3.2 Optimizations studied

The MRSL algorithm acts in an optimized way, with respect to performance, when compared with other existing learning algorithms in the literature. It works directly without considering the number of states of the variables, not suffering from any combinatory impact that can be implied by them in the search and score of the best net-

work structure. In order to further optimize the performance of the algorithm some considerations and heuristics can also be adopted. In the very first iterations of each variable, a control can be included in order to decrease the combinatory space to be covered and, consequentially its execution time.

Firstly, from the values obtained in the correlations of degree one (number of parents equals to one) of the variable X_i with its precedents ($X_1, \ldots, X_{i-1}$), it is already possible to observe which, amongst the variables, presents a higher level of correlation with X_i. This is important as, whenever a new arc can be added in the network structure, the new combination of parents found will have as component, compulsorily, the attribute (or combination of attributes, if the number of arcs is higher than 2) found previously. Thus, only the future regressions for models having as component the nodes assigned previously will be made.

Not only that, but if in the correlations coefficients $\overline{R}^2$ present values with low significance or close to or below zero, the search for a better configuration and admission of new arcs can cease, as the following ones will also obey the same trend.

Another aspect that can be manipulated, is the indication by the user specialist in the domain of a minimum degree of significance to be verified for the admission of a new arc in the structure.

As described thus far, the algorithm fixates the attributes as the dependent variables and study its relations with the preceding variables according to the ordering set as input, also applying strategies to diminish the combinatory search space. The causality analysis of the algorithm will be further focused now, presenting a means to model the network without the need for a previous ordering of the variables.

3.3 Causality

The regression model implemented studies the correlation and dependence among the variables, having the ordering of the variables as main aspect to attribute the direction of the dependence. The sequential order of the variables would be, at this point, very important, given that, by itself, the dependence relations among the variables does not necessarily implies on a causal relation.

The study of causality is then applied with the analysis of Granger [22]. Considering the hypothesis that *X* can cause *Y* ($X \rightarrow Y$), the test is established between a restricted regression, in which *Y* is a function of only its past values; and an unrestricted regression, where *Y* is a function of the values of *Y* and *X*. The functions for the restricted and unrestricted regressions are represented in (8) and (9), respectively.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \ldots + \alpha_s Y_{t-s} + \varepsilon \quad (\mathbf{8})$$

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \ldots + \alpha_s Y_{t-s} + \beta_1 X_{t-1} + \ldots + \beta_s X_{t-s} + \varepsilon \quad (\mathbf{9})$$

The hypothesis of causality from the analysis of both regressions is then made with the *F* test (Equation 7). So that, if the calculated value of *F* is higher than its critical values [21] the causality from *X* to *Y* is accepted.

With the advent of the causality analysis, it is possible to establish a new search heuristic, incorporating it with the model detailed on section 3.1, without the need of a previous ordering of the variables. Initially verifying the variables from which the attribute present dependences, and then studying the direction of its causality.

3.4 Performance evaluation

The evaluation of the model was made considering two aspects: the quality of the Bayesian network found by the algorithm, that is, the representativity of the network regarding the domain; and its computational performance.

For comparing the analysis regarding the quality of the generated network, the *Chest Clinic* [23] database was used as application example (usually known as *Asia*), which denotes a problem of a fictitious medical diagnosis, of whether a patient has tuberculosis, lung cancer or bronchitis, based on his X-Ray, dyspnea, visit to Asia and smoking status. The database possesses 8 binary variables and its Bayesian network presents 8 arcs connecting them (Figure 1).

The database was submitted to our learning algorithm to obtain the structure of the Bayesian network. For a quality comparative of the generated network, results from other search and score algorithms were used; the algorithms used were the K2, created by Cooper and Herskovitz (1992), which to this date is still a great reference among the existing algorithms for learning of Bayesian networks, being one of the most trustworthy and successful learning algorithms [24]. The results of the following algorithms existing in literature were also used as comparative for the quality of the generated network for the *Asia* database: Tetrad [8], Bayes9 [7], BayesN [6] and Genetic-Fuzzy [2].

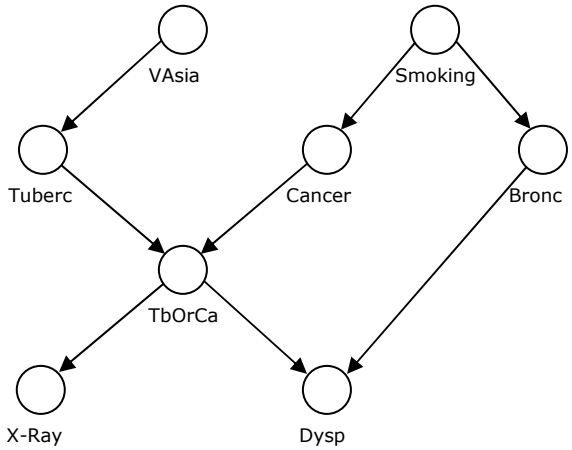

Fig. 1. Bayesian network of the Asia database

Table 2 compares the result achieved by our algorithm (MRSL) with the one from the original (the *golden network*, as it is also referenced in literature) Bayesian network of the Asia database, as well as the results obtained by others five existing algorithms in the literature (K2, Tetrad, Bayes9, BayesN and Genetic-Fuzzy). The *Total*

column presents the number of arcs found by each algorithm. The column *Correct* contains the number of arcs that were correctly found. The column *Additional* presents the number of arcs that were found and that are not in the original network. And the column *Absent* presents the number of arcs that were not found and are present in the original network.

Table 2. Comparative of the results obtained for the Asia database

Algorithms	Total	Correct	Additional	Absent
MRSL	8	8	0	0
K2	8	7	1	1
Genetic-Fuzzy	9	8	1	0
BayesN	8	5	3	3
Bayes9	4	4	0	4
Tetrad	4	4	0	4

Our algorithm presented results better than the other algorithms, finding an identical structure to the golden network of the *Asia* database, with no additional nor absent arcs. Being followed by the results presented by the K2 and Genetic-Fuzzy algorithms.

Table 3 presents the results for the structure learning of the *Asia* network with datasets of different sizes; now studying the capacity of the algorithm for obtaining the original structure of the network when different volumes of data are available.

Table 3. Results for the structures obtained according to the amount of data available

Num. of records	Total	Correct	Additional	Absent
100	9	7	2	1
200	7	7	0	1
500	8	7	1	1
1,000	8	8	0	0
5,000	8	8	0	0
10,000	8	8	0	0

The results obtained (Table 3) showed that the algorithm has a good capacity of learning, even when working with a small amount of data, finding the original structure of the network with a data amount of 1,000 records onwards.

For the performance evaluation of the algorithm, the analysis was made using as testbed experiment the model presented by the *Asia* network, which is composed of 8 variables and 1,000 records, comparing the results obtained with the ones presented by the K2 algorithm.

The tests made here seeks to verify the performance of the algorithm using as parameter the discretized states of the database variables, that is, the number of possible states that each attribute can assume. The performance tests were made analyzing the execution time for both algorithms over the database, with the attributes (initially binary) discretized from the two initial states until a maximum of ten. The obtained results (Table 4) denote the execution times of the algorithms without considering the time spent for reading the database into the memory.

Table 4. Execution times (in seconds) obtained by the algorithms

Num. of states	MRSL	K2
2	0.08	0.1
3	0.08	0.14
4	0.08	0.24
5	0.08	0.51
6	0.08	1.35
7	0.08	3.51
8	0.08	9.10
9	0.08	20.05
10	0.08	44.48

Tables 5 and 6 present the same tests, now also considering an increase in the number of records of the database to 5,000 and 10,000 respectively. Table 7 presents the values, considering only the discretization space of 10, for a better visualization of the gradual behavior in the increase of the execution time between the algorithms.

Table 5. Execution times (in seconds) obtained by the algorithms for 5,000 records

Num. of states	MRSL	K2
2	0.42	0.48
3	0.42	0.68
4	0.42	0.93
5	0.42	1.32
6	0.42	2.26
7	0.42	4.53
8	0.42	10.27
9	0.42	21.32
10	0.42	45.42

Table 6. Execution times (in seconds) obtained by the algorithms for 10,000 records

Num. of states	MRSL	K2
2	0.84	0.96
3	0.84	1.33
4	0.84	1.78
5	0.84	2.32
6	0.84	3.39
7	0.84	5.81
8	0.84	11.74
9	0.84	22.87
10	0.84	47.05

Table 7. Execution times (in seconds) obtained with a number of states set to 10

Num. of records	MRSL	K2
1000	0.08	44.48
5000	0.42	45.42
10.000	0.84	47.05

As it could be verified by the obtained results, the structure learning algorithm based on multiple regressions outperforms on both aspects analyzed: with respect to the quality of the Bayesian network induced by the algorithm as well as in its computational performance. The algorithm uses in its structure statistical models with a fundamental theory, especially concerning the prediction and correlation analysis of the variables; and that, due to its nature, works in an optimized way, improving the performance as the number of states assumed for the variables increases; which is a common characteristic for databases that represent real world domains.

4 Final remarks

The possibility to represent graphically the structure of the patterns obtained from the data, as well as the exploratory character of the analysis allowed by the Bayesian networks, enables to indicate more deeply the relationship between the variables of a domain, favoring the increase of the comprehensibility of the discovered patterns, as well as the identification of the usefulness and relevance of these patters.

In this paper, a new technique for modeling the graphical structure of a Bayesian network was presented, using multiple regressions as the method for analyzing the correlations among the attributes. The algorithm proposed implements a method for the structure learning which quantifies, based on mathematical models of regression, the level of the existing dependence from the variables. Initiating with a network without arcs and adding them in accordance with the identified correlations, as the search space of possible existing structures for the network is covered.

The tests carried to study the performance of the algorithm presented promising results in the observed aspects: regarding the quality of the generated network, when comparing with other learning algorithms, obtaining a representative structure of the Bayesian network, even when a reduced amount of data is available and; regarding its execution performance, achieving better execution times for bases with increasing volumes of data and, particularly, for its method of treating the attributes, discrete or continuous, not having its performance compromised as the number of possible states of the variables increases.

References

1. Huang, H., Song, H., Tian, F., Lu, Y. e Wang, Q. A comparatively research in incremental learning of Bayesian networks, Intelligent Control and Automation. Fifth World Congress on, Volume 5, (2004) 4260-4264.
2. Morales, M. M., Dominguez, R. G., Ramirez, N. C., Hernandez, A. G. e Andrade, J. L. J. A method based on genetic algorithms and fuzzy logic to induce Bayesian networks, Computer Science, 2004. ENC 2004. Proceedings of the Fifth Mexican International Conference in, (2004) 176-180.
3. Larrañaga, P. Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters, IEEE Journal on Pattern Analysis and Machine Intelligence, 18(9), USA (1996) 912-926.

4. Rissanen, J. Modeling by shortest data description, Automatica, vol. 14 (1978) 465-471.
5.Akaike, H. A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (6), (1974) 716-723.
6.Morales, M. M., Ramírez, N. C., Andrade, J. L. J. e Domínguez, R. G. Bayes-N: an algorithm for learning Bayesian networks from data using local measures of information gain applied to classification problems, In: MICAI 2004: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence, Vol.2972, Springer Verlag, Germany (2004).
7.Ramírez, N. C. Building Bayesian networks from data: a constraint based approach, PhD Thesis, University of Sheffield (2001).
8.Spirtes, P. R., Shcheines, R. e Clark, G. TETRAD II: Tools for Discovery. Lawrence Erlbaum Associates, Hillsdale, NJ., USA (1994).
9.Cooper, G. e Herskovitz, E. A Bayesian Method for the Induction of Probabilistic Networks from Data, Machine Learning, 9, (1992) 309-347.
10.Shetty, S.; Song, M. Structure learning of Bayesian networks using a semantic genetic algorithm-based approach, Information Technology: Research and Education, 2005, ITRE 2005. 3rd International Conference on, (2005) 454-458.
11.Gang Li, Fu Tong, Honghua Dai, Evolutionary structure learning algorithm for Bayesian network and Penalized Mutual Information metric, Data Mining, ICDM 2001, Proceedings IEEE International Conference on, (2001) 615-616.
12.Xiao-Lin Li, Sen-Miao Yuan, Xiang-Dong He, Learning Bayesian networks structures based on extending evolutionary programming, Machine Learning and Cybernetics, Proceedings of 2004 International Conference on, Volume 3, (2004) 1594-1598.
13.Gamez, J. A.; de Campos, L. M.; Moral, S. 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, (2002) 105-131.
14.Handa, H.; Katai, O. 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, (2003) 436-439.
15.Peng, H.; Ding, C. Structure search and stability enhancement of Bayesian networks, Data Mining, ICDM 2003. Third IEEE International Conference on, (2003) 621-624.
16.Chickering, D. M.; Heckerman, D.; Meek, C. Large-Sample Learning of Bayesian Networks is NP-Hard, Journal of Machine Learning Research, Volume 5, (2004) 1287-1330.
17.Cheng, J., Bell, D. e Liu, W. Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory, Artificial Intelligence. Volume 137, No.1-2, (2002) 43-90.
18.Robinson, R. W. Counting unlabeled acyclic digraphs, Proceedings of the Fifth Australian Conference on Combinatorial Mathematics, (1976) 28-43.
19.Herskovits, E. Computer-Based Probabilistic Networks Construction, Ph.D. Thesis, Medical Information Sciences, University of Pittsburgh (1991).
20.Hair, J. F. JR., Anderson, R. E., Tatham, R. L. e Black, W. C. Multivariate data analysis, Prentice-Hall (1998).
21.Rice, J. A. Mathematical Statistics and Data Analysis, 2nd Edition, Duxbury Press (1995).
22.Santana, A. C. Quantitative Methods in Econometry, UFRA (2003).
23.Lauritzen, S. L. e Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems, In: J. Royal Statistics Society B, 50(2), (1988) 157-194.
24.Yang, S. e Chang, K. Comparison of Score Metrics for Bayesian Network Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part A 32, (2002) 419-428.