Induction Learning Techniques Applied to Bayesian Networks Optimization

P. Britos, P. Felgaer, D. Rodríguez and R. García-Martínez

Software & Knowledge Engineering Center. Graduate School. Buenos Aires Institute of Technology
Intelligent Systems Laboratory. School of Engineering. University of Buenos Aires

rgm@itba.edu.ar

(Paper received on July 07, 2007, accepted on September 1, 2007)

Abstract: A bayesian network is a directed acyclic graph in which each node represents a variable and each arc a probabilistic dependency, they are used to provide: a compact form to represent the knowledge, and flexible methods of reasoning. Obtaining a bayesian network from data is a learning process that is divided in two steps: structural learning and parametric learning. In this paper, we define an automatic learning method that optimizes the bayesian networks applied to classification, using a hybrid method of learning that combines the advantages of the induction techniques of the decision trees (TDIDT - C4.5) with those of the bayesian networks. the resulting method is applied to prediction in health domain.

1 Introduction

The learning can be defined as "any process through as a system improves its efficiency". The ability to learn is considered as a central characteristic of the "intelligent systems" [Fritz *et al.*, 1989; García-Martínez & Borrajo, 2000], and for this a lot of effort and dedication was invested in the investigation and the development of this area. The development of the knowledge based systems motivated the investigation in the area of the learning with the purpose of automating the process of knowledge acquisition which considers one of the main problems in the construction of these systems.

The data mining [Perichinsky & García-Martínez, 2000; Perichinsky *et al.*, 2000; Perichinsky *et al.*, 2001; Perichinsky *et al.*, 2003] are the set of techniques and tools applied to the non-trivial process of extract and present/display implicit knowledge, previously unknown, potentially useful and humanly comprehensible, from large data sets, with object to predict of automated form tendencies and behaviors; and to describe of automated form models previously unknown, [Chen *et al.*, 1996; Mannila, 1997; Piatetski-Shapiro *et al.*, 1991]. The term Intelligent data mining, [Evangelos & Han, 1996; Michalski *et al.*, 1998] is the application of automatic learning methods, [Michalski *et al.*, 1983; Holsheimer & Siebes, 1991], to discover and enumerate present patterns in the data. For these, they were developed a great number of methods of analysis of data based on the statistic [Michalski *et al.*, 1982]. In the time

© H. Sossa, R. Barrón and E. Felipe (Eds.)
Special Issue in Neural Networks and Associative Memories
Research in Computing Science 28, 2007, pp. 225-234

in which the amount of information stored in the databases was increased, these methods began to face problems of efficiency and scalability and is here where appears the concept of data mining. One of the differences between a traditional analysis of data and the data mining are that first it supposes that the hypotheses already are constructed and validated against the data, whereas the second supposes that the patterns and hypotheses automatically are extracted of the data.

The tasks of the data mining can be classified in two categories: descriptive data mining and predictive data mining [Piatetsky-Shapiro *et al.*, 1996; Han, 1999]; some of the most common techniques of data mining are the decision trees (TDIDT), the production rules and neuronal networks. On the other hand, an important aspect in the inductive learning, is the one to obtain a model that represents the knowledge domain and that is accessible for the user, in particular, is important to obtain the dependency data between the variables involved in the phenomenon, in the systems where it is desired to predict the behavior of some unknown variables based on certain known variables, a representation of the knowledge that is able to capture this information on the dependencies between the variables is the bayesian networks [Cowell *et al.*, 1990; Ramoni & Sebastiani, 1999].

A bayesian network is a directed acyclic graph in which each node represents a variable and each arc a probabilistic dependency, in which specifies the conditional probability of each variable given its parents; the variable at which it points the arc is dependent (cause-effect) of the variable in the origin of this one. The topology or structures of the network gives information on the probabilistic dependencies between the variables but also on conditional independences of a variable (or set of variables) given another or other variables, these independences simplify the representation of the knowledge (less parameters) and the reasoning (propagation of the probabilities).

Obtaining a bayesian network from data is a learning process that is divided in two phases: the structural learning and the parametric learning [Pearl, 1988]. First of them, consists of obtaining the structure of the bayesian network, that means, the relations of dependency and independence between the involved variables. The second phase has the purpose of obtain the a priori and conditional probabilities from a given structure.

The bayesian networks [Pearl, 1988] are used in diverse areas of application like medicine [Beinlich *et al.*, 1989], sciences [Bickmore & Timothy, 1994; Breese & Blake, 1995], and economy [Ezawa *et al.*, 1995]. They provide a compact form to represent the knowledge and flexible methods of reasoning -based on the probabilistic theories- able to predict the value of non-observed variables and to explain the observed ones. Some characteristics of the bayesian networks are that they allow to learn dependency and causality relations, they allow to combine knowledge with data [Heckerman *et al.*, 1995; Diaz & Corchado, 1999] and they can handle incomplete databases [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996].

The bayesian networks are designed to find the dependence and independence relations between all the variables that conform the study domain, this allows to make predictions on the behavior of anyone of the unknown variables based on the values of the well-known variables; this estimates that any variable of the database can behave as incognito or as evidence according to the case.

Many practical tasks can be reduced to classification problems: medical diagnosis and pattern recognition are only two examples.
The bayesian networks can make the classification task -a particular case of prediction- that it is characterized to have a single variable of the database (class) that is desired to predict, whereas all the others are the data evidence of the case that is desired to classify. A great amount of variables in the database can exist; some of them directly related to the class variable but also other variables that have not direct influence on the class.
In this work, a method of automatic learning is defined that helps in the pre-selection of variables, optimizing the configuration of the bayesian networks in classification problems.

2 Methodology

In order to solve the problem of the bayesian networks applied to the classification task, in this work we use a hybrid learning method that combines the advantages of the induction techniques of the decision trees (TDIDT – C4.5) with those of the bayesian networks. For it, we integrate to the process of structural and parametric learning of the bayesian networks, a previous process of pre-selection of variables. In this process, it is chosen from all the variables of the domain, a subgroup with the purpose of generating the bayesian network for the particular task of classification and this way, optimizing the performance and improving the predictive capacity of the network.
The method for structural learning of bayesian networks is based on the algorithm developed by Chow and Liu (1969) to approximate a probability distribution by a product of probabilities of second order, which corresponds to a tree. The joint probability of variables can be represented like:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i) P(X_i | X_{j(i)}) \quad (1)$$

where $X_{j(i)}$ it is the cause or parent of X_i.

Consider the problem like one of optimization and it is desired to obtain the structure of the tree that comes near more to the "real" distribution. A measurement of the difference of information between the real distribution (P) and the approximate one (P^*) is used:

$$I(P, P^*) = \sum_x P(X) \log(P(X) / P^*(X)) \quad (2)$$

Then the objective is to minimize I. A function based on the mutual information between pairs of variables is defined as:

$$I(X_i,X_j)=\sum_x P(X_i,X_j)\log P(X_i,X_j)/P(X_i)P(X_j)) \qquad (3)$$

Chow (1968) demonstrates that the information difference is a function of the negative of the sum of the mutual information (weights) of all the pairs of variables that constitute the tree. Reason why to find the more similar tree is equivalent to find the tree with greater weight. Based on that, the algorithm to determine the optimal bayesian network from data is the following one:

1. Calculate the mutual information between all the pairs of variables (n (n - 1) / 2).
2. Sort the mutual information in descendent order.
3. Select the arc of greater value as the initial tree.
4. Add the next arc while it does not form cycles. If it is thus, reject.
5. Repeat (4) until all the variables are included (n - 1 arcs).

Rebane and Pearl (1989) extended the algorithm of Chow and Liu for poly-trees. In this case, the joint probability is:

$$P(X)=\prod_{i=1}^{n} P(X_i \mid X_{j1(i)},X_{j2(i)},...,X_{jm(i)}) \qquad (4)$$

where $\{X_{j1(i)},X_{j2(i)},...,X_{jm(i)}\}$ is the set of parents for the variable X_i.

In order to compare the results obtained when applying the complete bayesian networks (RB-Complete) and the preprocessed bayesian networks with induction algorithms C4.5 (RB-C4.5), we used the databases "Cancer" and "Cardiology" obtained at the Irving Repository of Machine Learning databases of the University of California [Murphy & Aha] and the database "Dengue" obtained at the University of Buenos Aires [Carbajo *et al.*, 2003].
Table 1 summarizes these databases in terms of amount of cases, classes, variables (excluding the classes), as well as the amount of resulting variables of the preprocessing with the induction algorithm C4.5.
The methodology used to carry out the experiments with each one of the evaluated databases, is detailed next.

1. Divide the database in two. One of control or training (approximately 2/3 of the total database) and the another one of validation (with the remaining data)
2. Process the control database with the induction algorithm C4.5 to obtain the subgroup of variables that will conform the RB-C4.5
3. Repeat for 10%, 20%, ..., 100% of the control database
   3.1. Repeat 30 times, by each iteration

3.1.1. Take randomly X% from the control database according to the percentage that corresponds to the iteration
3.1.2. With that subgroup of cases of the control database, make the structural and parametric learning of RB-Complete and the RB-C4.5
3.1.3. Evaluate the predictive power of both networks using the validation database
3.2. Calculate the average predictive power (from the 30 iterations)
4. Graph the predictive power of both networks (RB-Complete and RB-C4.5) based on the cases of training

The step (1) of the algorithm makes reference to the division of the database in the control and the validation ones. In most cases, the databases obtained from the mentioned repositories were already divided.
For the pre-selection of variables by the induction algorithms C4.5 of the step (2), we introduced each one of the control databases in a decision trees TDIDT generating system. From there, we obtained the decision trees that represent each one of the analyzed domains. The variables that integrate this representation conform the subgroup that were considered for the learning of the preprocessed bayesian networks.
Next (3) a ten iteration process begins, in each one of these iterations processed 10%, 20%, 100% of the control database for the networks structural and parametric learning. This way, could be analyzed not only the difference in the predictive capacity of the networks, but also how evolved this capacity when we learn with greater amount of cases.
The objective of the repetitive structure of the step (3.1) is to minimize the accidental results that do not correspond with the reality of the model in study. It is managed to minimize this effect, taking different data samples and average the obtained values.
In the steps (3.1.x) it is made the structural and parametric learning of the RB-Complete and the RB-C4.5 from the subgroup of the control database (both networks are obtained from the same subgroup of data). Once obtained the network, it is come to evaluate the predictive capacity with the validation databases. This database is scan and for each row, all the evidence variables are instantiated and it is analyzed if the inferred class by the network corresponds with the indicated one in the file. Since the bayesian network does not make excluding classifications (it means that it predicts for each value of the class the probability of occurrence), is considered like the inferred class, the class with the greater probability. The predictive capacity corresponds to the percentage of cases classified correctly respect to the total evaluated cases.
In the point (3.2) it is calculated the predictive power of the network, dividing the obtained values through all the made iterations.
Finally, in the step (4) it is come to graph the predictive power average of both bayesian networks based on the amount of training cases.

Database	Variables	Variables C4.5	Classes	Control cases	Validation cases	Total cases
Cancer	9	6	2	500	199	699

Cardiology	6	4	2	64	31	95
Dengue	11	5	4	1.414	707	2.121

Table 1 – Databases description

3 Results

The experimental results were obtained by the application of the methodology previously mentioned to each one of the test databases.
As it can be observed in Figure 1 ("Cancer" domain), the predictive power of the RB-C4.5 is superior to the one of RB-Complete throughout all its points. Also, it is possible to observe how this predictive capacity is increased, almost always, when it takes more cases of training to generate the networks. Finally, it is observed that from the 350 cases of training the predictive power of the networks become stabilized reaching its maximum level.
When analyzing the graph of Figure 2 corresponding to the database "Cardiology", also an improvement on the RB-C4.5 can be observed respect to RB-Complete. Although the differences between the values obtained with both networks are smaller that in the previous case, the hybrid algorithm presents a better approach to the reality that the other one. It is important to emphasize that in this case, the improvement level is minimize when the set of cases used for the learning process is increased.
For the database "Degue" corresponding to Figure 3, an improvement in the predictive power of the proposed network is observed. The RB-C4.5 makes the classification with a 10% better precision than the other network.

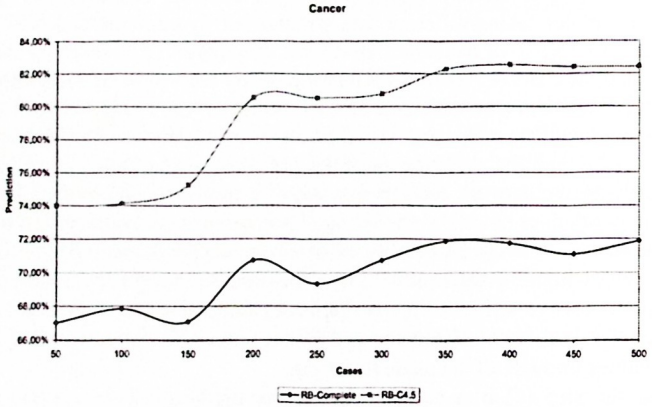

Figure 1 - Graph of the predictive power for the database "Cancer"

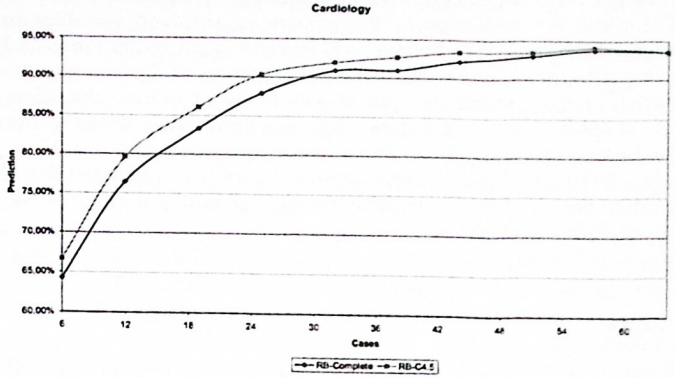

Figure 2 - Graph of the predictive power for the database "Cardiology"

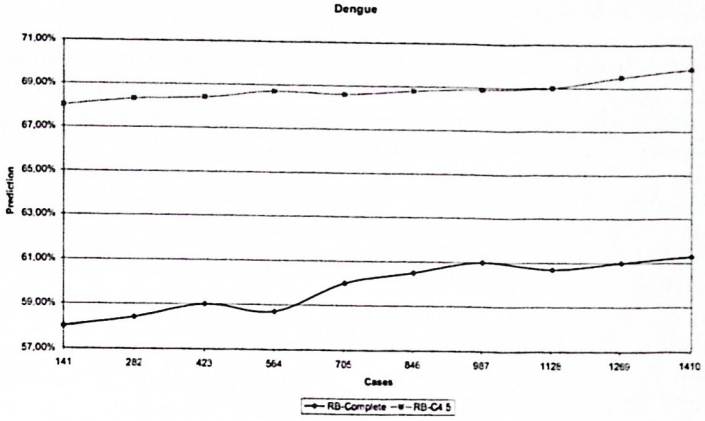

Figure 3 - Graph of the predictive power for the database "Dengue"

4 Discussion and Conclusions

As it is possible to observe, all the graphs that represent the predictive power based on the amount of cases of training are increasing. This phenomenon occurs independently of the domain of data used and the evaluated method (RB-Complete or RB-C4.5). Of the analysis of the results obtained in the experimentation, we can (experimentally) conclude that the learning hybrid method used (RB-C4.5) generates an improvement in the predictive power of the network with respect to the obtained one without making the preprocessing of the variables (RB-Complete).

In another aspect, the RB-C4.5 has a lesser amount of variables (or at the most equal) that RB-Complete, this reduction of the amount of involved variables produces a simplification of the analyzed domain, which carry out two important advantages; first, they facilitate the representation and interpretation of the knowledge removing parameters that do not concern on a direct way to the objective (classification task). Second, it simplifies and optimizes the reasoning task (propagation of the probabilities) which originates the improvement of the processing speed.
In conclusion, from the obtained experimental results, we concluded that the hybrid learning method proposed in this paper optimizes the configurations of the bayesian networks in classification tasks.

5 References

1. Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F. (1989). *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks*. In proceedings of the 2nd European Conference on Artificial Intelligence in Medicine.
2. Bickmore, Timothy W. (1994). *Real-Time Sensor Data Validation*. NASA Contractor Report 195295, National Aeronautics and Space Administration.
3. Breese, John S., Blake, Russ (1995). *Automating Computer Bottleneck Detection with Belief Nets*. Proceedings of the Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, pp 36-45.
4. Carbajo, A., Curto, S., Schweigmann, N. (2003). *Distribución espacio-temporal de Aedes aegypti (Diptera: Culicidae). Su relación con el ambiente urbano y el riesgo de transmisión del virus dengue en la Ciudad de Buenos Aires*. Departamento de Ecología, Genética y Evolución. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires
5. Chen, M., Han, J., Yu, P. (1996). *Data mining: An overview from database perspective*. IEEE Transactions on Knowledge and Data Eng.
6. Cowell, R., Dawid, A., Lauritzen, S., Spiegelhalter, D. (1990). *Probabilistic Networks and Expert Systems*. Springer, New York, NY.
7. Diaz, F., Corchado, J.M. (1999). *Rough sets bases learning for bayesian networks*. International workshop on objetive bayesian methodology, Valencia, Spain.
8. Evangelos, S., Han, J. (1996). *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, EE.UU.
9. Ezawa, Kazuo J., Schuermann, Til (1995). *Fraud/Uncollectible Debt Detection Using a Bayesian Network Based Learning System: A Rare Binary Outcome with Mixed Data Structures*. Proceedings of the Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, pp 157-166.
10. Fritz, W., García-Martínez, R., Rama, A., Blanqué, J., Adobatti, R., Sarno, M. (1989). *The Autonomous Intelligent System*. Robotics and Autonomous Systems. Elsevier Science Publishers. Holanda. Volumen 5. Número 2. Páginas 109-125.
11. García-Martínez, R., Borrajo, D. (2000). *An Integrated Approach of Learning, Planning and Executing. Journal of Intelligent and Robotic Systems*. Volumen 29, Número 1, Páginas 47-78. Kluwer Academic Press.

12.Han, J. (1999). *Data Mining.* Urban and Dasgupta (eds.), Encyclopedia of Distributed Computing. Kluwer Academic Publishers.
13.Heckerman, D., Chickering, M., Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data.* Machine learning 20: 197-243
14.Heckerman, D. (1995). *A tutorial on learning bayesian networks.* Technical report MSR-TR-95-06, Microsoft research, Redmond, WA.
15.Heckerman, D., Chickering, M. (1996). *Efficient approximation for the marginal likelihood of incomplete data given a bayesian network.* Technical report MSR-TR-96-08, Microsoft Research, Microsoft Corporation.
16.Holsheimer, M., Siebes, A. (1991). *Data Mining: The Search for Knowledge in Databases.* Report CS-R9406, ISSN 0169-118X, Amersterdam, The Netherlands.
17.Mannila, H. (1997). *Methods and problems in data mining.* In Proc. of International Conference on Database Theory, Delphi, Greece.
18.Michalski, R.S., Baskin, A.B., Spackman, K.A. (1982). *A Logic-Based Approach to Conceptual Database Analysis.* Sixth Annual Symposium on Computer Applications on Medical Care, George Washington University, Medical Center, Washington, DC, EE.UU.
19.Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (1983). *Machine learning I: An AI Approach.* Morgan Kaufmann, Los Altos, CA.
20.Michalski, R.S., Bratko, I., Kubat, M. (1998). *Machine Learning and Data Mining, Methods and Applications.* John Wiley & Sons Ltd, West Sussex, England.
21.Murphy, P.M., Aha, D.W. *UCI Repository of Machine Learning databases.* Machine-readable data repository, Department of Information and Computer Science, University of California, Irvine.
22.Pearl, J. (1988). *Probabilisic reasoning in intelligent systems.* Morgan Kaufmann, San Mateo, CA.
23.Perichinsky, G., García-Martínez, R. (2000). *A Data Mining Approach to Computational Taxonomy.* Proceedings del Workshop de Investigadores en Ciencias de la Computación. Páginas 107-110. Editado por Departamento de Publicaciones de la Facultad de Informática. Universidad Nacional de La Plata. Mayo.
24.Perichinsky, G., García-Martínez, R., Proto, A. (2000). *Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining.* CD del VI Congreso Argentino de Ciencias de la Computación. (\cacic2k\cacic\sp\is-039\IS-039.htm). Ushuaia. Octubre 2 al 6.
25.Perichinsky, G., García-Martínez, R., Proto, A., Sevetto, A, Grossi, D. (2001). *Data Mining: Supervised and Non-Supervised Intelligent Knowledge Discovery.* Proceedings del II Workshop de Investigadores en Ciencias de la Computación. Mayo. Editado por Universidad Nacional de San Luis en el CD Wicc2001:\Wiccflash\Areas\IngSoft\Datamining.pdf
26.Perichinsky, G., Servetto, A., García-Martínez, R., Orellana, R., Plastino, A. (2003). *Taxomic Evidence Applying Algorithms of Intelligent Data Minning Asteroid Families.* Proceedings de la International Conference on Computer Science, Software Engineering, Information Technology, e-Bussines & Applications. Pág. 308-315. Río de Janeiro (Brasil). ISBN 0-9742059-3-7.
27.Perichinsky, G., Servente, M., Servetto, A., García-Martínez, R., Orellana, R., Plastino, A. (2003). *Taxonomic Evidence and Robustness of the Classification*

Applying Intelligent Data Mining. Proceedings del VIII Congreso Argentino de Ciencias de la Computación. Pág. 1797-1808.

28.Piatetski-Shapiro, G., Frawley, W.J., Matheus, C.J. (1991). *Knowledge discovery in databases: an overview*. AAAI-MIT Press, Menlo Park, California.

29.Piatetsky-Shapiro, G., Fayyad, U.M., Smyth, P. (1996). *From data mining to knowledge discovery*. AAAI Press/MIT Press, CA.

30.Ramoni, M., Sebastiani, P. (1996). *Learning bayesian networks from incomplete databases*. Technical report KMI-TR-43, Knowledge Media Institute, The Open University.

31.Ramoni, M., Sebastiani, P. (1999). *Bayesian methods in Intelligent Data Analysis. An Introducction*. Pages 129-166. Physica Verlag, Heidelberg.