International Journal of Modern Physics C
Vol. 17, No. 3 (2006) 447–455

World Scientific
www.worldscientific.com

PREDICTION IN HEALTH DOMAIN USING BAYESIAN NETWORKS OPTIMIZATION BASED ON INDUCTION LEARNING TECHNIQUES

PABLO FELGAER

Intelligent Systems Lab. School of Engineering
University of Buenos Aires
Paseo Colón 850 4th Floor, South Wing, (1063) Buenos Aires, Argentina
pfelgaer@esamericas.net

PAOLA BRITOS

Software & Knowledge Engineering Center
Graduate School, Buenos Aires Institute of Technology
Av. Madero 399, (1106) Buenos Aires, Argentina
pbritos@itba.edu.ar

RAMÓN GARCÍA-MARTÍNEZ

Software & Knowledge Engineering Center
Graduate School, Buenos Aires Institute of Technology
Av. Madero 399, (1106) Buenos Aires, Argentina
rgm@itba.edu.ar

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 it 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.

Keywords: Bayes; induction learning; classification; hybrid intelligent systems.

1. Introduction

The learning can be defined as "any process through as a system improves its efficiency". The ability to learn is considered a central characteristic of the "intelligent systems",[1,2] 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.

Data mining[3–6] is the set of techniques and tools applied to the non-trivial process of extracting and presenting/displaying implicit knowledge, previously unknown, potentially useful and humanly comprehensible, from large data sets, with object to predict automated form tendencies and behaviors; and to describe automated form models previously unknown.[7–9] The term intelligent data mining[10,11] is the application of automatic learning methods[12,13] to discover and enumerate present patterns in the data. For these, a great number of data analysis methods were developed, based on the statistic.[14] In the time in which the amount of information stored in the databases was increased, these methods began to face problems of efficiency and scalability. This is where the concept of data mining appears. One of the differences between a traditional analysis of data and the data mining is that the first supposes that the hypotheses are already constructed and validated against the data, whereas the second supposes that the patterns and hypotheses are automatically extracted from the data.

The tasks of the data mining can be classified in two categories: descriptive data mining and predictive data mining;[15,16] 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 to obtain a model that represents the knowledge domain that is accessible for the user, it is particularly important to obtain the dependency data between the variables involved in the phenomenon; in the systems that need 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.[17,18]

A Bayesian network is a directed acyclic graph in which each node represents a variable and each arc represents a probabilistic dependency which specifies the conditional probability of each variable given its parents; the variable to which the arc points to is dependent (cause-effect) on 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 into two phases: the structural learning and the parametric learning.[19] The first 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[19] are used in diverse areas of application like medicine,[20] sciences,[21,22] and economy.[23] 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 are able to learn the dependency and causality relations, able to combine knowledge with data,[24,25] and can handle incomplete databases.[26–28]

The Bayesian networks represent the dependence and independence relations between all the variables that form the study domain. Base on this, probabilistic reasoning methods are used to make predictions of the value of any unknown variables based on the values of the known variables.

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 we desire to predict, whereas all the others are the data evidence of the case that we desire 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. This method 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 the process of structural and parametric learning of the Bayesian networks to a previous pre-selection process of variables. In this process, from all the variables of the domain we chose 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 the second order, which corresponds to a tree. The joint probability of variables can be represented like:

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

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 closer to the "real" distribution. A measurement of the difference of information between the real distribution (P) and the approximate

one (P^*) is used:

$$I(P, P^*) = \sum_x P(X) \log(P(X)/P^*(X)) \,. \tag{2}$$

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

$$I(X_i, X_j) = \sum_{i=1} P(X_i, X_j) \log(P(X_i, X_j)/P(X_i)P(X_j)) \,. \tag{3}$$

Chow (1968) demonstrates that 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 as follows:

(i) Calculate the mutual information between all the pairs of variables $(n(n-1)/2)$.
(ii) Sort the mutual information in descendent order.
(iii) Select the arc with greater value as the initial tree.
(iv) Add the next arc if it does not form cycles. Reject if it does.
(v) Repeat (iv) 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:

$$P(X) = \prod_{i=1}^{n} P(X_i | X_{j1(i)}, X_{j2(i)}, \ldots, X_{jm(i)}) \,, \tag{4}$$

where $\{X_{j1(i)}, X_{j2(i)}, \ldots, X_{jn(i)}\}$ 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 "Cancer" and " Cardiology" databases obtained from the Irving Repository of Machine Learning databases of the University of California[29] and the database "Dengue" obtained at the University of Buenos Aires.[30]

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.

Table 1. Databases description.

Database	Variables	Variables C4.5	Clases	Control cases	Validation cases	Total cases
Cancer	9	6	2	500	199	699
Cardiology	6	4	2	64	31	95
Dengue	11	5	4	1414	707	2121

The method used to carry out the experiments with each one of the evaluated databases, is detailed next.

(i) Divide the database in two. One of control or training (approximately 2/3 of the total database) and another one of validation (with the remaining data).
(ii) Process the control database with the induction algorithm C4.5 to obtain the subgroup of variables that will conform the RB-C4.5.
(iii) Repeat for 10%, 20%, ..., 100% of the control database.
  (a) Repeat 30 times, by each iteration.
    (i) Take randomly X% from the control database according to the percentage that corresponds to the iteration.
    (ii) With that subgroup of cases of the control database, make the structural and parametric learning of RB-Complete and the RB-C4.5.
    (iii) Evaluate the predictive power of both networks using the validation database.
  (b) Calculate the average predictive power (from the 30 iterations).
(iv) Graph the predictive power of both networks (RB-Complete and RB-C4.5) based on the cases of training.

The step (i) of the algorithm makes reference to the division of the control and the validation databases. 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 step (ii), we introduced each one of the control databases in a decision tree TDIDT generating system. From there, we obtained the decision tree that represents each one of the analyzed domains. The variables that integrate this representation conforms the subgroup that was considered for the learning of the preprocessed Bayesian networks.

Next (iii) a ten-iteration process begins, in each one of these iterations, it processed 10%, 20%, ..., 100% of the control database for the networks structural and parametric learning. This way, we could analyze not only the difference in the predictive capacity of the networks, but also how this capacity has evolved when we learn with greater amount of cases.

The objective of the repetitive structure of the step (a) is to minimize the accidental results that do not correspond with the reality of the model in study. We managed to minimize this effect, taking different data samples and average the obtained values.

In the steps (a)i., (a)ii. and (a)iii., the structural and parametric learning of the RB-Complete and the RB-C4.5 is made from the subgroup of the control database (both networks are obtained from the same subgroup of data). Once we obtained the network, we have to evaluate the predictive capacity with the validation databases. This database is scanned 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. The predictive capacity corresponds to the percentage of cases classified correctly respect to the total evaluated cases.

In step (b), the predictive power of the network is calculated by dividing the obtained values through all the iterations.

Finally in step (iv), it is come to graph the predictive power average of both Bayesian networks based on the amount of training cases.

3. Results

The experimental results were obtained by the application of the methodology previously mentioned to each of the test databases.

In the "Cancer" domain we predict, based on tumor characteristics, the type of tumor. As can be observed in Fig. 1 the predictive power of the RB-C4.5 is superior to the one of RB-Complete throughout all its points. Furthermore, 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 after 350 training cases, the predictive power of the networks become stabilized at its maximum level.

In the "Cardiology" domain, we predict a disease based on symptoms. The graph of Fig. 2 shows an improvement on the RB-C4.5 can be observed with respect to RB-Complete. Although the differences between the values obtained with both networks are smaller than in the previous case, the hybrid algorithm presents a better approach to reality that the other one. It is important to emphasize that in

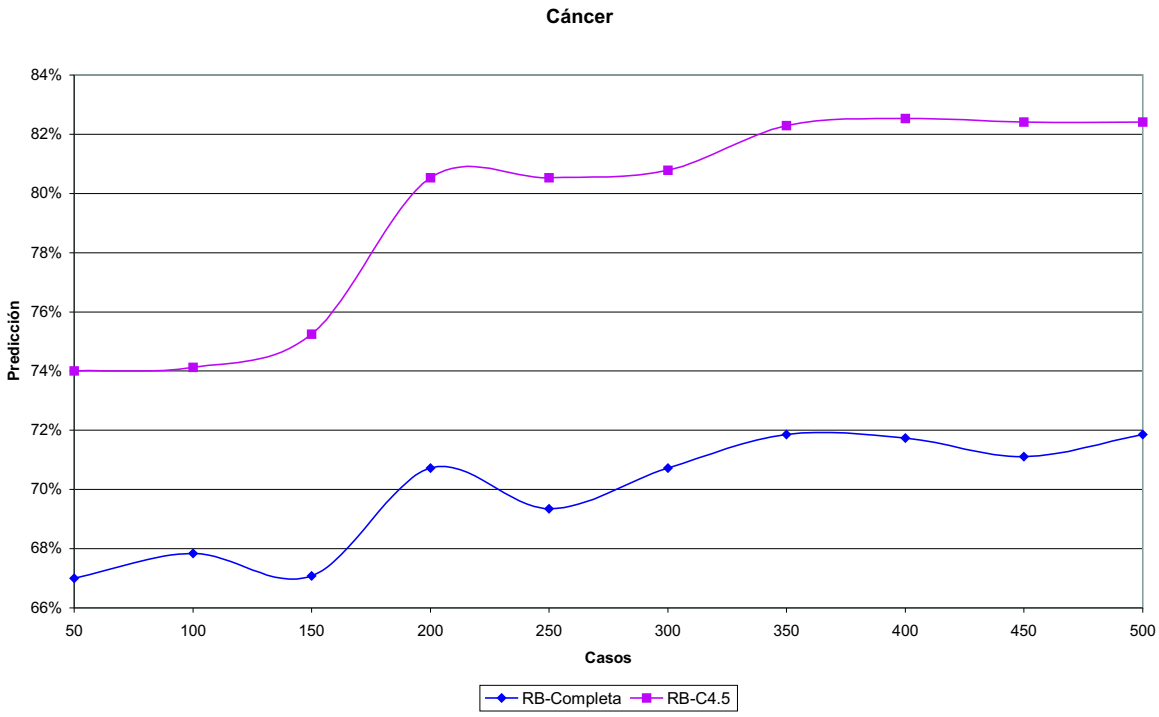

Fig. 1. The predictive power for the "Cancer" database.

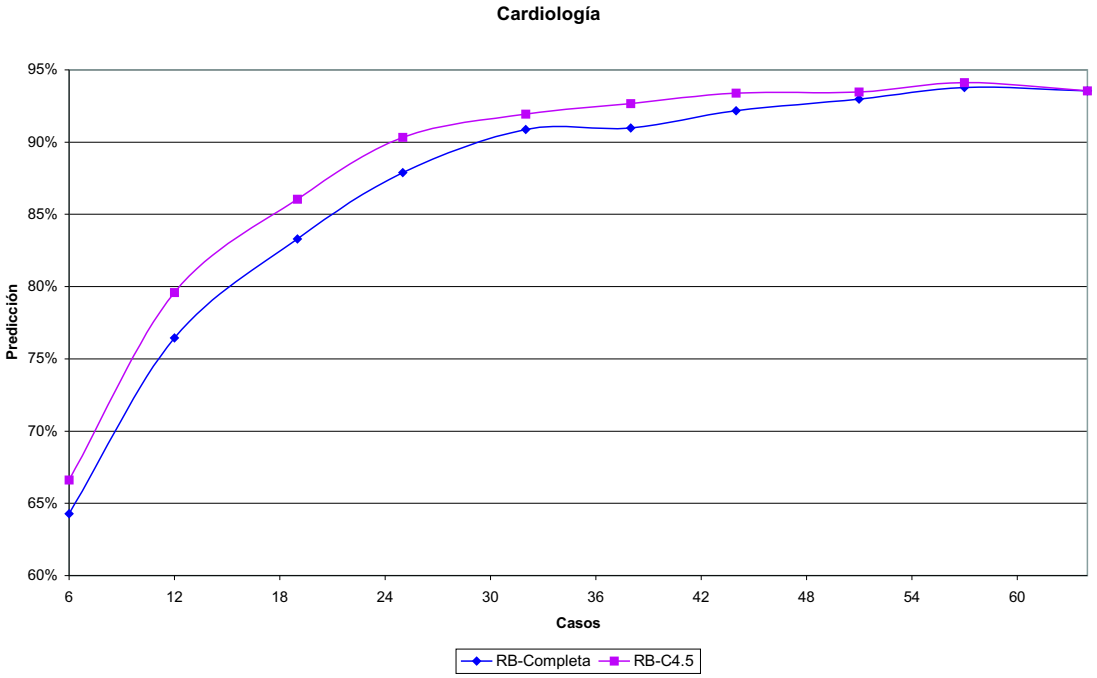

Fig. 2. The predictive power for the "Cardiology" database.

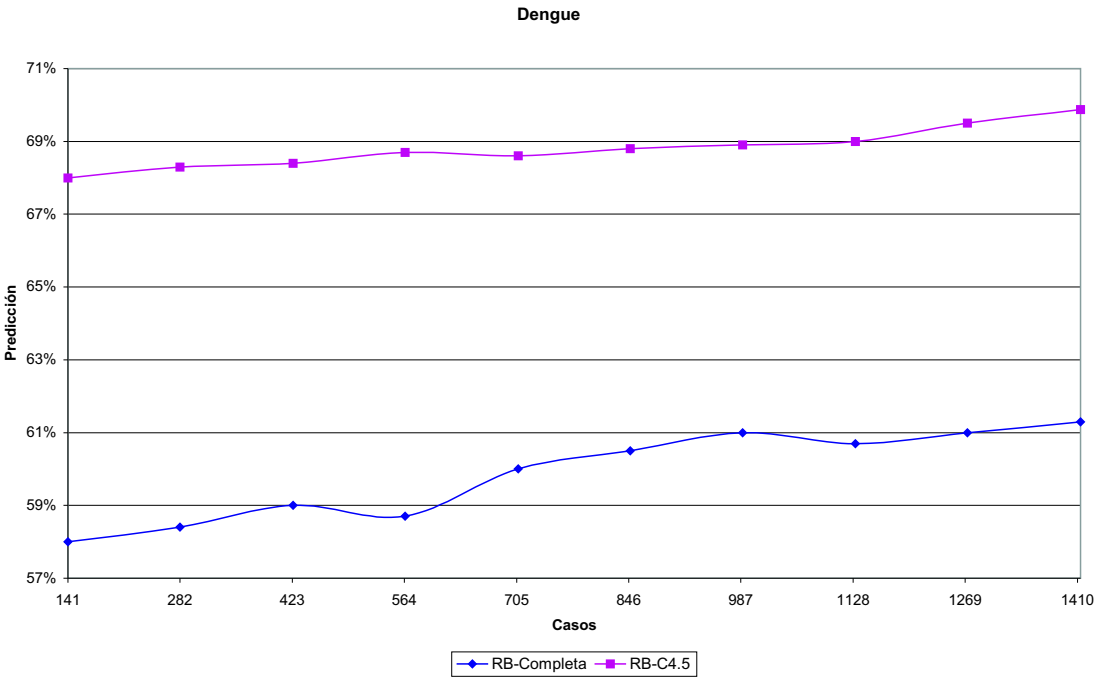

Fig. 3. The predictive power for the "Dengue" database.

this case, the improvement level is minimized when the set of cases used for the learning process is increased.

In "Dengue" domain we predict the distribution of the disease based on ambiental characteristics. Figure 3 shows an improvement in the predictive power of

the proposed network. The RB-C4.5 makes the classification with a 10% better precision than the other network.

4. Discussion and Conclusions

As it is possible to observe, all the graphs that represent the predictive power as a function of the amount of cases of training are increasing. This phenomenon occurs independent of the domain of data used and the evaluated method (RB-Complete or RB-C4.5). From 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 one obtained without making the preprocessing of the variables (RB-Complete).

In another aspect, the RB-C4.5 has fewer variables (or at most equal) than RB-Complete, this reduction in the amount of involved variables produces a simplification of the analyzed domain, which results in two important advantages; firstly, they facilitate the representation and interpretation of the knowledge removing parameters that do not concern in a direct way the objective (classification task). Secondly, it simplifies and optimizes the reasoning task (propagation of the probabilities) which is fundamental to 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.

References

1. W. Fritz, R. García-Martínez, A. Rama, J. Blanqué, R. Adobatti and M. Sarno, *Robot. Auton. Syst.* **5**, 109 (1989).
2. R. García-Martínez and D. Borrajo, *J. Intell. Robot. Syst.* **29**, 47 (2000).
3. G. Perichinsky and R. García-Martínez, *Proc. Workshop Comput. Sc. Researchers* (La Plata University Press, Buenos Aires, 2000), p. 107.
4. G. Perichinsky, R. García-Martínez and A. Proto, Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining, CD of the VI Comput. Sc. Argentinean Congr. (Ushuaia, 2000).
5. G. Perichinsky, R. García-Martínez, A. Proto, A. Sevetto and D. Grossi, Data Mining: Supervised and Non-Supervised Intelligent Knowledge Discovery, *Proc. II Workshop Computes Sc. Researchers* (San Luis University Press, San Luis, 2001).
6. G. Perichinsky, A. Servetto, R. García-Martínez, R. Orellana and A. Plastino, Taxomic Evidence Applying Algorithms of Intelligent Data Minning Asteroid Families, *Proc. Int. Conf. Comput. Sci.*, *Software Eng.*, *Information Technology*, *e-Bussines & Applications* (Rio de Janeiro, 2003), p. 308.
7. M. Chen, J. Han and P. Yu, *IEEE Trans. Knowledge and Data Eng.* **8**, 866 (1996).
8. H. Mannila, Methods and problems in data mining, *Proc. of Int. Conf. on Database Theory* (Delphi, Greece, 1997).
9. G. Piatetski-Shapiro, W. J. Frawley and C. J. Matheus, *Knowledge Discovery in Databases*: *An Overview* (AAAI-MIT Press, Menlo Park, California, 1991).
10. S. Evangelos and J. Han, *Proc. 2nd Int. Conf. Knowledge Discovery and Data Min.* (Portland, United States, 1996).

11. R. S. Michalski, I. Bratko and M. Kubat, *Machine Learning and Data Mining, Methods and Applications* (John Wiley & Sons Ltd, West Sussex, England, 1998).
12. R. S. Michalski, J. G. Carbonell and T. M. Mitchell, *Machine learning I: An AI Approach* (Morgan Kaufmann, Los Altos, CA, 1983).
13. M. Holsheimer and A. Siebes, Data mining: The search for knowledge in databases, Report CS-R9406 (University of Amsterdam, Amsterdam, 1991).
14. R. S. Michalski, A. B. Baskin and K. A. Spackman, A logic-based approach to conceptual database analysis, *6th Annu. Symp. Comput. Appli. Med. Care* (George Washington University, Medical Center, Washington, DC, 1982).
15. G. Piatetsky-Shapiro, U. M. Fayyad and P. Smyth, *From Data Mining to Knowledge Discovery* (AAAI Press/MIT Press, CA, 1996).
16. J. Han, *Data Mining, Urban and Dasgupta* (Encyclopedia of Distributed Computing, Kluwer Academic Publishers, 1999).
17. R. Cowell, A. Dawid, S. Lauritzen and D. Spiegelhalter, *Probabilistic Networks and Expert Systems* (Springer, New York, 1990).
18. M. Ramoni and P. Sebastiani, *Bayesian methods in Intelligent Data Analysis: An Introduction* (Physica Verlag, Heidelberg, 1999).
19. J. Pearl, *Probabilisic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, 1988).
20. I. A. Beinlich, H. J. Suermondt, R. M. Chavez and G. F. Cooper, The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks, *Proc. 2nd Eur. Conf. Arti. Intell. Medicine* (Vienna, 1989).
21. T. W. Bickmore, Real-Time Sensor Data Validation, NASA Contractor Report 195295 (National Aeronautics and Space Administration, United States, 1994).
22. J. S. Breese and R. Blake, Automating Computer Bottleneck Detection with Belief Nets, *Proc. Conf. Uncertainty Arti. Intell.* (San Francisco, CA, 1995), p. 33.
23. K. J. Ezawa and T. Schuermann, Fraud/uncollectible debt detection using a Bayesian network based learning system: A rare binary outcome with mixed data structures, *Proc. Conf. Uncertainty Arti. Intell.* (San Francisco, CA, 1995), p. 157.
24. D. Heckerman, M. Chickering and D. Geiger, *Machine learning* **20**, 197 (1995).
25. F. Diaz and J. M. Corchado, Rough sets bases learning for Bayesian networks, *International workshop on objetive Bayesian methodology* (Valencia, Spain, 1999).
26. D. Heckerman, A tutorial on learning Bayesian networks, Technical report MSR-TR-95-06 (Microsoft research, Redmond, 1995).
27. D. Heckerman and M. Chickering, Efficient approximation for the marginal likelihood of incomplete data given a Bayesian network, Technical report MSR-TR-96-08 (Microsoft Research, Microsoft Corporation, 1996).
28. M. Ramoni and P. Sebastiani, Learning Bayesian networks from incomplete databases, Technical report KMI-TR-43 (Knowledge Media Institute, The Open University, 1996).
29. P. M. Murphy and D. W. Aha, UCI Repository of machine learning databases, Machine-readable data repository, Department of Information and Computer Science (University of California, Irvine).
30. A. Carbajo, S. Curto and N. Schweigmann, 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, Buenos Aires (2003).