Bayesian Networks Optimization Based on Induction Learning Techniques

Paola Britos[1], Pablo Felgaer[2] and Ramon Garcia-Martinez[3]

Abstract 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 with those of the bayesian networks.

1 Introduction

Data mining tasks can be classified in two categories: descriptive data mining and predictive data mining; 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 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 [1]. 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.

[1] Paola Britos
PhD Program, Computer Science School, La Plata University. CAPIS-ITBA. pbritos@itba.edu.ar
[2] Pablo Felgaer
Intelligent Systems Laboratory. School of Engineering. University of Buenos Aires. pfelgaer@fi.uba.ar
[3] Ramon Garcia-Martinez
Software & Knowledge Engineering Center (CAPIS), ITBA. rgm@itba.edu.ar

Obtaining a bayesian network from data is a learning process that is divided in two phases: the structural learning and the parametric learning. 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 to obtain the a priori and conditional probabilities from a given structure. Some characteristics of the bayesian networks are that they allow to learn dependency and causality relations, they allow to combine knowledge with data [2] and they can handle incomplete data [1] [3]. 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 Proposed hybrid learning method

We propose a hybrid learning method that combines the advantages of the induction decision trees techniques 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 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 Xj(i) is the cause or parent of Xi. 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:

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

In this context, 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 shown on table 1.

Table 1. Algorithm to determine the optimal bayesian network

```
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:

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

where {Xj1(i), Xj2(i),…, Xjm(i)} is the set of parents for the variable Xi. 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 [4]. Table 2 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 2. Databases description

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

The algorithm used to carry out the experiments with each one of the evaluated databases, is detailed in table 3. 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 perform the subgroup that was 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. 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.

Table 3. Algorithm used to carry out the experiments

```
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
```

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.

3 Results

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.

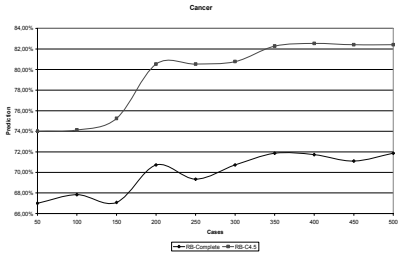

Fig. 1. Results on database "Cancer"

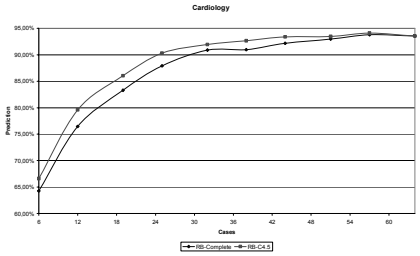

Fig. 2. Results on database "Cardiology"

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.

References

1. Ramoni, M., Sebastiani, P. Bayesian *methods in Intelligent Data Analysis. An Introducction*. Pages 129-166. Physica Verlag, Heidelberg. (1999).
2. Diaz, F., Corchado, J. *Rough sets bases learning for bayesian networks*. International workshop on objetive bayesian methodology, Valencia, Spain. (1999).
3. Heckerman, D., Chickering, M. *Efficient approximation for the marginal likelihood of incomplete data given a bayesian network*. Technical report MSR-TR-96-08, Microsoft Research, Microsoft Corporation. (1996).
4. Murphy, P., Aha, D. *UCI Repository of Machine Learning databases. Machine-readable data repository*, http://mlearn.ics.uci.edu/MLRepository.html. Accessed March, 28th, (2007).