Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches

Basilio Sierra, Pedro Larrañaga

Abstract

In this work we introduce a methodology based on genetic algorithms for the automatic induction of Bayesian networks from a file containing cases and variables related to the problem. The structure is learned by applying three different methods: The Cooper and Herskovits metric for a general Bayesian network, the Markov blanket approach and the relaxed Markov blanket method. The methodologies are applied to the problem of predicting survival of people after 1, 3 and 5 years of being diagnosed as having malignant skin melanoma. The accuracy of the obtained models, measured in terms of the percentage of well-classified subjects, is compared to that obtained by the so-called Naive–Bayes. In the four approaches, the estimation of the model accuracy is obtained from the 10-fold cross-validation method.

Bayesian network; Genetic algorithm; Structure learning; Model search; 10-Fold cross-validation

1. Introduction

Expert systems, one of the most developed areas in the field of artificial intelligence, is that of computer programs designed to help or replace humans tasks where human experience and knowledge are scarce and unreliable. Although there are domains where tasks can be specified by logic rules, other domains are characterized by inherent uncertainty. Probability was not taken into account for some time as a reasoning method for expert systems trying to model uncertain domains, because the computational requirements were considered to be too expensive. At the end of the 1980s, Lauritzen and Spiegelhalter [14] showed that these difficulties can be overcome by exploiting the modular character of the graphical models associated with the so-called probabilistic expert systems, which in this work are called Bayesian networks.

Bayesian networks (BNs) [9,13,16] constitute a probabilistic framework for reasoning under uncertainty. From an informal perspective, BNs are directed acyclic graphs (DAGs), where the nodes are random variables and the arcs specify the independence assumptions that must be held between the random variables. BNs are based upon the concept of conditional independence among variables. This concept makes possible a factorization of the probability distribution of the n-dimensional random variable $(X_1, \ldots, X_n)$ in the following way:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|pa(x_i))$$

where x_i represents the value of the random variable X_i and $pa\ (x_i)$ represents the value of the random variables parents of X_i.

Thus, in order to specify the probability distribution of a BN, one must give prior probabilities for all root nodes (nodes with no predecessors) and conditional probabilities for all other nodes, given all possible combinations of their direct predecessors. These numbers in conjunction with the DAG, specify the BN completely. Once the network is constructed it constitutes an efficient device to perform probabilistic inference. This probabilistic reasoning inside the net can be carried out by exact methods, as well as by approximate methods. Nevertheless, the problem of building such a network remains. The structure and conditional probabilities necessary for characterising the network can be either provided externally by experts or obtained from an algorithm which automatically induces them.

In this paper, a methodology for automatically inducing Bayesian networks is introduced. This methodology is based on genetic algorithms (GAB) and attempts to obtain from a database of cases the best structure of the Bayesian network.

The rest of the paper is organized as follows. In Section 2 some structure learning methods are reviewed, taking special interest in the method proposed by Cooper and Herskovits [5] as well as in the learning of the best Markov blanket (MB) of the variable to be classified; the MB concept was proposed by Pearl [16]. Section 3 introduces GAs, while Section 4 presents the structure learning methodology integrating both the Bayesian network model (Cooper and Herskovits, Markov blanket, relaxed Markov blanket) and the adaptive searching process characteristic of the GAs. In Section 5 we present the results obtained from applying the previous methodology to a database of cases, which contains information on 311 patients diagnosed as having malignant skin melanoma. The induced Bayesian network is used for classifying patients according to their prognosis of survival after one, three and five years of being diagnosed. These results are compared to those obtained by the called Naive–Bayes paradigm. Section 6 presents the conclusions.

2. Structure learning in Bayesian networks

2.1. Introduction

During the last 5 years, a good number of algorithms whose aim is to induce the structure of the Bayesian network that better represents the conditional independence relationships in a database of cases have been developed. In our opinion, the main reason for continuing the research in the structure learning problem is that modeling the expert knowledge has become an expensive, unreliable and time-consuming job.

The different approaches to the structure learning mentioned here are related with multiple connected networks and have been grouped according to the necessity or not of imposing order on the variables. See Heckerman et al. [7] for a good review.

Assuming order among variables means that a variable X_i can have the variable X_j as parent only if, in the established order among the variables, X_j precedes X_i. With this restriction, the cardinality of the space that contains all the structures is given by $2^{\binom{n}{2}}$, where n is the number of variables in the system. Some methods under this restriction are those developed by Cooper and Herskovits [5] and Bouckaert [4].

If we do not assume ordering between the nodes the cardinality of the search space is bigger and it is given by Robinson's formula [18]:

$$f(n) = \sum_{i-1}^{n} (-1)^{i+1}\binom{n}{i}2^{i(n-i)}f(n-i); \quad f(0) = 1, \quad f(1) = 1$$

Several authors have been working under these general assumptions. Among them, Bouckaert [3] and Provan and Singh [17].

2.2. The K2 algorithm

As will be seen in Section 4, one of the proposed approaches-based on GAs-use the CH metric introduced by Cooper and Herskovits [5] for evaluating the goodness of a BN structure, as well as the K2 algorithm developed by the same authors. K2 is an algorithm that creates and evaluates a BN from a database of cases once an ordering between the system variables is given. The CH metric is used for the evaluation of the network that it constructs. K2 searches, given a database D for the BN structure $\mathrm{B}_{\mathrm{S}^*}$. with maximal $P(\mathrm{B}_\mathrm{S}, D)$, where $P(\mathrm{B}_\mathrm{S}, D)$ is as described in the following theorem proved in [5].

Theorem 1. Let Z be a set of n discrete variables, where a variable x_i in Z has r_i possible value assignments: $(v_{i1}, \ldots, v_{ir_i})$. Let D be a database containing m cases, where each case contains a value assignment for each variable in Z. Let $\mathrm{B_S}$ denote a BN structure containing only the variables in Z. Each variable x_i in $\mathrm{B_S}$ has a set of parents, which are represented with a list of variables π_i. Let w_{ij} denote the jth unique instantiation of π_i relative to D. Suppose there are q_i such unique instantiations of π_i. Define N_{ijk} to be the number of cases in D in which variable x_i has the value v_{ik} and π_i is instantiated as w_{ij}. Let $N_{ij} = \Sigma_{k=1}^{r_i} N_{ijk}$. If given a BN model, the cases occur independently, there are no cases that have variables with missing values and the density function $f(\mathrm{B}_P|\mathrm{B_S})$ is uniform, then it follows that $P(\mathrm{B_S}|D) = P(\mathrm{B_S})\Pi_{i=1}^{n} g(i, \pi_i)$, where $g(i, \pi_i) = ((r_i - 1)!)/((N_{ij} + r_i - 1)!)\Pi_{k=1}^{r_i} N_{ijk}!$

The K2 algorithm assumes that an ordering on the variables is available and that, a priori, all structures are equally likely. For every node, it searches for the set of parent nodes that maximizes $g(i, \pi_i)$—CH metric. K2 is a *greedy* heuristic. It starts by assuming that a node does not have parents, after which in every step it adds incrementally that parent whose addition most increases the probability of the resulting structure. K2 stops adding parents to the nodes when the addition of a single parent cannot increase the probability. Obviously, this approach does not guarantee the selection of a structure with the highest probability.

2.3. Markov blanket approach

The Naive–Bayes method takes into account the fact that there is a special variable to be classified. However, this approach does not manage in an adequate manner the intrinsic semantics of BNs.

Taking into account that in a BN, any variable is influenced only by its Markov blanket (MB), i.e. its parent variables, its children variables and the parent variables of its children variables, it seems to be intuitive to do the search in the set of structures that are MB of the special variable. This concept of variable associated MB has been used to facilitate the simulation of the BN by Gibbs sampling and can be formally established by the next theorem [15].

Theorem 2. The probability distribution of each variable x_i in the network, conditioned on the state of all other variables is given by the product:

$$P(x_i|\mathbf{Z}_{x_i}) = \alpha P(x_i|\pi_{x_i}) \prod_j P(\omega_{ij}|\pi_{ij}(x_i))$$

where α is a normalizing constant, independent of x_i and x_i, $\mathbf{Z}_{x_i}$, π_i, ω_{ij} and $\pi_{ij}(x_i)$ denote any consistent instantiations of X, $Z_X = Z - X$, Π_X, Ω_j and Π_{ij} respectively, where Z is the set of all variables, Π_i the set of X's parents, Ω_i the set of X's children and Π_{ij} the set of parents of Ω_i.

2.4. Relaxed Markov blanket

Due to the overfitting obtained with the Markov blanket approach during the evaluation of the models using 10-fold crossvalidation, we have relaxed the Markov blanket concept in order to simplify the requirements and obtain simpler networks.

This is effected by two relaxations: (1) not all the variables of the model have to be part of the Markov blanket of the variable to be classified; and (2) we avoid some special structures that could be obtained. Any variable that is a parent of the variable to be classified cannot be the parent of one child of the variable to be classified and conversely, one variable can be parent of only one child of the variable to be classified.

3. Genetic algorithms

The computing complexity inherent in a great number of real problems of combinatorial optimization has motivated the development of heuristic methods that try to tackle these problems successfully. A heuristic is a procedure which will give a good solution—not necessarily the optimal—to problems which can be catalogued as difficult, if an attempt is made to solve them obtaining the exact solution. Although there are heuristics developed for specific problems, in the past there has been an explosion in the application of what we could call metaheuristics, because their formulation is independent of the problem to solve. Among the most studied metaheuristics we quote simulated annealing, Tabu search and GAs.

GAs [6] are adaptive methods that can be used for solving problems of search and optimization. They are based on the genetic processes of living organisms. Through generations the populations evolve in nature according to the principles of natural selection and survival of the fittest postulated by Darwin. Imitating this process, the GAs are capable of creating solutions for real world problems.

GAs use a direct analogy with the natural behaviour. They work with a population of individuals, each individual representing a feasible solution to a given problem. To each individual we assign a value or score according to the goodness of that solution. The better the adaptation of the individual to the problem, the more probable is that the individual will be selected for reproduction, crossing its genetic material with another individual selected in the same way. This cross will produce new individuals—offspring of the previous generation—which share some of the features of their parents. In this way a new population of feasible solutions is produced, replacing the previous one and verifying the interesting property of having greater proportion of good features than the previous population. Thus, through generations good features are propagated in the population. Favouring the cross of the fittest individuals, the most promising areas of the search space are being explored. If the GAs are well designed, the population will converge to an optimal solution of the problem.

Fig. 1 summarizes the pseudocode for the so-called abstract genetic algorithm. In it the parent selection does not need to be made by aligning to each individual a value proportional to its objective function, as is usual in the so-called simple genetic algorithm. This selection can be carried out by any function that selects parents in a natural way. It is worth noticing that descendants are not necessarily the next generation of individuals, but that this generation is made by the union of parents and descendants. This is why we need the operations of extension and reduction in the cycle.

4. Genetic algorithms in the induction of Bayesian networks

In this approach, each individual in the GA will be a BN structure.

4.1. Notation and representation

Denoting with D the set of BN structures for a fixed domain with n variables and the alphabet S being $\{0, 1\}$, a BN structure can be represented by an $n \times n$ connectivity matrix $\mathbf{C}$, where its elements, c_{ij}, verify:

$$c_{ij}\begin{cases} 1 & \text{if } j \text{ is a parent of } i \\ 0 & \text{otherwise.} \end{cases}$$

4.2. Assuming an ordering between the nodes

In this case, the connectivity matrices of the network structures are triangulated (i.e. elements under the diagonal are all 0) and therefore the genetic operators are closed operators with respect to the DAG conditions. We represent an individual of the population by the string:

$$c_{21}c_{31}c_{41}\ldots c_{n1}, \ldots c_{32}c_{42}\ldots c_{n2}, \ldots c_{n-1n-2}, c_{nn-2}, c_{nn-1}.$$

With this representation in mind, we show how the crossover and mutation operators work by using simple examples (Fig. 2).

```
begin AGA
    Make initial population at random
    WHILE NOT stop DO
      BEGIN
      Select parents from the population.
      Produce children from the selected parents.
      Mutate the individuals.
      Extend the population by adding the children to it.
      Reduce the extended population.
      END
    Output the best individual found.
end AGA
```

Fig. 1. The pseudo-code of the abstract genetic algorithm.

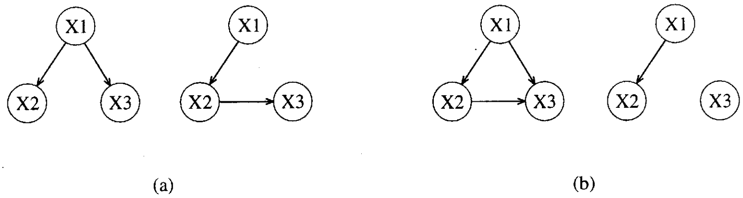

Fig. 2. With order assumption: crossing over two BN structures.

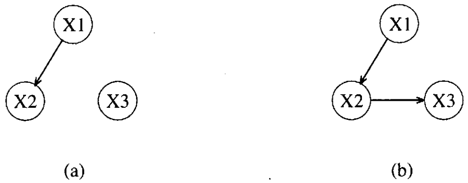

Fig. 3. With order assumption: mutating a BN structure.

Example 1. Consider a domain of three variables on which the two BN structures of Fig. 2(a) are defined. Using the above described representation, the networks are represented by the strings: 110 and 101. Suppose now that the two network structures are crossed over and that the crossover point is chosen between the second and the third part; this gives the offspring strings 111 and 100. Hence, the created offspring structures are those presented in Fig. 2(b).

Example 2. Consider the DAG of Fig. 3(a); it is represented by the string 100. Suppose that the third part is altered by mutation. This gives the string 101, which corresponds with the graph of Fig. 3(b).

4.3. Without assuming an ordering between the nodes

If no ordering assumption on the variables is made, we represent an individual of the population by the string:

$$c_{11}c_{21}\ldots c_{n1}c_{12}c_{22}\ldots c_{n2}\ldots c_{1n}c_{2n}\ldots c_{nn}.$$

As can be seen in the following examples, the genetic operators are not closed operators with respect to the DAG conditions.

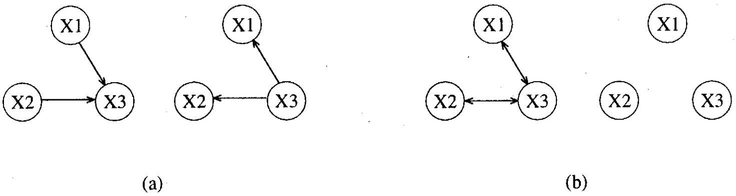

Fig. 4. Without order assumption: the crossover operator is not a closed operator.

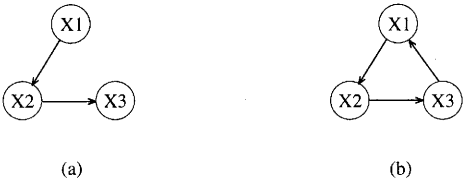

Fig. 5. Without order assumption: the mutation operator is not a closed operator.

Example 3. Consider a domain of three variables on which the two BN structures of Fig. 4(a) are defined. Using the above described representation, the networks are represented by the strings: 001001000 and 000000110. Suppose now that the two network structures are crossed over and that the crossover point is chosen between the sixth and the seventh part. This gives the offspring strings 001001110 and 000000000. Hence, the created offspring structures are the ones presented in Fig. 4(b). We see that the first offspring structure is not a DAG.

Example 4. Consider the DAG of Fig. 5(a). It is represented by the string 010001000. Suppose that the seventh bit is altered by mutation. This gives string 010001100, which corresponds to the cyclic graph of Fig. 5(b). To assure the closeness of the genetic operators we introduce a repair operator, which transforms the child structures that do not satisfy the DAG conditions into DAGs, by randomly eliminating the edges that invalidate the DAG conditions.

This approach has been evaluated empirically with a simulation of the ALARM network [2]. For details see Larranaga et al. [10,11]. Another approach, in which the individuals of the population are orderings, has been proposed by Larranaga et al. [12]. Table 1 shows the number of individuals evaluated in each approach. Obviously, the search space is smaller in the MB and RMB approaches.

5. Predicting survival in malignant skin melanoma

5.1. The malignant skin melanoma

In spite of the advances achieved in recent years in the treatment of cancer, the prognosis of patients having developed skin melanoma has changed very little. The

Table 1
Number of individuals evaluated by GAs in each approach

Approach	Individuals evaluated		
	1 Year	2 Years	3 Years
CH-GA	181051	180961	146813
MB	3343	3371	3284
RMB	3375	3279	3335

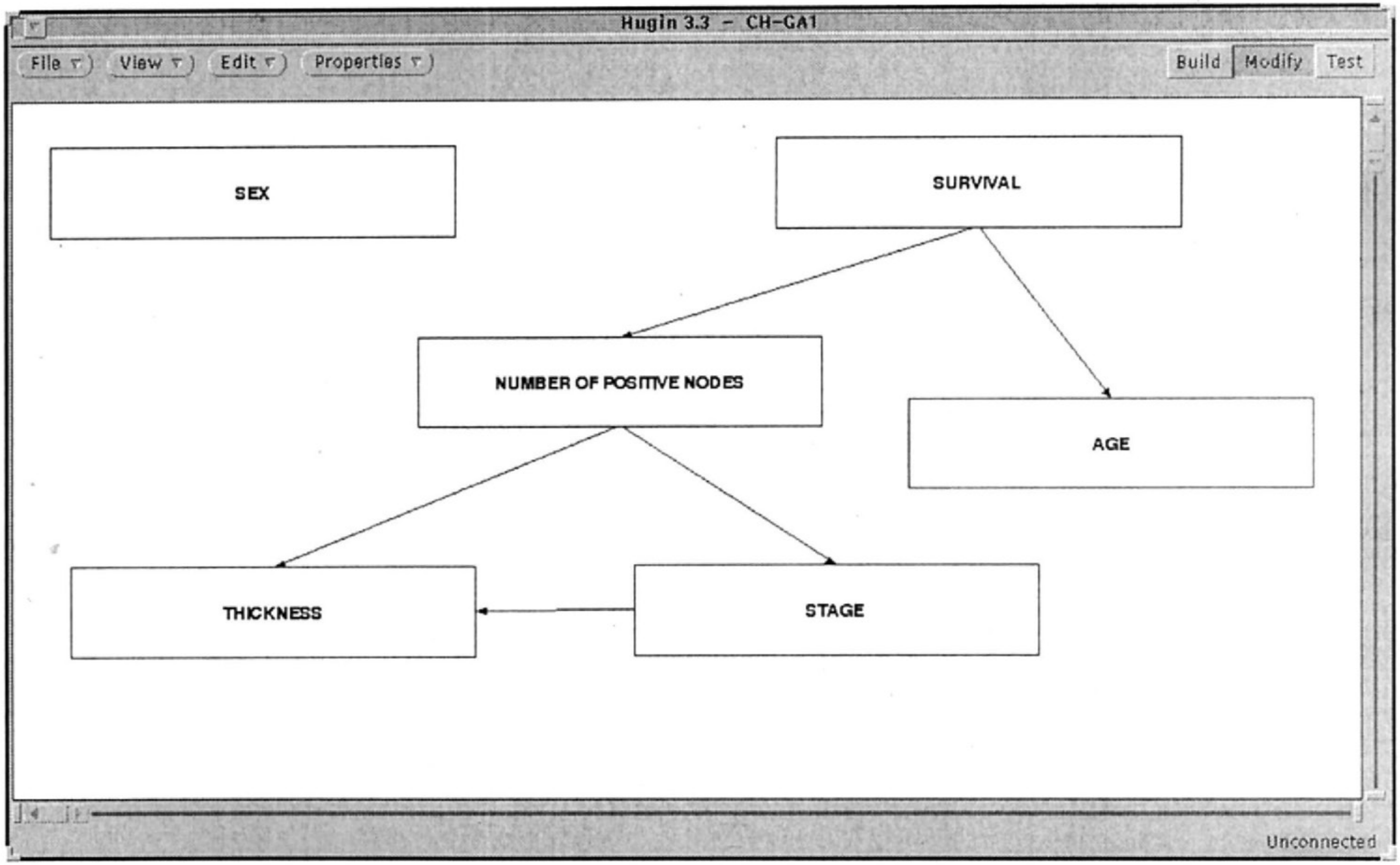

Fig. 6. The a posteriori most probable structure for the 1 year case.

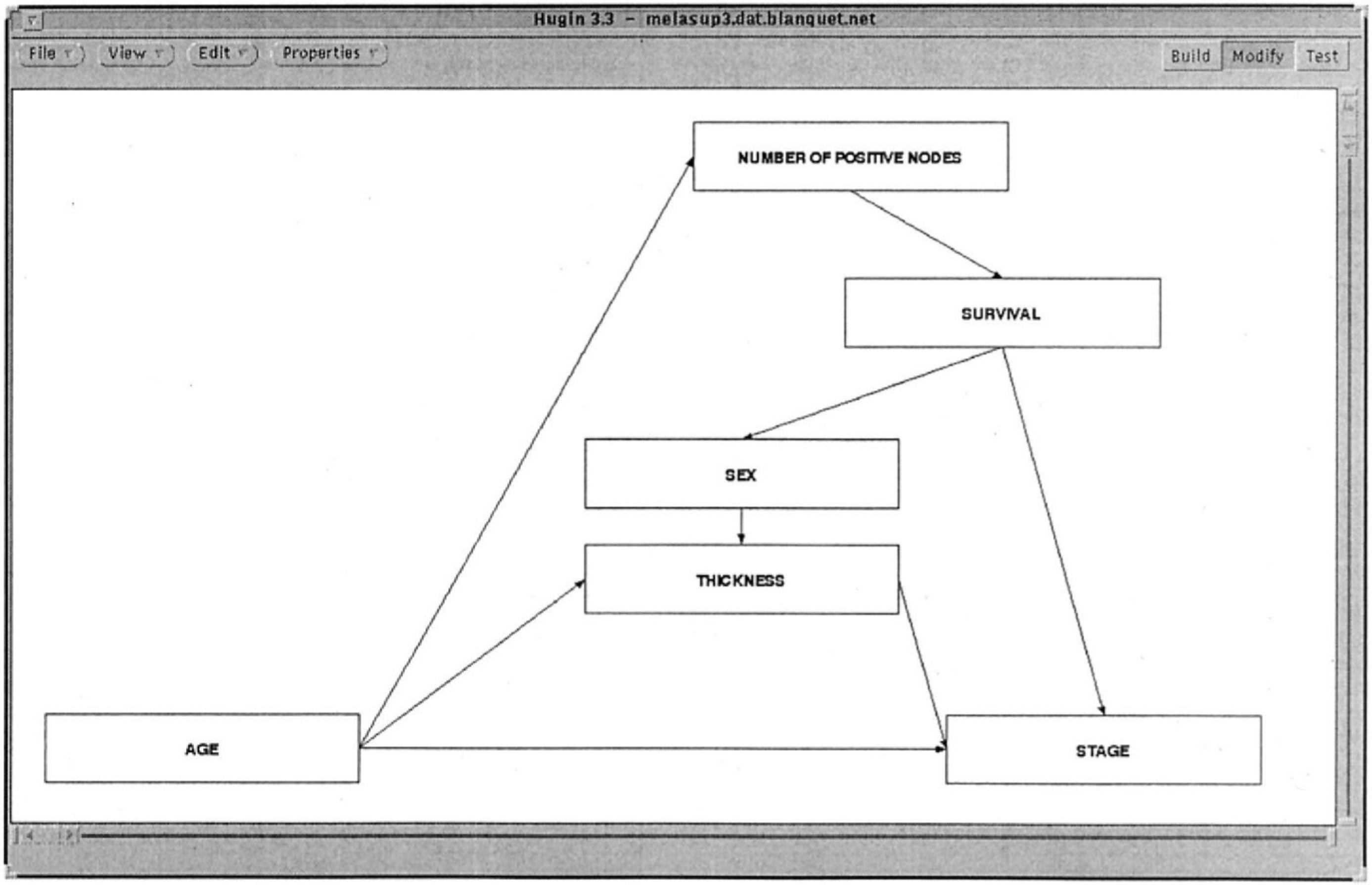

Fig. 7. The Markov blanket obtained for the 3 year case.

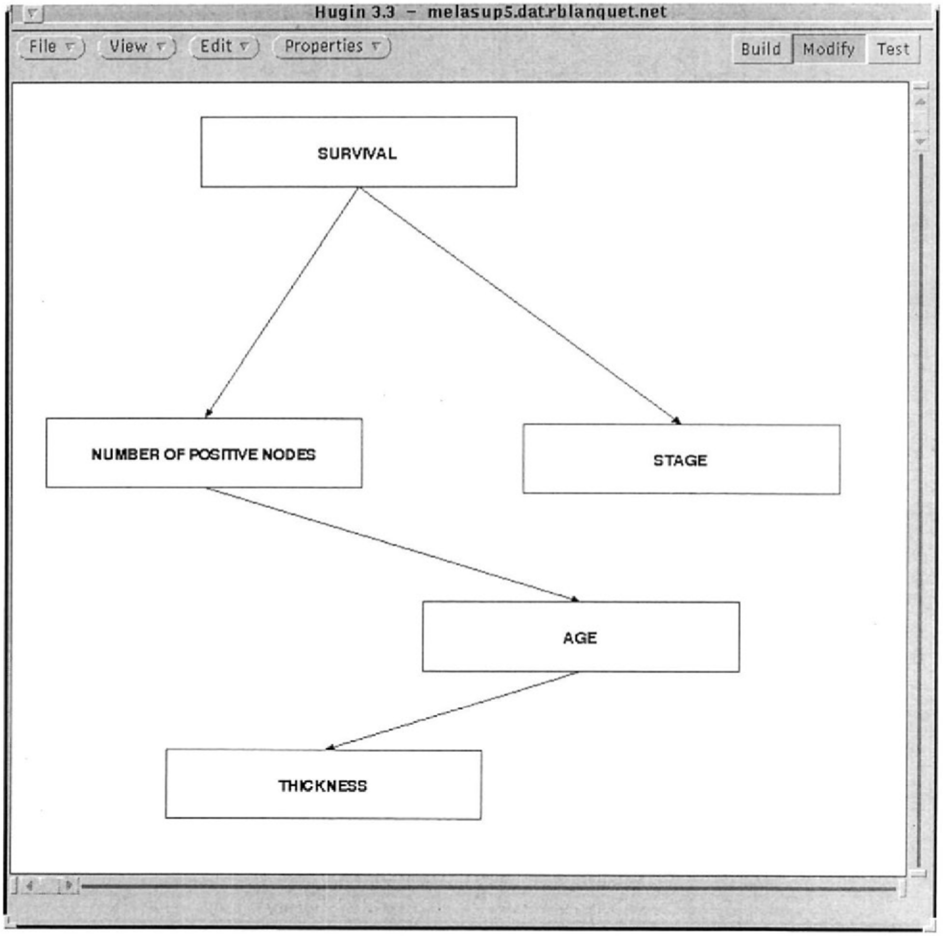

Fig. 8. The reduced Markov blanket obtained for the 5 year case.

incidence of the disease has continuously grown over the last decade. Annual incidence has increased and the progressive reduction of the ozone layer, if not stopped, will increase it even more.

Experimental data and the results of epidemiological studies suggest two main risk factors: sun exposure along with phenotype characteristics of the individual. Thus, for example, continuous exposure to the sun represents an odds ratio of 9, while acute intermittent exposure has an associated odds ratio of 5.7.

Malignant skin melanoma is a rather uncommon tumour in our country. It entails between 8 and 10% of the total malignant tumours that affect the skin. According to the Cancer Register of the Basque Country [8], in 1990 the rate of incidence was 2.2 for every 100000 people for males and three for every 100000 for females.

The database contains 311 cases-diagnosed at the Oncological Institute of Gipuzkoa in the period between 1 January, 1988 and 31 December, 1995 and for each case we have information about eight variables. The five predictor variables are: sex (two categories), age (five categories), stage (four categories), thickness (four categories) and number of positive nodes (two categories). The variable to predict has two categories taking into account if the person survives or not 1, 3 or 5 years after being diagnosed as having malignant skin melanoma.

5.2. The models

Four models have been taken into account. First, we have induced a BN structure using GAs, as explained in Section 4. In order to get it, we have searched the space of all structures without imposing any order restriction among the variables. Therefore, we have tried to find, given a database of cases, the a posteriori most probable structure. The second model used is the search of the best Markov blanket of the variable to be classified and the goal of the GA is to maximize the percentage of correctly classified cases. The third model is a relaxation of the Markov blanket concept and again we use the well classified percentage as goal function. The fourth model is the so-called Naive–Bayes. This model assumes independence among predictor variables. In both models the estimations of the rate of well-classified individuals have been obtained using 10-fold cross-validation [19]. The propagation of the evidence has been carried out using the HUGIN software [1].

Model I. The a posteriori most probable structure. CH-GA. Fig. 6 shows the structure of the Bayesian network induced by the genetic algorithm. It corresponds to the predictions of survival after one year of being diagnosed.

Model II. The Markov blanket of the variable to be classified. MB-GA. Fig. 7 shows the Markov blanket obtained for the three years case. In the proposed approach, Markov blanket induced by genetic algorithms MB-GA, individuals in GA are BN structures that constitute MB for the variable to be classified. We have introduced to the GA one operator that guarantees that the obtained children comply with a MB of the variable to be classified. In order to do this, we have to force every variable of the problem to be parent, child or parent of a child of the special variable. The accuracy of all generated structures is measured using the well-classified percentage obtained by applying the evidence propagation of the HUGIN software.

Model III. The relaxed Markov blanket. RMB-GA. Fig. 8 shows the relaxed Markov blanket obtained for the 5 years case. Not all the variables are part of the Markov blanket of the variable to be classified.

Model IV. Naive–Bayes classifier. N–B. In spite of the strong assumptions of independence upon which the model is built, Naive–Bayes classifier has proved itself competitive against other more refined classifiers. It is assumed that all variables are conditionally independent given the value of the variable to predict. Therefore, the model ignores the correlations among variables which can prejudice its predictive capacity. Fig. 9 gives the structure of the BN corresponding to the Naive–Bayes. This structure is common to the three classification problems. Table 2 shows that the estimations obtained by two of the Naive–Bayes models are inferior to those obtained by the other approach.

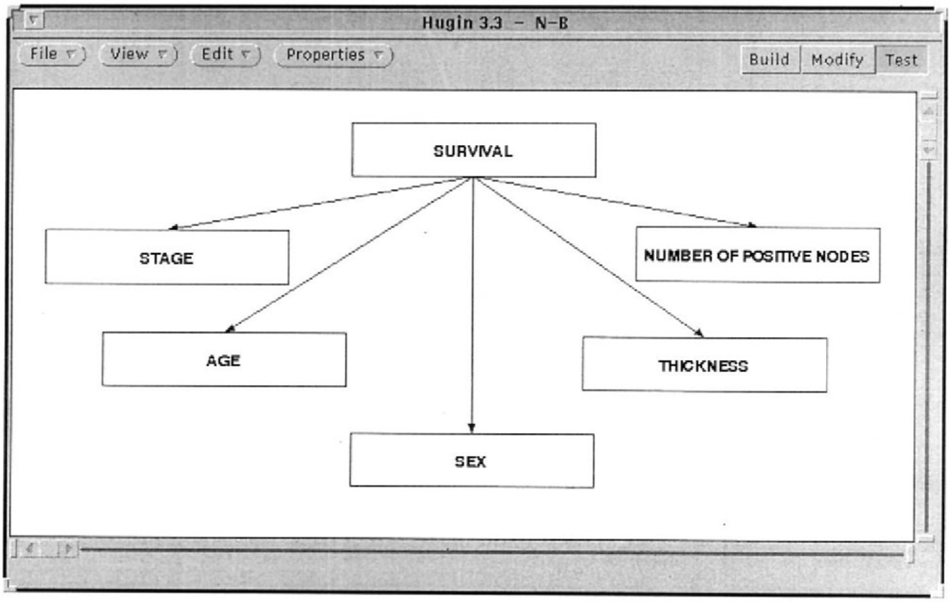

Fig. 9. The Naive–Bayes classifier.

Table 2
Distribution of survival/no survival (%)

	Yes	No
1 Year	3.06	6.94
3 Years	81.95	18.05
5 Years	7.28	32.72

Table 3
Accuracy of the different approaches for the prediction of survival 1, 3 and 5 years after being diagnosed

Survival of malignant skin melanoma (%)			
	1 Year	3 Years	5 Years
CH-GA	93.06	81.95	69.57
MB	94.28	83.90	78.88
RMB	93.47	83.85	74.53
N-B	91.43	79.02	71.43

Table 4
Hamming distance between the different BN structures

		CH-GA			MB			RMB		
		1 Year	3 Years	5 Years	1 Year	3 Years	5 Years	1 Year	3 Years	5 Years
CH-GA	1 Year	0	7	7	11	12	11	14	16	14
	3 Years		0	0	16	9	16	7	11	7
	5 Years			0	16	9	16	7	11	7
MB	1 Year				0	17	0	17	17	17
	3 Years					0	17	12	12	12
	5 Years						0	17	17	17
RMB	1 Year							0	8	0
	3 Years								0	8
	5 Years									0

Table 3 gives estimations of the rate of success in classification obtained by each of the previous models. Distribution of survival/no survival in the original databases are given in Table 2. Table 4 shows the Hamming distance between the different BNs obtained.

6. Conclusions and further research

Different methods of induction of Bayesian networks has been introduced. These methods are based on intelligent search made by genetic algorithms. One method uses the CH metric and tries to find the a posterior) most probable Bayesian network structure given the database of cases and the other two search for the best-restricted model using the well classified percentage as goal function.

The Bayesian network structures induced by these method have been empirically compared to the Naive–Bayes structures in one classification problem consisting of the prediction of survival of individuals after 1, 3 or 5 years of being diagnosed as having malignant skin melanoma. We can see that the Markov blanket approach seems to be the best once 10-fold crossvalidation is performed. However, in other cases, we have obtained some overfitting with this method.

In the future, we plan to add to the other methods some restrictions in order to introduce to the obtained Bayesian network some knowledge given by a human expert in the form of conditional dependencies or independencies that the variables should obey.

Acknowledgements

This work was supported by the CICYT under grant TIC97-1135- C04-03, by the grant UPV 140.226-EA186/96 from the University of the Basque Country and by the grant PI 96/12 from the Gobierno Vasco-Departamento de Educación, Universidades e Investigación.

References

[1] Andersen SK, Olesen KG, Jensen FV, Jensen F. HUGIN-a shell for building Bayesian belief universes for expert systems. In: 11th Int. Joint Conf. on Artificial Intelligence, 1989:1128–1133.

[2] Beinlinch IA, Suermondt HJ, Chavez RM, Cooper GF. The ALARM monitoring system: a case study with two probabilistic inference techniques for belief networks. In: Proc. 2nd European Conf. on Artificial Intelligence in Medicine, 1989:247–256.

[3] Bouckaert RR. Optimizing causal orderings for generating DAGs from data. In: Uncertainty in Artificial Intelligence. Proc. 8th Conf., 1992:9–16.

[4] Bouckaert RR. Properties of Bayesian belief networks learning algorithms. In: Uncertainty in Artificial Intelligence. 10th Annual Conf., 1994:102–109.

[5] Cooper GF, Herskovits EA. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9:309-347.

[6] Goldberg DE, Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[7] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. In: Technical Report MSR-TR-94-09, Microsoft, 1994.

[8] Izarzugaza MI. Informe del registro de Cáncer de Euskadi 1990. Osasunkaria 1994:8–11

[9] Jensen FV. Introduction to Bayesian networks. University College of London, UK, 1996.

[10] Larranaga P, Poza M, Yurramendi Y, Murga R, Kuijpers C. Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters. IEEE Trans Pattern Anal Mach Intell 1996;18:912-926.

[11] Larranaga P, Murga R, Poza M, Kuijpers C. Structure learning of Bayesian networks by hybrid genetic algorithms. In: Fisher D, Lenz H-J, editors. Learning from Data: AI and Statistics V, Lecture Notes in Statistics 112. New York: Spriger, 1996:165–174.

[12] Larranaga P, Kuijpers C, Murga R, Yurramendi Y. Learning Bayesian network structures by searching for the best ordering with genetic algorithms. IEEE Trans Syst Man Cybern 1996;26:487-493.

[13] Lauritzen SL, Graphical Models. London: Oxford University Press, 1996.

[14] Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application on expert systems. J Royal Stat Soc B 1988;50:157-224.

[15] Pearl J. Evidential reasoning using stochastic simulation of causal models. Artif Intell 1987

[16] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann, 1988.

[17] Provan GM, Singh M. Learning Bayesian networks using feature selection. In: Fisher D, Lenz H-J, editors. Learning from Data: AI and Statistics V, Lecture Notes in Statistics 112. New York: Spriger, 1996:291–300.

[18] Robinson RW. Counting unlabeled acyclic digraphs. In: Little CHC, editor. Lectures Notes in Mathematics 622: Combinatorial Mathematics V. New York: Springer, 1977:28–43.

[19] Stone M. Cross-validation choice and assessment of statistical procedures. J Royal Stat Soc 1974;36:111-147.