Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 38 (2005) 376–388
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS

Rosa Blanco [a,*], Iñaki Inza [a], Marisa Merino [b], Jorge Quiroga [c], Pedro Larrañaga [a]

[a] Department of Computer Science and Artificial Intelligence, University of Basque Country, P.O. Box 649, E-20080 San Sebastián, Spain
[b] Basque Health Service–Osakidetza, Comarca Guipúzcoa–Este, Av. Navarra 14, E-20013 San Sebastián, Spain
[c] Faculty of Medicine, University Clinic of Navarra, E-31080 Pamplona, Spain

Received 17 November 2004
Available online 4 June 2005

Abstract

The transjugular intrahepatic portosystemic shunt (TIPS) is a treatment for cirrhotic patients with portal hypertension. A subgroup of patients dies in the first 6 months and another subgroup lives a long period of time. Nowadays, no risk factors have been identified in order to determine how long a patient will survive. An empirical study for predicting the survival rate within the first 6 months after TIPS placement is conducted using a clinical database with 107 cases and 77 variables. Applications of Bayesian classification models, based on Bayesian networks, to medical problems have become popular in the last years. Feature subset selection is useful due to the heterogeneity of the medical databases where not all the variables are required to perform the classification. In this paper, filter and wrapper approaches based on the feature subset selection are adapted to induce Bayesian classifiers (naive Bayes, selective naive Bayes, semi naive Bayes, tree augmented naive Bayes, and k-dependence Bayesian classifier) and are applied to distinguish between the two subgroups of cirrhotic patients. The estimated accuracies obtained tally with the results of previous studies. Moreover, the medical significance of the subset of variables selected by the classifiers along with the comprehensibility of Bayesian models is greatly appreciated by physicians.

Keywords: Bayesian classification models; Filter approach; Wrapper approach; Transjugular intrahepatic portosystemic shunt; Survival prediction

1. Introduction

During the last few decades, researchers in artificial intelligence have developed new machine learning methods that construct predictive models from data, obtaining promising results in many clinical areas [1–4].

Due to the heterogeneity of medical variables [5] (collected via several methods), some variables could be irrelevant or redundant. The predictive accuracy of supervised classification models are not monotonic regarding the inclusion of features [6]. Irrelevant and redundant variables would reduce predictive performance.

In medical domains, several variables are based on invasive tests. These medical techniques could be painful for patients and the test result could only confirm an expected diagnosis. Moreover, not all the tests have the same economical cost. Thus, some of them are so painful and expensive that they are only carried out when strictly required.

In other cases, the relation between the number of instances and the number of variables is unbalanced, i.e., the number of examples are insufficient in comparison with the number of variables. Any classifier built under these conditions tends to overfit the input database, that is, the classifier correctly labels nearly all the examples seen but misclassifies many of the unseen examples. In spite of the undetectable misclassification

* Corresponding author. Fax: +34 943015590.
E-mail address: rosa@si.ehu.es (R. Blanco).
URL: http://www.sc.ehu.es/ccwbayes/members/rosa (R. Blanco).

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.05.004

error over the input database, the misclassification error of unseen instances is too large. A tool for dimensionality reduction could improve the classification results.

In an effort to solve these problems (decreased accuracy, extra cost, and overfitting), feature subset selection (FSS) aims to find the 'best' subset of variables with the 'best' accuracy for a given classification task. FSS has been tackled with success in several medical areas [7,8] with an increase in accuracy and a decrease in acquisition cost due to the reduction of the medical tests. A classification model with a small number of variables may be used more quickly and easily, with a rise in the interpretability and understanding of the classification models.

In the western world, 90% of the cases of portal hypertension are caused by cirrhosis of the liver. Portal hypertension has serious consequences, i.e., gastroesophageal varices, hepatic encephalopathy, hypersplenism, and ascites. The bleeding originated by gastroesophageal varices is a significant cause of mortality (approximately 30–50% at the first bleeding) [9,10].

The transjugular intrahepatic portosystemic shunt (TIPS) is a non-surgical method resulting in decompression of the portal system. A prosthesis is placed between the portal and the suprahepatic veins by means of an angiographic method. In spite of the number of studies carried out, the relationship between TIPS and the survival of treated patients is almost unknown.

Our medical staff identified a subgroup of patients who died within 6 months after TIPS placement whereas the rest of patients lived on for long periods. No risk factors have been determined to distinguish between the two subgroups.

The choice of a 6-month period is based on critical reasons. Medical factors like stenosis of the shunt and rebleeding would complicate the analysis. Moreover, the medical study does not show important variations in mortality after the first 6 months after TIPS placement.

Traditionally, Pugh's modification of the Child–Turcotte classification (referred to as the Child–Pugh classification) has been used to assess risk in patients undergoing portosystemic shunt surgery [11]. Despite its traditional use to asses the seriousness of liver disease, it has inherent problems when applied to patients undergoing TIPS. It cannot be used to predict the survival of the patients within a certain period of time. Several difficulties and inaccuracies when using Child's classification have been detailed in [12].

Based on a dataset of patients collected at the University Clinic of Navarra, Spain, this study sets out to predict survival within 6 months after TIPS placement coupled with a reduction in the number of features required. For this purpose, several machine learning methods related to Bayesian networks (naive Bayes, selective naive Bayes, semi naive Bayes, tree augmented naive Bayes, and k-dependence Bayesian classifier) are adapted to perform a FSS process and applied to the patients' database, which contains structured and standardized medical histories, clinical examinations, and complementary tests.

This paper is organized as follows. The data is described in Section 2. In Section 3, the Bayesian classifiers performed to carry out the study are presented, including both filter and wrapper approaches. The experimental results of the study are reported in Section 4. Finally, a brief set of conclusions and future work in this area are described in Section 5.

2. Patients: cases and variables

From May 1991 to September 1998, 134 patients suffering from liver cirrhosis underwent TIPS placement at the University Clinic of Navarra, Spain. In all the cases, the diagnosis of cirrhosis was based on liver histology. However, only 127 patients were included in the study due to medical reasons.

The indications for TIPS placement were prophylaxis of rebleeding (68 patients), refractory ascites (28 patients), prophylaxis of bleeding (11 patients), acute bleeding refractory to endoscopic and medical therapy (10 patients), portal vein thrombosis (9 patients), and Budd–Chiari syndrome (1 patient).

The prospective study includes 107 patients because 20 patients underwent liver transplants within the first 6 months after TIPS placement. Bearing in mind that the aim of the study is to predict survival within the first 6 months after TIPS placement, the follow-up of these patients was rejected on the day of the transplant. The inclusion of patients who underwent liver transplants influence the results. The survival prediction of the Bayesian classification models might be modified by the surgical mortality related to transplantation. On the other hand, transplant patients may live longer than patients who do not undergo TIPS. According to a study [13], survival in patients who undergo transplantation is significantly improved compared with those who do not undergo transplantation.

Patients who are recipients of transplantation contribute data that is considered correctly censored in survival analysis. Building a model to predict survival in 6 months after TIPS placement, these cases have to be removed to avoid biases. In order to predict patients who will undergo liver transplant, a study of factors such as the availability of a matching donor, the position in the queue for transplantation, and the probability that the patient will benefit from the transplant, is needed. Besides, such a study would require a large number of transplanted patients.

The database contains 77 clinical findings—see Table 1—for each patient. These 77 attributes were measured before TIPS placement. A new binary variable called *vital-status* is created. It reflects whether or not the patients died within the first 6 months after the

Table 1
Attributes of the database

History finding attributes:		
Age	Gender	Height
Weight	Etiology of cirrhosis	Indication of TIPS
Bleeding origin	Number of bleedings	Prophylactic therapy with propranolol
Previous sclerotherapy	Restriction of proteins	Number of hepatic encephalopathies
Type of hepatic encephalopathy	Ascites intensity	Number of paracenteses
Volume of paracenteses	Dose of furosemide	Dose of spironolactone
Spontaneous bacterial peritonitis	Kidney failure	Organic nephropathy
Diabetes mellitus		
Laboratory finding attributes:		
Hemoglobin	Hematocrit	White blood cell count
Serum sodium	Urine sodium	Serum potassium
Urine potassium	Plasma osmolarity	Urine osmolarity
Urea	Plasma creatinine	Urine creatinine
Creatinine clearance	Fractional sodium excretion	Diuresis
GOT	GPT	GGT
Alkaline phosphatase	Serum total bilirubin (mg/dl)	Serum conjugated bilirubin (mg/dl)
Serum albumin (g/dl)	Platelets	Prothrombin time (%)
Partial thrombin time	PRA	Proteins
FNG	Aldosterone	ADH
Dopamine	Norepinephrine	Epinephrine
Gammaglobulin		
CHILD score		
PUGH score		
Doppler sonography:		
Portal size	Portal flow velocity	Portal flow right
Portal flow left	Spleen length (cm)	
Endoscopy:		
Size of oesophageal varices	Gastric varices	Portal gastropathy
Acute hemorrhage		
Hemodynamic parameters:		
Arterial pressure (mm Hg)	Heart rate (beats/min)	Cardiac output (L/min)
Free hepatic venous pressure	Wedged hepatic venous pressure	Hepatic venous pressure gradient (HVPG)
Central venous pressure	Portal pressure	Portosystemic venous pressure gradient
Angiography:		
Portal thrombosis		

placement of TIPS. The values for this variable correspond to both classes of the problem. Within the first 6 months after TIPS placement, 33 patients died and 74 survived for a longer period, thus reflecting that the utility and consequences of TIPS were not the same for all the patients. Hence, the objective of the study is to build Bayesian classification models that discriminate between these two subgroups of patients.

The study was approved by the Local Ethics Committee and informed oral consent was obtained from all patients.

3. Feature selection in Bayesian classifiers

3.1. Preliminaries

The main goal of a supervised classification algorithm is to build a classification model using a dataset. The classification model can be seen as a function, γ, that assigns class values, $\{1, 2, \ldots, r_0\}$, to an instance vector, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, that is

$$\gamma : (x_1, \ldots, x_n) \rightarrow \{1, 2, \ldots, r_0\}.$$

In the case of 0/1 loss function, the Bayesian classifier assigns the *most a posteriori probable class* to a given instance, that is

$$\gamma(\boldsymbol{x}) = \arg\max_c p(c|x_1, \ldots, x_n).$$

Using the Bayes formula [14], $p(c|x_1, \ldots, x_n)$ can be calculated as follows:

$$p(c|x_1, \ldots, x_n) = \frac{p(c, x_1, \ldots, x_n)}{p(x_1, \ldots, x_n)} = \frac{p(c)p(x_1, \ldots, x_n|c)}{p(x_1, \ldots, x_n)}.$$

As we assume the equiprobability of all the instances, then:

$$p(c|x_1, \ldots, x_n) \propto p(c)p(x_1, \ldots, x_n|c).$$

Bayesian classification models by means of Bayesian networks [15–18] could provide interpretability and simplicity to the supervised classification task. Medical researchers are usually unfamiliar with supervised classification techniques. The graphical representation of Bayesian classifiers is intuitive, allowing them to understand the underlying probabilistic classification process.

A model hierarchy of increasing complexity can be established for the Bayesian classifiers, where the naive Bayes is at the bottom and a general Bayesian network is at the top of this hierarchy. The restrictions imposed on naive Bayes, selective naive Bayes, semi naive Bayes, tree augmented naive Bayes, and k-dependence Bayesian classifier are due to the type of relations between variables that they consider. In spite of their limitations, Bayesian classifiers provide a set of properties that can be appreciated by medical staff. Their graphical structure facilitates the interpretability and understanding by clinicians, reflecting probabilistic relationships between domain variables. The conditioned and marginal probabilities of the model can be of interest to physicians who want to better understand the uncertainty of the studied medical domain. Another interesting characteristic of these classifiers is that, when computational time is a critical factor, these Bayesian classifiers may be quickly learned from a database by means of the original inductors. Furthermore, once the Bayesian classification model is induced, it is quickly able to obtain a prediction for an unseen example.

In FSS, the aim of the search process was to maximize the accuracy of the classifier. The functions known as *filter* functions carry out this goal by looking only at the intrinsic characteristic of the data and measuring the power to discriminate among the classes. The induction process based on these filter functions is usually called the *filter approach*. However [6], reports that FSS should depend not only on the features and the concept to learn, but also on the classifier. Thus, the *wrapper* approach has been developed: when a feature subset is selected by the search algorithm, its estimated accuracy—estimated using the supervised classification model—is considered to guide the search process. The previous ideas borrowed from FSS are adapted in this paper to the induction of Bayesian classification models. Fig. 1 shows the differences between the wrapper and filter approaches for Bayesian classifiers induction.

When the filter approach is considered, a function independent of the characteristics of the classifier must be set. The *mutual information* of two variables is a filter function defined as $I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$, where the statistics are computed from the data. It can be described as the reduction in uncertainty about one variable due to the knowledge of the other variable. The mutual information between a predictive variable X and the class C, $I(X,C)$, that is, the reduction in uncertainty about C due to the knowledge of the predictive variable X is largely applied for classification purposes. It is known that $2N\ I(X,C)$ (where N is the

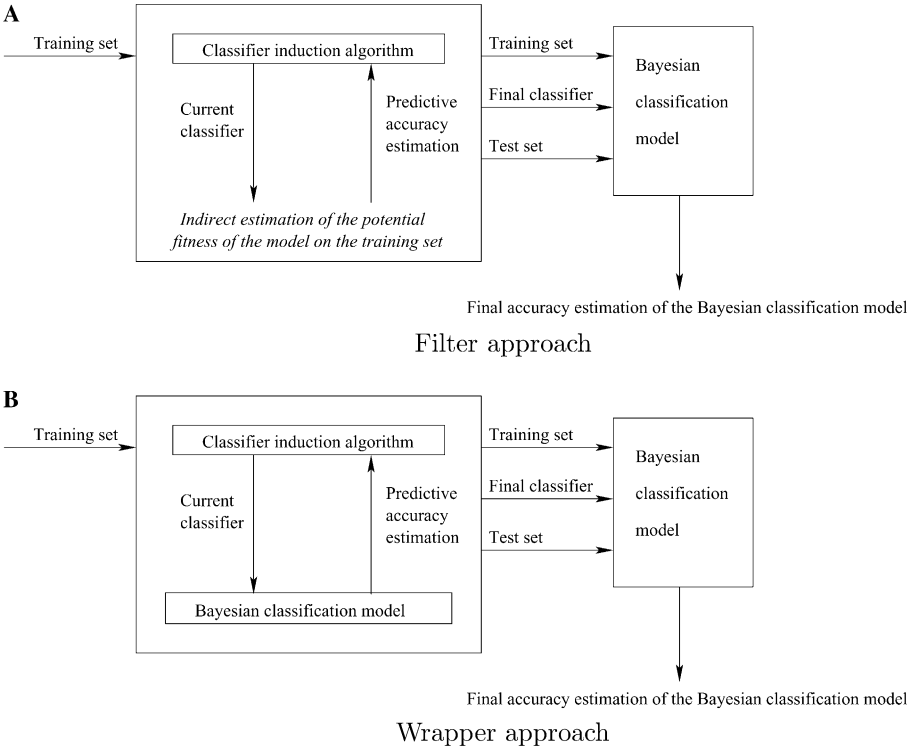

Fig. 1. General schemes for classifier induction: *Filter* (A) and *Wrapper* (B) approaches.

number of input instances) asymptotically follows a $\chi^2_{(r_i-1)(r_0-1)}$ distribution, where r_i is the number of values of the X_i variable and r_0 is the number of values of the class.

This result can be applied to run a FSS. However, in this paper, it is adapted to guide the induction of different types of Bayesian classifiers with an implicit feature selection. Thus, only the variables which meet the criterion with a certain level of significance, α, are taken into account for the induction of the Bayesian classifier. The aim of the search process is to maximize the accuracy of the Bayesian classification models indirectly by means of the percentile of the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test.

When a wrapper approach is considered in the construction of the classifier, the estimated accuracy of the seen instances is the function which guides the search process.

3.2. Naive Bayes

Naive Bayes [19] is a Bayesian supervised classification algorithm built from the assumption of conditional independence of the predictive variables given the class. Although this assumption is violated in numerous occasions in real domains, this fact does not degrade the performance of the paradigm in many situations [20,21]. Making this assumption, the prediction of the class for an unseen instance is simplified. In Fig. 2, a graphical representation of the structure of a naive Bayes is shown.

The naive Bayes classifier uses the Bayes theorem to predict the category for each unseen instance. To classify a new patient represented by $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, in our two-category problem—survive or not more than 6 months—$C = \{c_0, c_1\}$, where c_0 states that the patient does not survive more than 6 months after TIPS placement and c_1 implies that the patient does, the naive Bayes classifier uses the following formula:

$$c^* = \arg \max_{c \in \{c_0, c_1\}} p(C = c) \prod_{i=1}^{n} p(X_i = x_i | C = c),$$

where $p(X_i = x_i | C = c)$ represents the conditional probability of $X_i = x_i$ given that $C = c$, when all features have discrete values.

The naive Bayes classifier has a large tradition of application in medical areas. Immediately after its presentation in the 60s, Bailey [22] proposes the use of this classifier to carry out medical diagnosis tasks. During the 70s and 80s, the naive Bayes became popular in medical applications. Nordyke et al. [23] presents the use of naive Bayes to automated diagnosis of thyroid dysfunction. Russek et al. [24] solves a classification task related to heart disease. An interesting study proposed by Kononenko [25] compares the result of naive Bayes, a tree inductor called *Assistant* and four medical specialists in four medical domains: location of primary tumors, persistent breast cancer, thyroid diseases, and rheumatism domain.

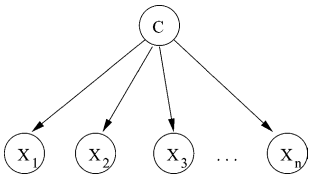

Fig. 2. Representation of the naive Bayes classifier. $p(c|x_1, \ldots, x_n) \propto p(c) \prod_{i=1}^{n} p(x_i|c)$.

3.3. Selective naive Bayes

In all problem domains, irrelevant features would degrade the predictive accuracy of learning algorithms. Variables in which the information contribution is overlapped or repeated would act in the same way. Due to the assumption of the independence of the variables given the class, algorithms such as naive Bayes are robust with respect to irrelevant variables but very sensitive to correlated variables.

Accuracy would be improved if the learning algorithm only used adequate variables [26], that is, not redundant variables that decrease the accuracy of the classifier. For this purpose, a variable selection process is required. FSS would be used to find a feature subset that maximized the predictive accuracy of the classification model built over this subset. From this point of view, FSS would be faced as a search problem where each point of the search space represents a variable subset [26].

This combination of FSS and naive Bayes is known as *selective naive Bayes* [26]. A Bayes classification model whose structure is similar to a naive Bayes is built, but not all the predictive variables are used by the classifier. Fig. 3 presents a selective naive Bayes structure.

When the filter method is used, bearing in mind the result referred to $2N\,I(X_i, C)$, a $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test is performed. The selective naive Bayes model is built with the predictive variables which pass the test. Fig. 4 shows the pseudo-code for this approach.

The forward sequential selection wrapper approach—see Fig. 5—starts with an empty set of variables. At each step the method adds the most accurate

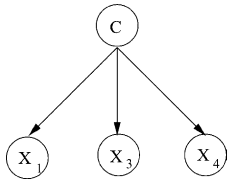

Fig. 3. Representation of the selective naive Bayes classifier for a domain with six predictive variables. $p(c|x_1, x_2, x_3, x_4, x_5, x_6) \propto p(c)p(x_1|c)p(x_3|c)p(x_4|c)$.

Selective NB_{fs}
Step 1: Let be the variable list, S, empty
Step 2: Repeat until all the variables have been seen
2.1: If $2N\,I(X_i, C)$ passes the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test then add X_i in S

Fig. 4. Pseudo-code for the filter approach to selective naive Bayes.

Selective NB_{ws}
Step 1: Let be the variable list, S, empty
Step 2: Repeat until non improvement is reached
2.1: Select the most accurate predictive variable not in S
2.2: Add the selected variable in S

Fig. 5. Pseudo-code for the wrapper approach to selective naive Bayes [26].

variable in terms of estimated accuracy. The algorithm stops when non-improvement is reached.

3.4. Semi naive Bayes

The selective naive Bayes algorithm is able to detect irrelevant and redundant variables, but it cannot notice dependencies between predictive variables. Furthermore, there are some well-known databases where the naive Bayes obtains a poor performance [27], perhaps because of its inability to discover any relationship between variables. A possible explanation of this matter is that the assumption of conditional independence is violated. To tackle this issue, the *semi naive Bayes* model [28] was proposed. By means of the constructive induction concept, a model with naive Bayes structure is built via wrapper greedy approach. The semi naive Bayes joins predictive variables in a super-variable via cartesian product, that is, a super-variable consisting of a cartesian product of a subset of variables. Fig. 6 shows the structure of a semi naive Bayes classification model.

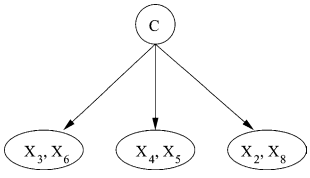

Fig. 6. Representation of the semi naive Bayes classifier for a domain with eight predictive variables. $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \propto p(c)p(x_3, x_6|c)p(x_4, x_5|c)p(x_2, x_8|c)$.

The induction of a semi naive Bayes classifier can be seen as a typical local search process where the objective function is accuracy. Starting with an empty structure and until non-improvement in the objective function is reached, at each step of the induction process the algorithm considers the best option between:

- adding a new variable X_{new} conditionally independent of the structure
- replacing a variable X_i by joining of X_i to a new variable X_{new}, resulting in $X_{\text{join}} = (X_i, X_{\text{new}})$

Fig. 7 shows the Forward Sequential Selection and Joining (FSSJ) algorithm proposed by Pazzani [27]. An analogous backward version, called Backward Sequential Elimination and Joining (BSEJ) was also proposed by the same author.

FSSJ
Step 1: Let be the variable list, S, empty
Step 2: Repeat until non improvement is reach
2.1: Select the 'best' option
(a): Consider each predictive variable not in S a new variable conditionally independent of the current classifier
(b): Consider joining each predictive variable not in S to each predictive variable in S
2.2: Add the selected variable in S

Fig. 7. Pseudo-code for the FSSJ algorithm [27].

Semi NB_{fs}
Step 1: Let be the variable list, S, empty
Step 2: Repeat until no variables $2N\,I(X_i, C)$ pass the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test
2.1: Select the option with higher percentile
(a): Consider each predictive variable X_{max} not in S such as $2N\,I(X_{max}, C)$ passes the $\chi^2_{(r_{max}-1)(r_0-1);1-\alpha}$ test as a new variable conditionally independent of the current classifier
(b): Consider joining each predictive variable X_{new} not in S to each predictive variable X_i in S such as $2N\,I((X_i, X_{new}), C)$ passes the $\chi^2_{((r_i r_{new})-1)(r_0-1);1-\alpha}$ test
2.2: Add the selected variable in S

Fig. 8. Pseudo-code for the filter approach to semi naive Bayes.

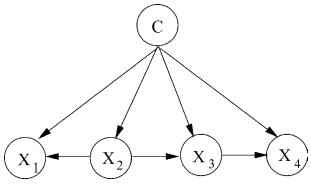

Fig. 9. Representation of the tree augmented naive Bayes classifier with four predictive variables. $p(c|x_1, x_2, x_3, x_4) \propto p(c)p(x_1|x_2, c)\,p(x_2|c)p(x_3|x_2, c)p(x_4|x_3, c)$.

The original formulation of the semi naive Bayes classifier is adapted to perform the search by means of the filter approach presented before. Therefore, the *FSSJ* schema is followed, but the 'best' option is the one with the highest statistical significance for the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test. It must be noted that, in the case of super-variables, the degrees of freedom of the performed test change. To summarize, when a new variable is considered to be part of the classifier, the $X_{\max}$ with the largest level of significance for $2N\,I(X_i, C)$ is selected. On the other hand, when a joining step is considered, the X_{join} with the largest level of significance for $2N\,I((X_i, X_{\text{new}}), C)$, where $X_{\text{join}} = (X_i, X_{\text{new}})$, is selected. Fig. 8 shows the pseudo-code for this approach to semi naive Bayes.

3.5. Tree augmented naive Bayes

The *tree augmented naive Bayes* (TAN) [29] builds a classifier where a probabilistic tree-like structure, built among the predictive variables, is extended with a naive Bayes structure. Fig. 9 displays an example of a TAN classifier structure.

The method proposed by Friedman et al. [29] forces to construct a connected tree structure with all the variables of the problem domain. In our proposed filter approach, not all the variables are taken into account to build the classification model. Moreover, a 'forest' structure is allowed: several tree structures can connect different and disjoint subsets of predictive variables.

In order to carry out the proposed filter approach—see Fig. 10—the subset of domain variables which pass the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test are selected to perform the Chow–Liu algorithm [30]. Then, the edges that pass the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test are added to the undirected graph. The non-existence of a statistician with a known distribution to fix the significance of $2N\,I(X_i, X_j|C)$ (where X_i and X_j are the variables related to the arc) means a problem. To solve this, in our approach it is required that $2N_c\,I(X_i, X_j|C = c)$ (N_c is the number of cases where $C = c$) passes the $\chi^2_{(r_i-1)(r_j-1);1-\alpha}$ test to some value c of the class variable.

TAN_{fs}
Step 1: Select the subset S of variables such as $2N\,I(X_i, C)$ passes the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test
Step 2: Compute $I(X_i, X_j\,|\,C)$ between each pair of variables in S with $i < j$, $i, j = 1, \ldots, |S|$
Step 3: Build a complete undirected graph in which the nodes are the predictive variables in S. If $2N_c\,I(X_i, X_j|C = c)$ passes the $\chi^2_{(r_i-1)(r_j-1);1-\alpha}$ test to some value c, then set the weight of an edge connecting X_i to X_j by $I(X_i, X_j\,|\,C)$
Step 4: Follow the Kruskall algorithm to build a maximum weighted spanning tree from the previous undirected graph. If the graph is not connected, repeat this step for each connected component
Step 5: Transform the resulting undirected tree(s) to a directed tree(s) by choosing a variable as the root and setting the direction of the edges
Step 6: Construct a TAN model by adding a C vertex and adding an arc from C to each predictive variable in S

Fig. 10. Pseudo-code for the filter approach to TAN.

TAN_{ws}
Step 1: Let be the variable list, S, empty
Step 2: Sequentially select the two most accurate predictive variables, and add them to S
Step 3: Repeat until non improvement is reached
3.1: Select the 'best' option
(a): Consider each predictive variable not in S a new variable conditionally independent of the current classifier
(b): Consider each arc which does not invalidate the forest TAN condition between two predictive variables in S
3.2: Add the corresponding decision

Fig. 11. Pseudo-code for the wrapper approach to TAN.

A wrapper method to obtain a TAN structure was proposed by Keogh and Pazzani [31]. In order to reduce the search space, their approach starts with a naive Bayes structure, and a 'legal' arc (restricting the number of parents: the class and another predictive variable) is added at each step until not accuracy improvement is attained.

In this paper, a different novel wrapper greedy approach is presented. As in the proposed filter approximation, a 'forest' structure can be obtained and not all the predictive variables must be involved in a single tree structure. Fig. 11 presents the novel wrapper algorithm suggested to obtain a TAN.

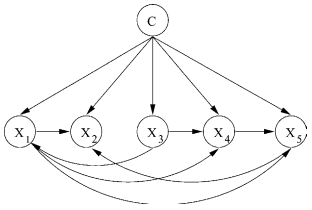

Fig. 12. Representation of a structure for a k-dependence Bayesian classifier with five predictive variables and $k = 2$. $p(c|x_1, x_2, x_3, x_4, x_5, x_6) \propto p(c)p(x_1|x_3, c)p(x_2|x_1, x_5, c)p(x_3|c)p(x_4|x_1, x_3, c)p(x_5|x_1, x_4, c)$.

The new wrapper TAN method can be seen as the semi naive Bayes wrapper approach where the joining of two variables are swapped to include an arc between two variables. This means that, starting with an empty set of variables, after the addition of two variables in a greedy way, the algorithm decides whether to add a new variable or to create an arc between two variables in the model.

3.6. k-Dependence Bayesian classifier

The tree augmented naive Bayes classification model is limited by the number of parents of the predictive variables. A predictive variable can have a maximum of two parents: the class and another predictive variable. The k-dependence Bayesian classifier (kDB) [32] tries to avoid this restriction by allowing a predictive variable to have up to k parents besides the class. In Fig. 12 a kDB structure can be seen.

The most restrictive condition of a kDB structure is the number of parents of a variable: a number of k must be fixed. A way to automatically identify a good value for k for each variable of a given problem is desirable. The proposed filter approach for kDB is analogous to

$k\text{DB}_{fs}$
Step 1: Select the subset V of variables such that $2N\,I(X_i, C)$ passes the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test
Step 2: Compute $I(X_i, X_j|C)$ between each pair of variables in V with $i < j$, $i, j = 1, \ldots, |V|$
Step 3: Let the used variable list, S, be empty
Step 4: Let the k-dependence Bayesian classifier being constructed, kDB, begin with a single class node, C
Step 5: Repeat until S includes all predictive variables in V
5.1: Select the variable X_{max} in V which is not in S and has the largest $I(X_{max}, C)$
5.2: Add a node to kDB representing X_{max}
5.3: Add an arc from C to X_{max} in kDB
5.4: Add $m = \min(|S|, k)$ arcs from m different variables X_j in S with the highest value for $I(X_{max}, X_j|C)$ if $2N_c\,I(X_{max}, X_j|C = c)$ passes the $\chi^2_{(r_i-1)(r_j-1);1-\alpha}$ test to some value c
5.5: Add X_{max} to S

Fig. 13. Pseudo-code for the filter approach to kDB.

kDB$_{ws}$
Step 1: Let be the variable list, S, empty
Step 2: Sequentially select the two most accurate predictive variables, and add them to S
Step 3: Repeat until non-improvement is reached
3.1: Select the 'best' option
(a): Consider each predictive variable not in S a new variable conditionally independent of the current classifier
(b): Consider each arc which not invalidate the kDB condition between two predictive variables in S
3.2: Add the corresponding decision

Fig. 14. Pseudo-code for the wrapper approach to kDB.

the TAN filter approach, where k is the maximum number of parents for each variable. Furthermore, several kDB structures can connect different and disjointed subsets of variables. The original algorithm is applied over the subset of predictive variables which pass the $\chi^2_{(r_i-1)(r_0-1);1-\alpha}$ test. In the same way, only the arcs whose $2N_C\, I(X_i, X_j|C=c)$ pass the $\chi^2_{(r_i-1)(r_j-1);1-\alpha}$ test to some value c of the class variable are added to the final model. Fig. 13 displays the filter approach to kDB classifier.

The original kDB algorithm is based on the calculation of mutual information and conditional mutual information statistics. Therefore, it can be regarded as a filter approach. A kDB wrapper approach tries to identify the most accurate subset of variables and arcs, where k is the maximum number of parents of a predictive variable. Fig. 14 shows the proposed wrapper approach to obtain a kDb structure.

The wrapper kDB procedure is similar to the wrapper TAN approach, but when the addition of an arc is considered, the kDB restrictions must not be violated.

4. Experimental results

The aim of this work is to reach the highest accuracy with a feature reduction when identifying the subgroup of surviving patients 6 months after TIPS placement, coupled with the reliability and satisfaction of the medical staff with the Bayesian classification models. We focus our empirical study on the accuracy of the proposed Bayesian classification models. However, the number of selected features and, in the case of the wrapper approaches, the number of evaluations required are also reported. The ROC curves of the proposed classifiers are presented. In order to validate the Bayesian classification models a *leave-one-out* cross-validation [33] are performed. When a wrapper approach is used, a 5-fold cross-validation is performed as the internal accuracy estimation which guides the search for the best model.

The parameters of all proposed Bayesian classification models are estimated applying the Laplace correction [34] to their maximum likelihood parameter estimations. The α parameter of the proposed filter approaches is fixed at $\alpha = 0.01$.

The study database contains missing data and continuous variables. The presented Bayesian classifiers are implemented to manage complete discrete data. Therefore, the imputation of the missing values is carried out replacing the missing value by the mean (when the variable is continuous) or the mode (when the variable is discrete), conditioned to the value of the class. Continuous variables of the dataset are discretized by means of the *Equal Frequency* [35] algorithm into two intervals.

The Elvira software [36] is used in the implementation of the presented Bayesian classification models. The cited imputation and discretization algorithms are included in the Elvira package.

Table 2 shows the estimated average accuracy (and its standard deviation) and the number of features of the induced classifier for the Bayesian classifiers are presented. In the case of the wrapper approaches, the number of evaluations due to the computational cost is also displayed. The results support the fact that not all the features are needed to learn an accurate classification model, that is, the feature reduction increases the accuracy.

Table 2
Average results: estimated accuracy and number of features of the induced classifier

	Accuracy	No. features	No. evaluations
Naive Bayes	88.78 ± 3.06	77	
TAN	88.78 ± 3.06	77	
kDB	88.78 ± 3.06	77	
Selective NB$_{fs}$	93.46 ± 2.40	11	
Semi NB$_{fs}$	93.46 ± 2.40	11	
TAN$_{fs}$	93.46 ± 2.40	11	
kDB$_{fs}$	93.46 ± 2.40	11	
Selective NB$_{ws}$	92.52 ± 2.55	3	302
Semi NB$_{ws}$	93.46 ± 2.40	5	1111
TAN$_{ws}$	90.65 ± 2.82	4	395
kDB$_{ws}$	92.52 ± 2.55	4	395

In order to compare statistical differences in the behavior of the different Bayesian classification models, the Mann–Whitney test [37] is performed. In spite of the fact that no statistically significant differences were observed, the filter approaches of all the Bayesian classification models improve the estimated accuracy with respect to original methods and wrapper approaches (except semi naive Bayes classifier). Achieved estimated accuracies are competitive with previous studies [13,38,39] and are considered 'acceptable' by physicians.

Results from this analys is involving leave-one-out cross-validation are optimistic. However, the purpose is to compare Bayesian classification models, so their performance ranks will be preserved even in the presence of overfitting.

In order to select a classifier, researchers must take into account not only the estimation of accuracy, but also the time complexity or the number of features included in the models.

A Receiver Operating Characteristic (ROC) analysis [40] is related to cost-benefit analysis of diagnosis decision making. Widely used in biomedical applications, ROC analysis provides tools to select classifiers independently from the cost context or the class distribution. A ROC curve supplies a concise graphic depiction of the overall performance of a classifier plotting the true positive rate against the false positive rate for the different possible cut-points.

Fig. 15 shows the ROC curves for the proposed Bayesian classification models comparing each classifier without feature selection to the corresponding filter and wrapper approaches. Looking at the curves figures of the ROC, it must be noted that the filter approaches of all the proposed Bayesian classification models reach slightly higher sensibility and specificity than wrapper approaches and the classifiers without feature selection.

Besides the accuracy and the ROC curves, the calibration concept could be taken into account to select a

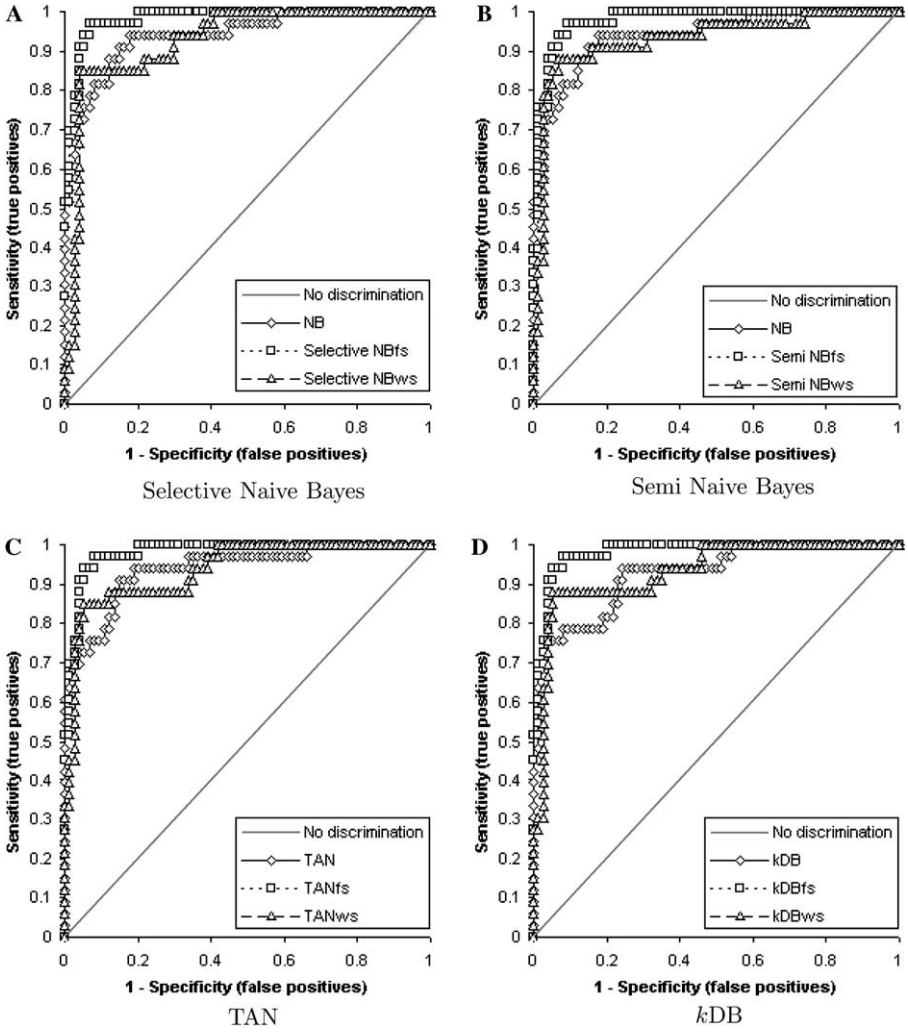

Fig. 15. ROC curves for the proposed Bayesian classification models comparing the models without FSS with the corresponding filter and wrapper approaches.

classification model. The calibration score describes how close the estimates of the model are to the true underlying probabilities [41]. As the true probability is unknown, we cannot directly calculate a calibration score.

The Brier score [42], following the formulation in [43], is defined as:

$$bs(D) = \frac{1}{N}\sum_{l=1}^{N}\sum_{k=1}^{r_0}(p(C = c_k|\boldsymbol{X} = \boldsymbol{x}_l) - \delta_{l,k})^2,$$

where

$$\delta_{l,k} = \begin{cases} 1 & c_l = c_k \\ 0 & \text{otherwise} \end{cases}$$

denotes whether c_l, the real value of the class in the instance $\boldsymbol{x}_l$ equals c_k, the estimated value of the class. The Brier score could be seen as the squares error of the class prediction for each instance.

The Brier scoring measure of each final classification model is used to estimate the classifier calibration. It is also calculated by means of a *leave-one-out* cross-validation. Table 3 shows the average Brier score (and standard deviation) for the proposed Bayesian classification models. Looking at the Brier score values, the filter approaches of the Bayesian classifiers can be considered the most calibrated models.

Table 4 shows the variables included in the final Bayesian classification model built over the entire dataset. Note that the proposed filter methods select the same 11 variables.

Although wrapper approaches are time-consuming algorithms, the achieved dimensionality reduction is significant in contrast to original algorithms and filter approaches. Physicians note that the dimensionality reduction affects data acquisition, reducing the extra costs of the medical tests (limiting the number of tests, the cost is automatically reduced) and the number of invasive medical techniques (endoscopy, angiography, and hemodynamic test in the majority of the Bayesian classifier models), which are not required. Most of the selected variables are related to the patient's history and laboratory findings. While this subgroup of variables (history and laboratory findings) are easily obtained without any significant inconvenience for patients, the inconveniences for patients increase with the rest of the medical tests, particularly with the hemodynamic test, where a catheter is introduced to examine the state of the vein. In spite of the requirement of a

Table 3
Average Brier score and its standard deviation for the proposed Bayesian classifiers

	Brier score
Naive Bayes	0.1016 ± 0.281
TAN	0.1034 ± 0.288
kDB	0.0990 ± 0.280
Selective NB_{fs}	0.0529 ± 0.193
Semi NB_{fs}	0.0553 ± 0.189
TAN_{fs}	0.0532 ± 0.198
$k DB_{fs}$	0.0536 ± 0.196
Selective NB_{ws}	0.0705 ± 0.185
Semi NB_{ws}	0.0597 ± 0.149
TAN_{ws}	0.0834 ± 0.211
$k DB_{ws}$	0.0780 ± 0.206

Table 4
List of variables included in the Bayesian classifiers

	Filter	Sel NB_{ws}	Semi NB_{ws}	TAN_{ws}	$k DB_{ws}$
History finding attributes:					
Etiology of cirrhosis			×		
Indication of TIPS		×	×	×	×
Spontaneous bacterial peritonitis	×				
Laboratory finding attributes:					
Plasma osmolarity	×	×	×	×	×
Creatinine clearance	×				
Serum albumin (g/dl)	×				
Aldosterone	×				
ADH	×				
Epinephrine				×	
PUGH score	×				
Doppler sonography:					
Portal flow left	×				
Portal flow velocity			×		
Spleen length (cm)	×				
Hemodynamic parameters:					
Arterial pressure (mm Hg)	×				
Heart rate	×	×	×	×	×
Hepatic venous pressure gradient (HVPG)					×

hemodynamic test to obtain the value of the HVPG variable, it is only introduced by the $k\text{DB}_{ws}$ classifier.

Comparing this analysis with a previous study of the same database presented in [39], the variables selected by the Bayesian classifiers are not exactly the same subset of features. However, the selected medical variables presented in this study are interrelated with the selected variables in [39].

Physicians acknowledge that 'subjective' variables (where the value of the variable is determined by the medical staff) were not selected in the wrapper Bayesian classification models. This means that the final results of the Bayesian classifiers are not influence by physicians' point of view unlike the traditionally used Child–Pugh score, which is a collection of five variables, where two of them are based on medical opinion.

In relation to the set of a posteriori probability distributions, the conditional probabilities related to Bayesian classification models roughly assert the previous medical knowledge about cirrhosis.

When the Bayesian classifiers are presented to the medical staff, physicians notice an improvement in comprehensibility and simplicity among the models induced by filter and wrapper approaches with respect to the original models induced by the whole set of variables. In other words, the dimensionality reduction, carried out in the filter and wrapper approaches, reduces the complexity of the final classifiers. This fact provides a useful classification tool that can be easily used in medical practice.

5. Conclusions and future work

A set of filter and wrapper approaches, which perform a feature selection process, to building Bayesian classification models are presented. As far as we know, a subset of these models is novel approaches. A real medical problem is tackled with the proposed Bayesian classification inducers. The filter and wrapper approaches to selective naive Bayes, semi naive Bayes, tree augmented naive Bayes, and k-dependence Bayesian classifier are proposed to estimate the survival of cirrhotic patients after TIPS placement.

The purpose of this work is not to formalize the framework for variable selection procedures, but rather to compare performance of FSS methods in different models that predict survival with IPS.

Although the wrapper approaches attain a larger dimensionality reduction (from 77 original variables, only 3 are required to build a model for some of the Bayesian classifiers considered), the average accuracy of the filter approaches is slightly better in general. These accuracy results are supported by the ROC curves, where the filter approaches show a higher sensibility and specificity.

The dimensionality reduction provides the medical staff with simpler and comprehensible models that can be used in everyday practice. Physicians note that this dimensionality reduction decreases the economical costs and, to the patients' further advantage, several medical tests are not required. Moreover, physicians note the non-occurrence of 'subjective' variables (those variables whose value is determined by the physicians) in the final classifiers.

In the future, to avoid the local optima of the greedy wrapper search the use of evolutionary computation (specifically the *estimation of distribution algorithms*) in combination with wrapper approaches is planned.

Principled comparisons of sensitivity of different types of models to the curse of dimensionality are rare. The purpose of this article is to define which methods seem to be best for the data set at hand. Other methods such as support vector machines may be evaluated in future work.

The Brier score can help to choose the 'best' classifier. It seems to be related to the accuracy of the Bayesian classifier. When this score is applied, the objective is to minimize the Brier score instead of maximize the accuracy. Furthermore, a multiobjective search can be performed where other measures (like *conditional log-likelihood* or *penalized log-likelihood*) can be included.

Acknowledgments

The authors thank the anonymous reviewers whose suggestions improved both the content and the presentation of this paper. This work is partially supported by the Ministry of Science and Technology, by the Fondo de Investigaciones Sanitarias, by the Basque Government, by the University of the Basque Country and by the Provincial Council of Guipuzkoa under Grants TIC2001-2973-C05-03, PI021020, ETORTEK-GENMODIS and ETORTEK-BIOLAN projects, 9/UPV 00140.226-15334/2003 and 760/2003, respectively.

References

[1] Kononenko I, Bratko I, Roška E. Experiments in automatic learning of medical diagnosis rules, in: Proceedings of the international school for the synthesis of expert's knowledge workshop, 1984.

[2] Ohmann C, Moustakis V, Yang Q, Lang K. Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. Artif Intell Med 1996;8:23–36.

[3] Cooper G, Aliferis C, Ambrosio R, Aronis J, Buchanan B, Caruana R, et al. An evaluation of machine learning methods for predicting pneumonia mortality. Artif Intell Med 1997;9:107–38.

[4] Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28–36.

[5] Cios K, Moore G. Uniqueness of medical data mining. Artif Intell Med 2002;26:1–24.
[6] Kohavi R, John G. Wrappers for feature subset selection. Artif Intell 1997;97(1–2):273–324.
[7] Draper D, Fouskakis D. A case study of stochastic optimization in health policy: problem formulation and preliminary results. J Global Optim 2000;18:399–416.
[8] Jelonek J, Stefanowski J. Feature subset selection for classification of histological images. Artif Intell Med 1997;20(11–13): 1202–9.
[9] Bornman P, Krige J, Terblanche J. Management of oesophageal varices. Lancet 1994;343:1079–84.
[10] Saunders J, Walters J, Davies P, Paton A. A 20-year prospective study of cirrhosis. Br Med J 1981;282:263–6.
[11] Pugh R, Murray-Lion I, Dawson J, Pictioni M, Williams R. Transection of the esophagus for bleeding oesophageal varices. Br J Surg 1973;60:646–9.
[12] Conn H. A peek at the Child–Turcotte classification. Hepatology 1981;1:1–7.
[13] Malinchoc M, Kamath P, Gordon F, Peine C, Rank J, ter Borg P. A model to predict survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology 2000;31(4):864–71.
[14] Bayes T. Essay towards solving a problem in the doctrine of chances, The Phylosophical Transactions of the Royal Society of London; 1764.
[15] Castillo E, Gutiérrez J, Hadi A. Expert Systems and Probabilistic Network Models. New York: Springer-Verlag; 1997.
[16] Jensen F. Bayesian Networks and Decision Graphs. New York: Springer Verlag; 2001.
[17] Neapolitan RE. Learning Bayesian Networks. Englewood Cliffs, NJ: Prentice-Hall; 2003.
[18] Pearl J. Probabilistic Reasoning in Intelligent Systems. Los Altos, CA: Morgan Kaufmann; 1988.
[19] Minsky M. Steps toward artificial intelligence. Trans Instit Radio Engineers 1961;49:8–30.
[20] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 1997;29(2–3):103–30.
[21] Hand D, You K. Idiot's Bayes—not so stupid after all? Int Stat Rev 2001;69:385–98.
[22] Bailey N. Probability methods of diagnosis based on small samples. Math Comput Sci Biol Med 1964:103–7.
[23] Nordyke R, Kulikowski C, Kulikowski C. A comparison of methods for the automated diagnosis of thyroid dysfunction. Comput Biomed Res 1971;4:374–89.
[24] Russek E, Kronmal R, Fisher L. The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. Comput Biomed Res 1983;16:537–52.
[25] Kononenko I. Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition, in: Wielinga B, Boose J, Gaines B, Shereiber G, van Someren M, editors. Current Trends in Knowledge Acquisition. IOS Press: Amsterdam; 1990.
[26] Langley P, Sage S. Induction of selective Bayesian classifiers, in: Proceedings of the tenth conference on uncertainty in artificial intelligence. Morgan Kaufmann; 1994, pp. 399–406.
[27] Pazzani M. Searching for dependencies in Bayesian classifiers, in: Artificial intelligence and Statistics IV, Lecture Notes in Statistics. New York: Springer-Verlang; 1997.
[28] Kononenko I. Semi-naïve Bayesian classifiers, in: Proceedings of the 6th European working session on learning, 1991, pp. 206s–19s.
[29] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29(2):131–64.
[30] Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Trans Inform Theory 1968;14:462–7.
[31] Keogh E, Pazzani M. Learning augmented Bayesian classifiers: a comparison of distributions-based and classification-based approaches, in: Uncertainty 99: The 7th international workshop on artificial intelligence and Statistics, 1999, pp. 225–30.
[32] Sahami M. Learning limited dependence Bayesian classifiers, in: Proceedings of the 2nd international conference on knowledge discovery and data mining, 1996, pp. 335–38.
[33] Stone M. Cross-validatory choice and assessment of statistical predictions (with discussion). J R Stat Soc B 1974;36:111–47.
[34] Laplace P. Philosophical Essays on Probabilities, Springer-Verlag; 1995, translation of A.I. Dale from the 5th French edition of 1825.
[35] Cattlet J. On changing continuous attributes into ordered discrete attributes, in: Proceedings of the European working session on learning, 1991, pp. 164–78.
[36] Elvira Consortium, Elvira: an environment for probabilistic graphical models, in: Gámez J, Salmerón A. editors. Proceedings of the first European workshop on probabilistic graphical models, 2002, pp. 222–30.
[37] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50–60.
[38] Chalasani N, Clark W, Martin L, Kamean J, Khan M, Patel N, et al. Determinants of mortality in patients with advanced cirrhosis after transjugular intrahepatic portosystemic shunting. Gastroenterology 2000;118:138–44.
[39] Inza I, Merino M, Larrañaga P, Quiroga J, Sierra B, Girala M. Feature subset selection by genetic algorithms and estimation of distributions algorithms. A case study in the survival of cirrhotic patients treated with TIPS. Artif Intell Med 2001;23:187–205.
[40] Hanley J, McNeil B. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[41] Vinterbo S. Predictive models in medicine: some methods for construction and adaptation, Norwegian University of Science and Technology, 1999, Ph.D. Thesis.
[42] Brier G. Verification of forecasts expressed in terms of probabilities. Mon Weather Rev 1950;78:1–3.
[43] van der Gaag L, Renooij S. Evaluation scores for probabilistic networks, in: Proceedings of the 13th Belgium–Netherlands conference on artificial intelligence, Amsterdam, The Netherlands: Universiteit van Amsterdam; 2001, pp. 109–16.