ELSEVIER Artificial Intelligence in Medicine 23 (2001) 187–205

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Feature subset selection by genetic algorithms and estimation of distribution algorithms A case study in the survival of cirrhotic patients treated with TIPS

I. Inza[a,*], M. Merino[b], P. Larrañaga[a], J. Quiroga[c], B. Sierra[a], M. Girala[c]

[a]*Department of Computer Science and Artificial Intelligence, P.O. Box 649, University of the Basque Country, E-20080 Donostia-San Sebastián, Spain*

[b]*Basque Health Service — Osakidetza, Comarca Gipuzkoa — Este, Avenida Navarra 14, E-20013 Donostia-San Sebastián, Spain*

[c]*Facultad de Medicina, University Clinic of Navarra, E-31080 Pamplona-Iruña, Spain*

Received 20 June 2000; received in revised form 7 July 2000; accepted 28 May 2001

Abstract

The transjugular intrahepatic portosystemic shunt (TIPS) is an interventional treatment for cirrhotic patients with portal hypertension. In the light of our medical staff's experience, the consequences of TIPS are not homogeneous for all the patients and a subgroup dies in the first 6 months after TIPS placement. Actually, there is no risk indicator to identify this subgroup of patients before treatment. An investigation for predicting the survival of cirrhotic patients treated with TIPS is carried out using a clinical database with 107 cases and 77 attributes. Four supervised machine learning classifiers are applied to discriminate between both subgroups of patients. The application of several feature subset selection (FSS) techniques has significantly improved the predictive accuracy of these classifiers and considerably reduced the amount of attributes in the classification models. Among FSS techniques, FSS–TREE, a new randomized algorithm inspired on the new EDA (estimation of distribution algorithm) paradigm has obtained the best average accuracy results for each classifier. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Feature subset selection; Transjugular intrahepatic portosystemic shunt; Estimation of distribution algorithm; Indications; Survival

[*] Corresponding author. Tel.: +34-943015026; fax: +34-943215396.
E-mail address: ccbincai@si.ehu.es (I. Inza).

PII: S0933-3657(01)00085-9

1. Introduction

Portal hypertension is a major complication of chronic liver disease. By definition, it is a pathological increase in the portal venous pressure which results in formation of portosystemic collaterals that divert blood from the liver to the systemic circulation. This is caused by both an obstruction to outflow in the portal flow as well as an increased mesenteric flow. In the western world, cirrhosis of the liver accounts for approximately 90% of the patients.

Of the sequelae of portal hypertension (i.e. varices, encephalopathy, hypersplenism, ascites), bleeding from gastro-oesophageal varices is a significant cause of early mortality (approximately 30–50% at the first bleed) [4,41].

Many efforts have been made over the past decades in the treatment of portal hypertension. This has resulted in an increasing number of randomized trials and publications but, unfortunately, therapeutic decision is not easy [10].

The transjugular intrahepatic portosystemic shunt (TIPS) is an interventional treatment resulting in decompression of the portal system by creation of a side-to-side portosystemic anastomosis. Since its introduction over 10 years ago [39,40] and despite the large number of published studies, many questions remain unanswered. Currently, little is known about the effects of TIPS on the survival of the treated patients.

Our medical staff has found that a subgroup of patients dies in the first 6 months after a TIPS placement and the rest survive for longer periods. Actually there is no risk indicator to identify both subgroups of patients. We are equally interested in the detection of both subgroups, giving the same relevance to the reduction of both error types.

A period of 6 months is chosen as we think that beyond that period, factors such as stenosis of the shunt and possible variceal rebleeding as a consequence, would compound the analysis. Furthermore, a critical criteria for choosing this period is that the average waiting time on a list for a liver transplant at the University Clinic of Navarra is approximately 6 months. The only published study [31] to identify a subgroup of patients who die within a period after a TIPS placement fixes the length of this period to 3 months. However, we think that our specific conditions really suggest lengthening this period to 6 months.

Traditionally, Pugh's modification of the Child–Turcotte score (referred to as the Child–Pugh score) has been used to assess risk in patients undergoing portosystemic shunt surgery [37]. Although it is a classic score to assess the level of seriousness of a patient's liver disease, it has inherent problems when applied to patients undergoing TIPS and it cannot be used to predict which patients will die within a certain period of time and which patients will survive that period. The several difficulties and innacuracies in applying the Child–Pugh score to predict survival periods have been detailed by Conn [8].

In 1980s and 1990s, researchers in artificial intelligence have developed new machine learning methods that construct predictive models from data, obtaining promising results in several clinical areas [9,18,35]. As far we know, these kinds of techniques have never been applied to TIPS indication or contraindication. Thus, we assume that the prediction of patient survival within 6 months after elective TIPS is well-suited to supervised machine learning methods.

We have performed a prospective study, building up a database which includes a structured and standardized history and clinical examination. In the study reported here, we concentrate on the task of predicting survival within 6 months after a TIPS placement of hospitalized patients from their findings before a TIPS setting. For this purpose, in the first step, four well-known supervised machine learning methods with a long tradition on medical applications such as a Naive Bayes classifier, a decision-tree technique, a rule-learning procedure and a nearest neighbor method are applied.

However, the used database has a large set of measured findings[1] and some of them seem to be irrelevant or redundant. It is well known that the accuracy of supervised machine learning methods is not monotonic regarding the inclusion of features [26]: irrelevant or redundant attributes, depending on the specific characteristics of the classifier, may degrade the accuracy of the classification model. Ohmann et al. [35], in a problem of an acute abdominal pain diagnosis, note that the high dimensionality of their study database is the major problem in order to improve the predictive accuracy of their supervised classification models. In this sense, given the entire set of attributes, we aim to find the attribute subset with the best predictive accuracy for a certain classifier. This problem is known in the machine learning community as the feature subset selection (FSS) problem and it has been tackled with success in different medical areas [13,21]. A reduction in the number of variables is of interest as classification models with a smaller number of variables may be more quickly and easily used by clinicians, as these models would require a lower data input [9]. Models with a relatively small number of variables may be more readily converted into paper-based models that could be used widely in current medical practice. Other interesting effects of the dimensionality reduction are the decrease in the cost of adquisition of the data and the rise in the interpretability and comprehensibility of the classification models.

Thus, in the second stage of the study, we apply two sequential and two genetic FSS techniques over the same survival predictive problem. We extend our comparison with the application of two new FSS procedures based on the new EDA (estimation of distribution algorithm) [33] paradigm.

Although the application of two new EDA-inspired FSS techniques refers to the specific medical problem, the proposed approach is general and can be used for other tasks where supervised machine learning algorithms face a high number of irrelevant and/or redundant features.

Costs of medical tests are not considered in the construction of classification models and predictive accuracy maximization is the principal goal of our research. As the cost of the TIPS placement is not insignificant, our study is developed to help physicians, counsel patients and their families before deciding to proceed with elective TIPS.

The paper is organized as follows. The study database is described in Section 2. The supervised classifiers included in the study and the FSS methods to improve their predictive accuracy are described in Section 3. Experimental results are presented in Section 4. The last section briefly summarizes the work and presents ways of future research in the field.

[1] We use the terms ‘finding’, ‘attribute’, ‘feature’ and ‘variable’ indistinctly. Since ‘finding’ is a medical word, ‘attribute’, ‘feature’ and ‘variable’ come from the Statistics community.

2. Patients: study database

The prospective study includes 127 patients with liver cirrhosis who underwent TIPS from May 1991 to September 1998 in the University Clinic of Navarra, Spain. The diagnosis of cirrhosis was based in liver histology in all cases.

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).

Table 1
Attributes of the study database

History finding attributes		
Age	Gender	Height
Weight	Etiology of cirrhosis	Indication of TIPS
Bleeding origin	Number of bleedings	Prophylactic therapy with popranolol
Previous sclerotherapy	Restriction of proteins	Number of hepatic encephalopathies
Type of hepatic encephalophaty	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)	Plateletes	Prothrombin time (%)
Parcial thrombin time	PRA	Proteins
FNG	Aldosterone	ADH
Dopamine	Norepineohrine	Epinephrine
Gamma-globulin		
Child score		
Pugh score		
Doppler sonography		
Portal size	Portal flow velocity	Portal flow right
Portal flow left	Spleen lenght (cm)	
Endoscopy		
Size of esophageal varices	Gastric varices	Portal gastropathy
Acute hemorrhage		
Hemodinamic 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	Portosystemis venous pressure gradient
Angiography		
Portal thrombosis		

The analysis includes 107 patients as 20 underwent liver transplants the first 6 months after TIPS placement. The follow-up of these transplanted patients was censored on the day of the transplant. This censoring was done to remove the effect of transplantation when modeling the 6-months survival of patients who undergo TIPS. If these patients were not censored, deaths due to surgical mortality related to transplantation might have influenced the selection of variables that are prognostic for the TIPS procedure. On the other hand, transplantation may prolong survival compared with patients who do not undergo TIPS. It is predictably found that survival in patients who undergo transplantation is significantly improved compared with those who do not undergo transplantation [31].

The database contains 77 clinical findings for each patient. These 77 attributes were measured before TIPS placement (see Table 1). A new binary variable is created, called *vital-status*, which reflects whether the patient died in the first 6 months after the placement of the TIPS or not: this variable reflects both classes of the problem. In the first 6 months after the placement of the TIPS, 33 patients died and 74 survived for a longer period, thus reflecting that the utility and consequences of the TIPS were not homogeneous for all the patients.

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

3. Methods of automatic knowledge acquisition

3.1. Supervised classifiers

In the study, four well-known machine learning supervised classifiers, with completely different approaches to learning, were applied to predict the survival of cirrhotic patients for the first 6 months after the setting of the TIPS. All the algorithms were selected due to their simplicity and their long standing tradition in medical diagnose studies.

The Naive–Bayes (NB) rule [5] uses the Bayes theorem to predict the category for each case, assuming that the attributes are independent given the category. To classify a new patient characterized by d attributes $\boldsymbol{X} = (X_1, X_2, \dots, X_d)$ in our two-category problem $C = \{c_1, c_2\}$, where c_1 implies that the patient survives more than 6 months and c_2 implies that the patient does not survive more than 6 months, the NB classifier applies the following rule:

$$c_{\mathrm{NB}} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{d} p(x_i|c_j)$$

where c_{NB} denotes the category value output by the NB classifier (in our problem, c_1 or c_2). The probability for nominal features is estimated from data using maximum likelihood estimation and applying the Laplace correction. A normal distribution is assumed to estimate the class conditional probabilities for continuous attributes. Unknown values in the test instance are skipped. Despite its simplicity, the NB rule has obtained better results than more complex algorithms in many medical domains. Many researchers think that the success of the NB rule is based on the idea that doctors, in order to make a diagnosis, collect the attributes in the same way as the NB rule uses them to classify: that is to say, independently with respect to the category.

The *CN2* [7] represents a classification model by a set of IF–THEN rules, where the THEN part represents the class predicted for the instances that match the conditions of the IF part. CN2 is based on an information theoretic approach with a significance metric to improve rule quality and to avoid overspecialization of results. When a significant rule is found, CN2 removes those examples it covers from the training set and adds the rule to the end of the rule list. To use induced rules to classify test examples, CN2 tries each rule in order until one is found whose conditions are satisfied by the example being classified. If no induced rules are satisfied, the final default rule assigns the most common class in the training set to the test case.

The *C4.5* [38] represents a classification model by a decision tree. The tree is constructed in a top–down way, dividing the training set and beginning with the selection of the best attribute in the root of the tree. The selection of the best attribute is based on an informatic theoretic approach. A descendant of the root node is then created for each possible value of the selected attribute, and the training cases are sorted to the appropiate descendant node. The entire process is then recursively repeated using the training cases associated with each descendant node to select the best attribute to test at that point in the tree. The process stops at each node of the tree when all cases in that point of the tree belong to the same category or the best split of the node does not surpass a fixed Chi-square significancy threshold. Then, the tree is simplified by a pruning mechanism to avoid overspecialization. Even though a decision tree can be converted into a set of IF–THEN rules, while CN2 rules are independent to each other, C4.5 rules are dependent on each other. The *IB1* [1] is a case-based, nearest-neighbor classifier. To classify a test instance, all training instances are stored and the nearest training instance regarding the test instance is found: its class is retrieved to predict this as the class of the test instance. To measure the distance between two instances, an euclidean distance measure is used for continuous attributes and an overlap metric for nominal ones.

Apart from the prediction accuracy, the explanation ability of the classifier is also very important. To support the diagnostic process in everyday practice, physicians need a classifier that is able to explain its decisions. Such transparent decisions are much more acceptable by physicians. For this reason, other promising techniques, such as neural nets, are not included among our classification models, due to their low human-transparency [32].

Due to the low number of patients (107), when no FSS method is applied, the *leave-one-out* [24] procedure is used to estimate the accuracy of supervised classifiers. In the leave-one-out technique, the learning algorithm is run k times, where k is the number of patients of the database. Each time $k-1$ instances are used for training and the remaining patient is used for testing, where each patient is used only once for testing. The leave-one-out estimate of accuracy is the overall number of correct classifications, divided by k, the number of patients in the dataset.

Experiments were run in a SUN-SPARC computer using MLC++ [25] Machine Learning library of programs for the presented classifiers.

3.2. Feature subset selection methods

The basic problem of machine learning is concerned with the induction of a model that classifies a given object into one of several known classes. In order to induce the classification model, each object is described by a pattern of *d* features. Here, the machine learning community has formulated the following question: "are all of these *d* descriptive features useful for learning the 'classification rule'?". On trying to respond to this question, we come up with the FSS [30] approach which can be reformulated as follows: "given a set of candidate features, select the 'best' subset in a classification problem". In our case, the 'best' subset will be the one with the best predictive accuracy.

Most of the supervised learning algorithms perform rather poorly when faced with many irrelevant or redundant (depending on the specific characteristics of the classifier) attributes. In this way, the FSS proposes additional methods to reduce the number of features so as to improve the performance of the supervised classification algorithm.

FSS can be viewed as a search problem [28], with each state in the search space specifying a subset of the possible features of the task. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice due to the large amount of computational effort required. In this way, any feature selection method must determine four basic issues that define the nature of the search process:

1. *The starting point in the space*. It determines the direction of the search. One might start with no features and successively add them, or one might start with all features and successively remove them. One might also select an initial state somewhere in the middle of the search space.
2. *The organization of the search*. It determines the strategy of the search. Roughly speaking, the search strategies can be *complete* or *heuristic* (see [30] for a review of FSS algorithms). The basis of the *complete search* is the systematic examination of every possible feature subset. Three classic complete search implementations are depth-first, breadth-first and branch and bound search [34]. On the other hand, among *heuristic* algorithms, there are *deterministic heuristic* and *non-deterministic heuristic* algorithms. Classic deterministic heuristic FSS algorithms are sequential forward selection (SFS) and sequential backward selection (SBS) [23], floating selection methods (SFFS and SFBS [36]) and best-first search [26]. They are deterministic in the sense that all the runs over the same data always obtain the same solution. *Non-deterministic heuristic* search appears in a motivation to avoid getting stuck in local maximum. Randomness is used to escape from local maximum and this implies that one should not expect the same solution from different runs. Two classic implementations of non-deterministic search engines are Genetic Algorithms (GA) [42] and Simulated Annealing [12]. Genetic Algorithms are possibly the most commonly used search engine in the FSS task.
3. *Evaluation strategy of feature subsets*. The evaluation function identifies the promising areas of the search space. The objective of FSS algorithm is its maximization. The search algorithm uses the value returned by the evaluation function for helping to guide the search. Some evaluation functions carry out this objective looking only at the characteristics of the data, capturing the relevance of each feature or set of features

to define the target concept: these type of evaluation functions are grouped below the *filter* strategy. However, John et al. [22] reported that when the goal of FSS is the maximization of the accuracy, the features selected should depend not only on the features and the target concept to be learned, but also on the learning algorithm. Thus, they proposed the *wrapper* concept: this implies that the FSS algorithm conducts a search for a good subset using the induction algorithm itself as a part of the evaluation function, the same algorithm that will be used to induce the final classification model. Once the classification algorithm is fixed, the idea is to train it with the feature subset found by the search algorithm, estimating the accuracy and assigning it as the value of the evaluation function of the feature subset. In this way, representational biases of the induction algorithm which are used to construct the final classifier are included in the FSS process. It is claimed by many authors [26,30] that the wrapper approach obtains better predictive accuracy estimates than the filter approach.

4. *Criterion for halting the search*. An intuitive approach for stopping the search will be the non-improvement of the evaluation function value of alternative subsets. Another classic criterion will be to fix an amount of possible solutions to be visited along the search.

3.2.1. Four classic feature subset selection methods

We have used the following well-known FSS methods in the experimentation phase.

- SFS is a classic hill-climbing search algorithm [23] which starts from an empty subset of features and sequentially selects features until no improvement is achieved in the evaluation function value. It performs the major part of its search near the empty feature set.
- SBE is another classic hill-climbing algorithm [23] which starts from the full set of features and sequentially deletes features until no improvement is achieved in the evaluation function value. It performs the major part of its search near the full feature set.
- GA search with one-point crossover (GA-o).
- GA search with uniform crossover (GA-u).

Genetic Algorithms (GAs) [19] are one of the best known techniques for solving optimization problems. The GA is a population based search method. First, a population of individuals[2] (in our case feature subsets) is generated, then promising individuals are selected, and finally new individuals which will form the new population are generated using crossover and mutation operators. On the other hand, SFS and SBE, instead of working with a population of solutions, try to optimize a single feature subset.

Although the optimal selection of parameters is still an open problem on GAs [17], for both GA algorithms, guided by the recommendations of Bäck [2], the probability of crossover is set to 1.0 and the mutation probability to $1/d$, being d the number of variables of the domain (these values are so common in the literature). Fitness-proportionate selection [16] is used to select individuals for crossover. The population size is set to

[2] The terms ‘individual’ and ‘solution’ are used indistinctly.

```
EDA
    $D_0 \leftarrow$ Generate $N$ individuals (the initial population) randomly.
    Repeat for $l = 1, 2, \ldots$ until a stop criterion is met.
        $D^s_{l-1} \leftarrow$ Select $S \leq N$ individuals from $D_{l-1}$ according to a selection method.
        $p_l(\mathbf{x}) = p(\mathbf{x}|D^s_{l-1}) \leftarrow$ Estimate the joint probability distribution of an individual
        being among the selected inviduals.
        $D_l \leftarrow$ Sample $N$ individuals (the new population) from $p_l(\mathbf{x})$.
```

Fig. 1. Main structure of the EDA approach.

1000 and the new population is formed by the best members from both old population and offspring.[3] The criterion for halting the genetic search is the following: GA-o and GA-u stop when in a sampled new population of solutions no individual is found with an evaluation function value that improves the best individual found in the previous generation. Thus, the best solution of the previous population is returned as the result of the genetic search. We want to control the risk of overspecialization with this severe criteria.

3.2.2. FSS–PBIL and FSS—TREE: two new EDA-inspired methods to select features

GAs are mainly criticized for three aspects [29]:

- the large number of parameters and their associated refered optimal selection or tuning process;
- the extremely difficult prediction of the movements of the populations in the search space;
- their incapacity to solve the well-known deceptive problems [16].

In an attempt to solve the previous problems using an evolutionary and population-based search method, the EDA [29,33] appears. In EDA, there are no crossover nor mutation operators: the new population is sampled from a probability distribution which is estimated from the selected individuals. Fig. 1 shows the basic structure of the EDA approach.

The EDA algorithm can be used to solve the FSS problem, representing each individual of the EDA search as a possible feature subset solution. A common notation can be used to represent an individual (or feature subset): for a full *d* feature problem, there are *d* bits in each individual, each bit indicating whether a feature is present (1) or absent (0).

[3] As 'offspring' is known the set of newly created solutions.

The main problem of EDA resides on how the joint d-dimensional probability distribution $p_l(\boldsymbol{x})$ is estimated. Obviously, the computation of 2^d probabilities (for a domain with d binary variables) is impractical. Bayesian networks [14] could be an attractive paradigm to make the factorization of the probability distribution of best individuals. However, due to the large amount of attributes in our database, a huge number of individuals is needed to induce a reliable Bayesian network [15].

In our study, we have used PBIL [3] and dependency-trees [6], two simple and well-known probabilistic models. PBIL assumes that all attributes of the database are independent to each other.[4] In PBIL, each variable is examined independently, and the probability distribution to sample each variable of an individual of the new population is learned in the following way:

$$p_l(x_i) = (1-\alpha)p_{l-1}(x_i|D_{l-1}) + \alpha p_{l-1}(x_i|D^{\mathrm{s}}_{l-1})$$

where $p_{l-1}(x_i|D_{l-1})$ is the probability distribution of the variable i in the old population, $p_{l-1}(x_i|D^{\mathrm{s}}_{l-1})$ the probability distribution of the variable i among selected individuals and 'α' a user parameter which guarantees the evolution. It is fixed to 0.5.

A dependency tree assumes some kind of dependencies among the attributes of the database, restricting $p_l(\boldsymbol{x})$ to factorizations in which the conditional probability distribution for any one bit depends on the value of, at the most, one other bit. In Bayesian network terms, this means we are restricting our probability models to networks in which each node can have one parent at the most. We use the method proposed by Chow and Liu [6] to find the optimal model within these restrictions. The induced tree is optimal in the sense that among all possible trees, its probabilistic structure maximizes the likelihood of selected solutions when they are drawn from any unknown distribution. We call the application to the feature subset selection problem of both search algorithms as FSS–PBIL and FSS–TREE. Thus, Fig. 2 summarizes both FSS–PBIL and FSS–TREE approaches: they only differ in the learned probabilistic model (fifth step).

As GA FSS approaches, FSS–PBIL and FSS–TREE are also randomized and population-based FSS algorithms. The absence of crossover and mutation operators (implicit to GAs) to evolve the population is one of their biggest attractions. A population size of 1000 individuals is setup for both algorithms and they use the same stop criteria as both GA approaches. They also form the new population from the best members from both old population and sampled new population.

3.2.3. Evaluation function of FSS methods

To assess the goodness of each proposed feature subset for a specific classifier, a wrapper approach is applied. In the same way as supervised classifiers when no feature selection is applied, this wrapper approach estimates, by the leave-one-out procedure, the goodness of the classifier using only the feature subset proposed found by the search algorithm. Thus, the study database is projected maintaining the values of the selected features and the class variable *vital-status* for the whole of 107 patients: over this projected dataset the goodness of the proposed feature subset using the specific classifier is estimated by the explained leave-one-out estimation technique.

[4] Assuming the following factorization: $p_l(\boldsymbol{x}) = \prod_{i=1}^{d} p_l(x_i)$.

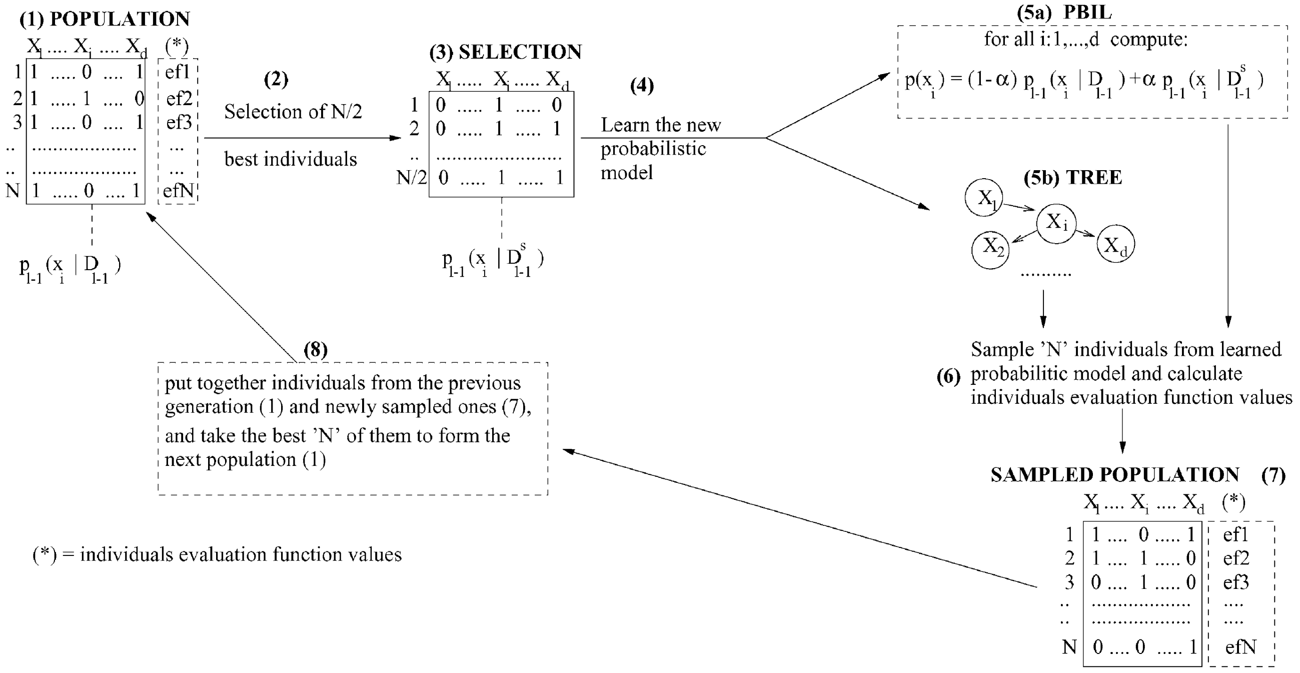

Fig. 2. FSS–PBIL and FSS–TREE algorithms. Both algorithms only differ in the fifth step.

Table 2
Estimated accuracy percentage coupled with the standard deviation of the leave-one-out process (first row) and the cardinalities of finally selected feature subsets (second row)[a]

	No-FSS	SFS	SBE	GA-o	GA-u	FSS–PBIL	FSS–TREE
NB	75.70 ± 4.17	82.24 ± 3.71	80.37 ± 3.86	85.05 ± 3.46	85.05 ± 3.46	85.05 ± 3.46	86.92 ± 3.28
	77	3	72	8	9	8	10
CN2	69.92 ± 4.96	81.15 ± 3.59	79.53 ± 4.27	84.97 ± 2.57	84.16 ± 2.52	83.22 ± 3.97	85.88 ± 2.85
	32	6	30	5	9	10	7
C4.5	61.66 ± 4.62	85.19 ± 2.66	80.00 ± 4.37	85.73 ± 4.22	86.11 ± 3.63	85.19 ± 2.66	87.95 ± 2.63
	6	3	6	3	4	3	5
IB1	67.96 ± 4.20	80.19 ± 4.19	72.59 ± 3.84	84.26 ± 2.86	85.93 ± 3.10	85.19 ± 2.66	86.85 ± 3.19
	77	2	74	8	5	5	5

[a] For randomized FSS algorithms, the results of the most accurate run are reflected.

4. Experiments

SFS and SBE are deterministic algorithms which are only run once for each classifier. Due to their randomized nature, GA-o, GA-u, FSS–PBIL and FSS–TREE are run 10 times for each classifier. Coupled with the leave-one-out estimation of the predictive accuracy of four classifiers without feature selection and SFS and SBE selection methods, Table 2 also reflects the leave-one-out accuracy estimation of the best run of each randomized FSS method. Apart from the standard deviation of the leave-one-out estimation, Table 2 reflects the cardinality of the best feature subset for each FSS method and classifier. Note that when no FSS method is applied, CN2 and C4.5 classifiers can discard a subset of the features on their own.

As randomized FSS techniques are run 10 times, Table 3 reflects the average accuracy percentages and cardinality of selected subsets of these 10 runs, coupled with the associated standard deviations of these averages.

Several conclusions can be extracted from the results of Tables 2 and 3.

- With the aid of FSS techniques, all supervised classifiers obtain better accuracy levels than the no-FSS approach using feature subsets that need <85% of the attributes

Table 3
Average accuracy percentages (first row) and cardinalities of finally selected feature subsets (second row) of 10 runs for randomized FSS methods[a]

	GA-o	GA-u	FSS–PBIL	FSS–TREE
NB	84.67 ± 0.51	84.48 ± 0.51	85.05 ± 0.00	86.28 ± 0.54
	10.2 ± 3.11	9.4 ± 2.88	8.8 ± 2.86	10.6 ± 0.57
CN2	84.43 ± 0.40	84.12 ± 0.06	83.10 ± 0.20	85.26 ± 0.54
	9.0 ± 3.46	11.0 ± 3.46	8.3 ± 1.52	8.3 ± 2.30
C4.5	85.30 ± 0.37	85.30 ± 0.70	85.09 ± 0.13	86.34 ± 1.38
	4.0 ± 1.73	3.3 ± 0.57	3.5 ± 0.52	3.7 ± 0.95
IB1	84.12 ± 0.09	84.53 ± 1.25	84.31 ± 1.00	85.51 ± 1.23
	5.80 ± 1.82	5.9 ± 1.85	4.1 ± 0.87	5.3 ± 1.25

[a] The standard deviations of accuracy and cardinality averages are also reported.

of the whole feature set. This dimensionality reduction and accuracy improvement is possible due to the large amount of correlations that appear among the findings of the study database. 162 pairwise correlations statistically significant at the $^{**}P < 0.01$ level and 168 correlations at $^{*}P < 0.05$ by means of the Pearson coefficient are detected among the medical findings of the study database. As a clear example, *Child score* and *Pugh score* variables are a linear combination of other five features of the study. FSS methods are able to mainly detect these groups of correlated features that hurt the accuracy level of supervised classifiers. Our physicians collected a large number of features they thought could affect the survival of the patients. However, FSS algorithms are able to discover these correlations with just a small portion of them, the supervised classifiers are able to acceptably discriminate between both categories of the problem.

- The dimensionality reduction is coupled with significant accuracy improvements regarding the accuracy of the whole feature set. For all supervised classifiers, when a cross-validated paired *t*-test [11] is applied between the no-FSS approach and each run of any FSS method (except SBE with NB and IB1), accuracy differences are always statistically significant at the $^{*}P < 0.05$ level. In the case of SBE with NB and IB1, accuracy differences are only statistically significant at the $P < 0.1$ level.
- Closely related to the previous points, the poorest predictive accuracy of SBE among FSS methods can be explained: SBE performs the major part of its search near the full feature set and as the database has the best accuracy levels near the empty feature set, it seems that SBE is unable to escape from local optimas in its optimization process and aid the supervised classifiers to discard the correlated features. Thus, the cardinalities of the feature subsets selected by SBE are close to those obtained by supervised classifiers without the aid of any FSS method (no-FSS).
- Due to the low number of patients and the intrinsically high standard deviation, it is unlikely to establish statistically significant accuracy differences among the FSS algorithms. Despite the non-significance of accuracy differences, randomized FSS methods achieve better accuracy levels for all supervised classifiers than SFS and SBE. However, in the case of the C4.5 classifier, the accuracy level achieved by SFS is close to the accuracy of randomized FSS methods. The explanation for these accuracy differences among sequential and population-based FSS methods appears in the works of Vafaie and De Jong [43] and Kudo and Sklansky [27], where the tendency of greedy-like sequential searches as SFS and SBE to get trapped on local peaks of the search space is highlighted: thus, the use of population based search methods as more robust search engines in the FSS problem is defended.
- The C4.5 tree classifier, due to its own pruning process, induces the model with least features. Without the aid of any FSS method, C4.5 builds a tree with just six features: however, this induced tree has in its root node, a feature (*prothrombin time*) that is hardly correlated to the class, but combining it on the tree with other features, the overall accuracy can not be largely improved. On the other hand, SFS and randomized FSS methods performing a slight dimensionality reduction, are able to discard this correlated feature and find more accurate feature combinations to build the tree (this phenomenon of correlated features that build poor decision trees is noted by Kohavi and John [26] in the *Corral* artificial dataset).

Table 4
For each supervised classifier, the amount of the best 10 randomized runs which belong to each FSS method

	GA-o	GA-u	FSS–PBIL	FSS–TREE
NB	0	0	0	10
CN2	2	1	0	7
C4.5	1	2	1	6
IB1	0	2	2	6

- The low standard deviation of the average accuracy of 10 runs for each randomized FSS method must be noted: this gives us an idea of the stability of the models induced by these methods. Among all the algorithms, FSS–TREE, apart from the single classification models with the best predictive accuracy for four supervised classifiers, achieves the best average accuracy results. Table 4 presents an intuitive comparison among four randomized methods. As each randomized FSS method is run 10 times, we have 40 runs among all randomized methods: Table 4 shows have many of the best 10 out of these 40 runs belong to each method. Table 4 clearly states the stability in the superior accuracy achieved by the predictive models induced by the features selected by FSS–TREE: for all supervised classifiers, more than half of the best 10 subsets found by randomized FSS methods belong to FSS–TREE.
- We note a high stability degree in the medical findings selected by FSS methods. Although the low standard deviation in the number of findings selected by randomized FSS methods gives us an idea of stability, our objective is to find out which specific attributes are selected by different runs of a randomized FSS method. As FSS–TREE is the method with the best average accuracy for all classifiers, its behaviour regarding the stability in the selection of features is analyzed. As FSS–TREE is run 10 times for each classifier, Table 5 reflects the amount of runs in which each medical finding is included in the final models. The attributes that appear in the best subset are also indicated. Table 5 shows a high degree of stability in the attributes selected by the 10 runs of FSS–TREE for each classifier. This stability in the selection of features also occurs for the rest of the randomized FSS methods. As the wrapper approach is used to assure the optimality of selected features with respect to the specific supervised classifier, we analyze each classifier separately.
  - ○ NB uses, in 10 runs, 16 different features with a mean of 10.6 ± 0.57 features per execution. Six out of 10 models are minor variations of each other and the rest do not highly vary. The presence of *parcial thrombin time, PRA* and *gamma-globulin* findings in all the executions must be noted.
  - ○ CN2 uses 21 different features with a mean of 8.3 ± 2.30 features per execution. The 10 models can be divided into three groups of near identical rule sets within each group of executions. The presence of the *previous sclerotherapy* finding is noted in all the executions.
  - ○ C4.5 uses eight different features with a mean of 3.7 ± 0.95 features per execution. All the models are minor variations of each other and six runs output the same tree. The *gamma-globulin* finding appears in all the trees.

Table 5
For FSS–TREE, this table lists the amount of runs that each medical finding appears in the models induced by each classifier

	NB	CN2	C4.5	IB1
History finding attributes				
Gender	6	0	0	3
Weight	4[a]	0	0	0
Etiology of cirrhosis	0	3	0	0
Indication of TIPS	0	4[a]	0	0
Previous sclerotherapy	0	10[a]	0	0
Restriction of proteins	0	0	3	0
Number of hepatic encephalopathies	0	0	3	0
Dose of furosemide	3	0	0	0
Spontaneous bacterial peritonitis	6	3	0	2
Kidney failure	0	0	0	4[a]
Organic nephropathy	4[a]	6	0	4
Laboratory finding attributes				
Hematocrit	3	0	2[a]	0
White blood cell count	4[a]	0	0	0
Urine sodium	0	0	0	3
Serum potassium	7[a]	0	0	3
Urine potassium	0	3	0	0
Plasma osmolarity	0	0	0	3
Urine osmolarity	0	4[a]	0	0
Urea	0	3	0	0
Creatinine clearance	3	0	0	0
Fractional sodium excretion	0	0	0	3
Diuresis	0	5[a]	0	0
GOT	0	0	8	0
GPT	0	3	0	0
Serum total bilirubin (mg/dl)	0	2	0	0
Serum conjugated bilirubin (mg/dl)	4[a]	0	0	0
Serum albumin (g/dl)	3	0	0	0
Plateletes	0	3	0	0
Prothrombin time (%)	0	3	5	4[a]
Parcial thrombin time	10[a]	4[a]	2[a]	0
PRA	10[a]	0	0	0
Proteins	6	0	0	0
FNG	0	3	0	0
Aldosterone	4[a]	0	0	0
Epinephrine	0	0	2[a]	3
Gamma-globulin	10[a]	3	10[a]	0
Child score	0	7[a]	0	10[a]
Pugh score	6	6	0	3
Doppler sonography				
Portal flow left	0	0	0	4[a]
Spleen lenght (cm)	3	0	0	0
Endoscopy				
Portal gastropathy	0	4[a]	0	0
Hemodinamic parameters				
Free hepatic venous pressure	4[a]	0	0	0
Wedged hepatic venous pressure	3	0	0	0

Table 5 (*Continued*)

	NB	CN2	C4.5	IB1
Hepatic venous pressure gradient (HVPG)	0	3	0	0
Central venous pressure	0	0	2[a]	0
Angiography				
Portal thrombosis	0	0	0	4[a]

[a] The findings that appear in the best found feature subset. We do not reflect the attributes that are not never selected by FSS–TREE.

○ IB1 uses 14 different features with a mean of 5.3 ± 1.25 features per execution. The *Child score* finding appears in all the models. The 10 models can be divided into three groups of near identical rule sets within each group.

When the classification models are presented to the medical staff, they noted a large improvement in comprehensibility among the models induced with the aid of FSS techniques and those that are constructed without FSS. The dimensionality reduction carried out by the FSS process has reduced the complexity and the amount of variables to be input to the classification models, converting them to paper-based models [9] which can be more easily used in everyday practice. Thus, by this dimensionality reduction, the confidence and acceptance in the models of our medical staff is increased.

Physicians, apart from the high accuracy levels, highlight the transparency and user-executable (ability for mental check) levels [32] of the graphical output produced by decision trees and the set of decision rules. They also judge, as medium–high, the same characteristics of the NB classifier. On the other hand, they qualify these qualities for the nearest-neighbor classifier as modest.

With the reduction in the number of needed measurements, an obvious reduction of the derived economic costs is achieved. We also have a lower amount of possibly troublesome medical test for the future patients.

The values of misclassification matrices of probabilities obtained by the FSS processes are also satisfactory for our physicians. Misclassification matrices do not show skewed predictions to the more numerous category in the datafile: as the category frequencies are not very skewed in the database (74 patients survive more than 6 months and 33 do not), this aids the classifiers to similarly reduce both type of errors when FSS processes are applied. This balanced situation in both error type reductions also increases the confidence

Table 6
Misclassification matrices of probabilities for the best run of FSS–TREE and the no-FSS approach for C4.5[a]

True class	Predicted class	
	Do not survive	Survive
Do not survive	0.10; 0.23	0.21; 0.08
Survive	0.18; 0.04	0.51; 0.65

[a] In each cell, the first value is for the no-FSS approach and the second for FSS–TREE. Results reflect the probabilities of the leave-one-out process.

of physicians in the results. As three randomized FSS methods obtain similar results for four classifiers, Table 6 shows the misclassification matrices of probabilities of the best run of FSS–TREE and the no-FSS approach for C4.5.

5. Summary and future work

A medical problem, the prediction of the survival of cirrhotic patients treated with TIPS, has been focused from a machine learning perspective, with the aim of obtaining a classification rule for the indication or contraindication of TIPS in cirrhotic patients. With the application of several feature selection techniques the predictive accuracy of applied classifiers is largely improved. Among feature selection techniques, FSS–TREE, a new randomized algorithm inspired on the new EDA paradigm, has obtained the best average accuracy results for each classifier. Coupled with this improvement, more compact models with fewer attributes, which could be easier understood and applied by our medical staff, have been obtained.

Although the new FSS EDA-inspired approach has been applied in this paper to the specific medical problem of TIPS indication, it has a general character and can be used for other kind of problems.

In the future, we plan to use a database with nearly 300 attributes to deal with the problem of survival in cirrhotic patients treated with TIPS, which also collects patients measurements 1 month after the placement of TIPS. For this work, we plan to apply other probability distribution factorization models that are different to PBIL and TREE in order to factorize the distribution of selected solutions in the EDA approach. We also plan to use more advanced probability estimation techniques, such as Bayesian networks [20], to study the relationships among the variables of the study database.

Acknowledgements

This work was supported by the PI 96/12 grant from Gobierno Vasco, Departamento de Educación, Universidades e Investigación and the grant UPV 140.226-EB131/99 from University of the Basque Country.

References

[1] Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Machine Learning 1991;6:37–66.

[2] Bäck T. Evolutionary algorithms is theory and practice. Oxford: Oxford University Press, 1996.

[3] Baluja S. Population-based incremental learning: a method for integrating genetic search based function optimization and competitive learning. Technical Report CMU-CS-94-163, Carnegie Mellon University, Pittsburgh, IL, 1994.

[4] Bornman PC, Krige JEJ, Terblanche J. Management of oesophageal varices. Lancet 1994;343:1079–84.

[5] Cestnik B. Estimating probabilities: a crucial task in Machine Learning. In: Proceedings of ECAI-90, 1990. p.147–9.

[6] Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Trans Inform Theory 1968;14:462–7.

[7] Clark P, Nibblet T. The CN2 induction algorithm. Machine Learning 1989;3:261–83.

[8] Conn HO. A peek at the Child–Turcotte classification. Hepatology 1981;1:1–7.

[9] Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, Spirtes P. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 1997;9:107–38.

[10] D'Amico G, Pagliaro L, Bosch J. The treatment of portal hypertension: a meta-analytic review. Hepatology 1995;22:332–54.

[11] Diettrich TG. Approximate statistical tests for comparing supervised learning algorithms. Neural Comput 1998;10:1895–924.

[12] Doak J. An Evaluation of feature selection methods and their application to computer security. Technical Report CSE-92-18, University of California at Davis, CA, 1992.

[13] Draper D, Fouskakis D. A case study of stochastic optimization in health policy: problem formulation and preliminary results. J Global Opt 2000;18:399–416.

[14] Etxeberria R, Larrañaga P. Global optimization with Bayesian networks. In: Proceedings of the II Symposium on Artificial Intelligence CIMAF '99. Special Session on Distributions and Evolutionary Optimization, 1999. p. 332–9.

[15] Friedman N, Yakhini Z. On the sample complexity of learning Bayesian networks. In: Proceedings of the Twelveth Conference on Uncertainty in Artificial Intelligence, 1996. p. 274–82.

[16] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley, 1989.

[17] Grefenstatte JJ. Optimization of ontrol parameters for genetic algorithms. IEEE Trans Syst Man Cybernetics 1986;1:122–8.

[18] Güvenir HA, Acar B, Demiröz G, Cekin A. A supervised Machine Learning algorithm for Arrhythmia analysis. Comput Cardiol 1997;24:433–6.

[19] Holland J. Adaptation in natural and artificial systems. Michigan: University of Michigan Press, 1975.

[20] Inza I, Larrañaga P, Sierra B, Etxeberria R, Lozano JA, Peña JM. Representing the behaviour of supervised classification learning algorithms by Bayesian networks. Pattern Recogn Lett 1999;20(11–13):1202–9.

[21] Jelonek J, Stefanowski J. Feature subset selection for classification of histological images. Artif Intell Med 1997;9:227–39.

[22] John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, 1994. p. 121–9.

[23] Kittler J. Feature set search algorithms. In: Chen CH, editor. Pattern recognition and signal processing. Alphen a/d Rijn: Sijthoff and Noordhoff, 1978. p. 41–60.

[24] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of IJCAI-95, 1995. p. 1137–43.

[25] Kohavi R, Sommerfield D, Dougherty J. Data mining using MLC++, a Machine Learning Library in C++. Int J Artif Intell Tools 1997;6:537–66.

[26] Kohavi R, John G. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[27] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. Pattern Recogn 2000;33:25–41.

[28] Langley P. Selection of relevant features in machine learning. In: Proceedings of the AAAI Fall Symposium on Relevance, 1994. p. 140–4.

[29] Larrañaga P, Etxeberria R, Lozano JA, Peña JM. Combinatorial optimization by learning and simulation of Bayesian networks. In: Proceedings of the Conference in Uncertainty in Artificial Intelligence, UAI-2000, 2000. p. 343–52.

[30] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Norwell, MA: Kluwer Academic Publishers, 1998.

[31] Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PCJ. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology 2000;31:864–71.

[32] Michie D. Personal models of rationality. J Statist Plann Inference 1990;25:381–99.

[33] Müehlenbein H, Paaß G. From recombination of genes to the estimation of distributions. Binary parameters. In: Lecture notes in computer science 1411: parallel problem solving from nature — PPSN IV, 1996. p. 178–87.

[34] Narendra P, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Trans Comput 1977;26:917–22.

[35] 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.

[36] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. Pattern Recogn Lett 1994;15:1119–925.

[37] Pugh RNH, Murray-Lion IM, Dawson JL, Pictioni MC, Williams R. Transection of the esophagus for bleeding oesophageal varices. Br J Surg 1973;60:646–9.

[38] Quinlan JR. Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[39] Róssle M, Richter GM, Nóldge G. Performance of an intrahepatic portocaval shunt (PCS) using a catheter technique — a case report, Hepatology 8:1988;1348 (abstract).

[40] Róssle M, Richter GM, Nóldge G, Palmaz JC, Wenz W, Gerok W. New operative treatment for variceal haemorrhage. Lancet 1989;2:153.

[41] Saunders JB, Walters JRF, Davies P, Paton A. A 20-year prospective study of cirrhosis. Br J Med 1981;282:263–6.

[42] Siedelecky W, Skalansky J. On automatic feature selection. Int J Pattern Recogn Artif Intell 1988;2:197–220.

[43] Vafaie H, De Jong K. Robust feature selection algorithms. In: Proceedings of the Fifth International Conference on Tools with Artificial Intelligence, 1993. p. 356–63.