Prototype Selection and Feature Subset Selection by Estimation of Distribution Algorithms. A case study in the survival of cirrhotic patients treated with TIPS

B. Sierra[1], E. Lazkano[1], I. Inza[1], M. Merino[2], P. Larrañaga[1], J. Quiroga[3]

[1] Dept. of Computer Science and Artificial Intelligence, University of the Basque Country, P.O. Box 649, E-20080 Donostia, Spain
e-mail: ccpsiarb@si.ehu.es
Web-site: http://www.sc.ehu.es/isg
[2] Basque Health Service - Osakidetza, Comarca Gipuzkoa - Este, Avenida Navarra 14, E-20013 Donostia - San Sebastián, Spain
[3] Facultad de Medicina, University Clinic of Navarra, E-31080 Pamplona - Iruña, Spain

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 six months after TIPS placement. An investigation for predicting the conduct of cirrhotic patients treated with TIPS is carried out using a clinical database with 107 cases and 77 attributes. We have applied a new Estimation of Distribution Algorithms based approach in order to perform a Prototype and Feature Subset Selection to improve the classification accuracy obtained using all the variables and all the cases. Used paradigms are K-Nearest Neighbours, Artificial Neural Networks and Classification Trees.

Keywords: Machine Learning, Prototype Selection, Feature Subset Selection, Transjugular Intrahepatic Portosystemic Shunt, Estimation of Distribution Algorithm, Indications.
[1]

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 porto-systemic 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) [2].

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 [17] 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 six 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. The only published study [12] to identify a subgroup of patients who die within a period after a TIPS placement fixes the length of this period to three months. However, we think that our specific conditions really suggest lengthening this period to six months.

In this paper a new technique is presented that combines Prototype selection and Feature Subset Selection by an Estimation of Distribution Algorithm (EDA) approach. Although the application of the new EDA-inspired technique 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 rest of the paper is organized as follows: the study database is described in section 2. Prototype Selection is introduced in section 3, while Feature Subset Selection (FSS) methods are described in section 4. Section 5 presents the Estimation of Distribution Algorithm, a new evolutionary computation paradigm. In section 6 the new approach used in this work is described, and experimental results are presented in section 7. Finally, the last section briefly summarizes the work and presents lines of future research in the field.

2 Patients. Study database

The analysis includes 107 patients. 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 six-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 [12].

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.

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		

3 Prototype selection

Usually, three main approaches are used to develop Prototype Selection Algorithms:

1. Filtration of cases.
   These approaches are introduced in the firsts research works about prototype selection, and use some kind of rule in order to incrementally determine which cases of the training database will be selected as prototypes and which

of them discarded as part of the model. Some of the works keep wrong classified cases (Hart[4], Aha[1]), other approaches keep typical instances as prototypes (Zhang[22]) and some other algorithms select the correctly classified instances (Wilson[21]).

2. Stochastical search.
   Among the stocastical search methods, some of them make use of Genetic Algorithms[7] to select prototypes. Cameron-Jones[3] offers an heuristic approach consisting on minimizing the lenght of the bit string that represents a group of well and wrong classified instances. These kind of algorithms have also been applied by Skalak[19].
3. Case weighing.
   In these approaches a computation is done to weight the cases based on well classified or on more abstract concepts (Djouadi y Boucktache[6]).

3.1 Hart's Condensed Nearest Neighbor algorithm

This algorithm constitutes the first formal proposal to built a condensation method (Condensed Nearest Neighbor, CNN, Hart[4]). The method defines the so called consistency concerned to the training set. It is said that a group of prototypes S is consistent with respect to another group D, if upon using S as the training set, it is possible to correctly classify the cases in D. It is desirable to obtain a condensed group of prototypes that is small and consistent. The corresponding algorithm is shown in Figure 1.

The operation mode of this method is very simple: it maintains in the training database those prototypes wrong classified taking as model the cases belonging to the S subset at each step. This is done under the presumption that wrong classified cases belong to the decision border. The CNN algorithm has a lineal order computational behavior in practice, obtaining quite reduced prototype subsets. It can not be assumed that the prototype subset obtained by applying this method is the minimal consistent group.

4 Feature Subset Selection

In supervised Machine Learning[13], the goal of a supervised learning algorithm is to induce a classifier that allows us to classify new examples $E* = \{e_{n+1}, ..., e_{n+m}\}$ that are only characterized by their n descriptive features. To generate this classifier we have a set of m samples $E = \{e_1, ..., e_m\}$, characterized by n descriptive features $X = \{X_1, ..., X_n\}$ and the class label $C = \{w_1, ..., w_m\}$ to which they belong ($w_i = 0$ or $w_i = 1$ in the two class problem we are working with). Machine Learning can be seen as a 'data-driven' process where, putting little emphasis on prior hypotheses than is the case with classical statistics, a 'general rule' is induced for classifying new examples using a learning algorithm. Many representations with different biases have been used to develop this

Hart Condensation Algorithm

```
START
  As input, the algorithm should receive:
    The training database, D, containing N cases,
    (x_i, θ_i), i = 1, ..., N,

  Initialize the prototype set S to an empty set
  FOR each case belonging to D, (x_i, θ_i) DO
    START
        IF the class given by the NN algorithm
        taking S as the training database is correct
        THEN
            Do not add the case to S
        ELSE
            Add the case to S
    END
  As output, the algorithm gives the set S containing the selected prototypes.
END
```

Fig. 1. Pseudo-code of the *Hart Condensation Algorithm.*

'classification rule'. Here, the Machine Learning community has formulated the following question: *"Are all of these d descriptive features useful for learning the 'classification rule'?"* Trying to respond to this question the Feature Subset Selection (FSS) approach appears, which can be reformulated as follows: *given a set of candidate features, select the best subset under some learning algorithm.*

This dimensionality reduction made by a FSS process can carry out several advantages for a classification system in a specific task:

- a reduction in the cost of acquisition of the data,
- improvement of the compressibility of the final classification model,
- a faster induction of the final classification model,
- an improvement in classification accuracy.

The attainment of higher classification accuracies is the usual objective of Machine Learning processes. It has been long proved that the classification accuracy of Machine Learning algorithms is not monotonic with respect to the addition of features. Irrelevant or redundant features, depending on the specific characteristics of the learning algorithm, may degrade the predictive accuracy of the classification model. In our work, FSS objective will be the maximization of the performance of the classification algorithm. In addition, with the reduction in the number of features, it is more likely that the final classifier is less complex and more understandable by humans.

Once the objective is fixed, FSS can be viewed as a search problem, 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 infeasible in practice because of the large amount of computational effort required. Many search techniques have been proposed to solve FSS problem when there is no knowledge about the nature of the task, carrying out an intelligent search in the space of possible solutions.

5 Estimation of Distribution Algorithms

Genetic Algorithms (GAs, see Holland [7]) are one of the best known techniques for solving optimization problems. Their use has reported promising results in many areas but there are still some problems where GAs fail. These problems, known as deceptive problems, have attracted the attention of many researchers and as a consequence there has been growing interest in adapting the GAs in order to overcome their weaknesses.

GAs are a population based method. A set of individuals (or candidate solutions) is generated, promising individuals are selected, and new individuals are generated using crossover and mutation operators.

An interesting adaptation of this is the Estimation of Distribution Algorithm (EDA) [14] (see Figure 3). In EDAs, there are neither crossover nor mutation operators, the new population is sampled from a probability distribution which is estimated from the selected individuals.

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. 2. Main scheme of the EDA approach.

In this way, a randomized, evolutionary, population-based search can be performed using probabilistic information to guide the search. In this way, both approaches (GAs and EDAS) do the same except that EDAs replaces genetic crossover and mutation operators by means of the following two steps:

1. a probabilistic model of selected promising solutions is induced,
2. new solutions are generated according to the induced model.

The main problem of EDA resides on how the probability distribution $p_l(\mathbf{x})$ is estimated. Obviously, the computation of 2^n probabilities (for a domain with n binary variables) is impractical. This has led to several approximations where the probability distribution is assumed to factorize according to a probability model (see Larrañaga et al. [11] or Pelikan et al. [15] for a review).

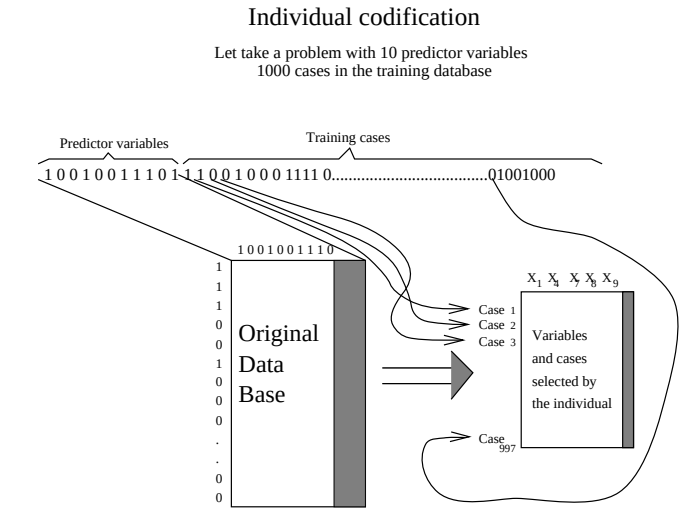

Fig. 3. Example of an individual codification.

6 New approach

In the new approach used in this paper, a binary individual codification is used for the Prototype Seleftion and FSS[2]. Let n be the number of predictor variables in the initial training database, and let m be number of cases. We use a binary individual of $(n+m)$ bits length, where the first n bits indicate the selected predictor variables and the last m the selected cases. In both cases, a bit value of 1 indicates that selection is done and a value of 0 means rejection.

In other words, if the i-th bit $= 1$

$$\begin{cases} i \leq n, \text{the} X_i \text{predictor variable is selected} \\ i > n, \text{ the } Case_{i-n} \text{is selected} \end{cases}$$

[2] The variable corresponding to the class is not included

else the i-th bit value being 0 indicates no selection of the corresponding variable or case. Figure 3 shows a graphical example of the new approach.

In other words, one individual $(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})$ is represented by a binary string of lenght $n+m$, where, for $i = 1, \ldots, n$:

$$x_i = \begin{cases} 0 \text{iff the } i\text{-th predictor variable is selected} \\ 1 \text{in other case} \end{cases}$$

and for $i = n+1, \ldots, n+m$:

$$x_i = \begin{cases} 0 \text{iff the } i-n\text{-th case is selected} \\ 1 \text{in other case} \end{cases}$$

7 Experimental results

Three different classification algorithms have been used in the experimentation done, in order to show that the main Prototype Selection - Feature Subset Selection by Estimation of Distribution Algorithm (PS-FSS-EDA) approach could be used with most Machine Learning paradigms. The algorithms used are the 1-Nearest Neighbor (1-NN)[4] and a backpropagation (BP) based Artificial Neural Network[13] and C4.4 classification tree[16].

Validation is done by the Leave-One-Out (LOO) technique, which consists on classifying each case of the database by using the model obtained by the rest of the training cases. In other words, it is the Stone[20] X-Fold Cross-validation method in which the X equals the number of cases of the training database, 107 in our case. This validation technique is applied usually when the size of the database is minor than 1000.

Table 2. Well classified cases obtained using all the cases and all the variables, estimated by the Leave-One-Out validation technique

Classifier	Variables	Number of cases	Well classified	Percentage
1-NN	77	107	68	63.55
BP	77	107	73	68.22
C4.5	77	107	68	63.55

Table 2 shows the classification accuracy obtained by each of the classifiers used with the training database using all the 107 cases and all the 77 predictor variables, estimated by the LOO method.

The new proposed PS-FSS-EDA approach have been used with the same database. Obtained results estimated by the LOO technique are shown in the Table 3.

As it could be seen, accuracy results obtained by the three classifiers by applying PS-FSS-EDA paradigm outperform those obtained using all the cases and all the variables.

Table 3. Well classified cases obtained by the PS-FSS-EDA approach, estimated by the Leave-One-Out validation technique

Classifier	Prototype average	Variable average	Well classified	Percentage
1-NN	5.12	37.38	83	77.53
BP	67.53	53.56	80	74.77
C4.5	?	?	76	71.03

When the classification models are presented to the medical staff, they noted a large improvement in aplicability among the models induced with the aid of FSS techniques and those that are constructed without FSS. Thus, by this dimensionality reduction, the confidence and acceptance in the models of our medical staff is increased. With the reduction in the number of needed measurements, an obvious reduction of the derived economic costs is achieved.

8 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 method for the indication or contraindication of TIPS in cirrhotic patients.

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 one month after the placement of TIPS. We also plan to use more advanced probability estimation techniques, such as Bayesian networks [8], to study the relationships among the variables of the study database. We are also planning to use a classifier combination technique[18] in order to improve obtained results by using more than one classifier.

Acknowledgments

This work was supported by the PI 1999-40 grant from Gobierno Vasco - Departamento de Educación, Universidades e Investigación and the grant 9/UPV/EHU 00140.226-/2084/2000 from University of the Basque Country.

References

1. D. Aha, D. Kibler y M.K. Albert (1991): Instance-Based learning algorithms. *Machine Learning* **6**, 37-66.
2. P.C. Bornman, J.E.J. Krige and J. Terblanche, Management of oesophageal varices, *Lancet* 343 (1994) 1079-1084.
3. M. Cameron-Jones (1995): Instance selection by encoding length heuristic with random mutation hill climbing. *IEEE Proceedings of the eighth Australian Joint Conference on Artificial Intelligence*, World Scientific, 99-106.

4. B.V. Dasarathy (1991): *Nearest Neighbor (NN) Norms: NN Pattern Recognition Classification Techniques.* IEEE Computer Society Press
5. T.G. Diettrich, Approximate statistical tests for comparing supervised learning algorithms, *Neural Computation* 10 (1998) 1895-1924.
6. P. Djouadi, E. Boucktache (1997): A fast algorithm for the Nearest-Neighbor Classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 3, 277-281.
7. J. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, 1975).
8. I. Inza, P. Larrañaga, B. Sierra, R. Etxeberria, J.A. Lozano and J.M. Peña, Representing the behaviour of supervised classification learning algorithms by Bayesian networks, *Pattern Recognition Letters*, 20 (11-13) (1999) 1201-1210.
9. I. Inza, P. Larrañaga, R. Etxeberria and B. Sierra (2000): "Feature subset selection by Bayesian network-based optimization", *Artificial Intelligence* **123**, 157-184.
10. R. Kohavi and G. John, Wrappers for feature subset selection, *Artificial Intelligence* 97 (1997) 273-324.
11. P. Larrañaga, R. Etxeberria, J.A. Lozano, B. Sierra, I. Inza, J.M. Peña, A review of the cooperation between evolutionary computation and probabilistic graphical models, in: Proceedings of the II Symposium on Artificial Intelligence CIMAF99, La Habana, Cuba, 1999, pp. 314-324.
12. M. Malinchoc, P.S. Kamath, F.D. Gordon, C.J. Peine, J. Rank and P.C.J. ter Borg, A model to Predict Poor Survival in Patients Undergoing Transjugular Intrahepatic Portosystemic Shunts, *Hepatology* 31 (2000) 864-871.
13. T. Mitchell (1997): *Machine Learning.* McGraw-Hill.
14. H. Müehlenbein and G. Paaß, 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) 178-187.
15. M. Pelikan, D.E. Goldberg, F. Lobo, A Survey of Optimization by Building and Using Probabilistic Model, IlliGAL Report 99018, Urbana: University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, 1999.
16. J.R. Quinlan (1993): "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers, Inc.* Los Altos, California
17. M. Rössle, V. Siegerstetter, M. Huber and A. Ochs, The first decade of the transjugular intrahepatic portosystemic shunt (TIPS): state of the art, *Liver* 18 (1998) 73-89.
18. B. Sierra, N. Serrano, P. Larrañaga, E.J. Plasencia, I. Inza, J.J. Jiménez, J.M. De la Rosa and M.L. Mora (2001): Using Bayesian networks in the construction of a multi-classifier. A case study using Intensive Care Unit patient data. *Artificial Intelligence in Medicine*. In press.
19. D.B. Skalak (1994): Prototipe and feature selection by Sampling and Random Mutation Hill Climbing Algortithms. *Proceedings of the Eleventh International Conference on Machine Learning*, NJ. Morgan Kaufmann. 293-301.
20. M. Stone (1974): Cross-validation choice and assessment of statistical procedures. *Journal Royal of Statistical Society* **36**, 111-147.
21. D.L. Wilson (1972): Asymptotic properties of nearest neighbour rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 2, 408-421.
22. J. Zhang (1992): Selecting Typical instances in Instance-Based Learning. *Proceedings of the Ninth International Machine Learning Workshop*, Aberdeen, Escocia. Morgan-Kaufmann, San Mateo, Ca, 470-479.