Annals of the University of Craiova, Math. Comp. Sci. Ser.
Volume 35, 2008, Pages 15–20
ISSN: 1223-6934

Features selection approach for non-invasive evaluation of liver fibrosis

Smaranda Belciug, Monica Lupsor, Radu Badea

Abstract. In many domains, a range of input variables are considered, not clearly which of them are most useful, or indeed are needed at all. Data are often collected on variables that are not only correlated, but also are large in number. This makes the data process, interpretation and detection of its structure difficult. Feature selection is a pattern recognition approach to choose important variables according to some criteria, in order to improve the decision process by removing the redundant information. The intent of this work is to provide a feature selection approach, based on the analysis of correlations between the explanatory (input) variables and the outcome variables, to improve the classification process of the liver fibrosis stages, using both the naive Bayes classifier and the probabilistic neural network model.

2000 Mathematics Subject Classification. 62C10; 92B20.
Key words and phrases. Feature selection, naive Bayes classifier, probabilistic neural network model, hepatic fibrosis.

1. Introduction

Curse of dimensionality [2] refers to the exponential growth of hyper-volume as a function of dimensionality. The curse of dimensionality causes ML models processing lots of redundant inputs to behave relatively badly. Since the dimension of the input space is high, the stand-alone software uses almost all its resources to represent irrelevant features, thus reducing its processing speed. A priori information can help dealing with the curse of dimensionality together with the use of standard statistical tools. Hepatic fibrosis is the major indicator of progressive liver disease and, therefore, an accurate assessment of fibrosis is vital for monitoring disease progression. Until recently, liver biopsy was the only way to evaluate liver fibrosis and it has traditionally been considered as the gold standard [3]. An obvious trend in clinical practice observed in the latter years consists of finding a correct method for liver fibrosis evaluation in a non-invasive way, both by biochemical tests as well as imaging methods, as an alternative to liver biopsy. The last technological discovery in the evaluation of liver fibrosis is the Fibroscan (Echosens, Paris), a specially adapted ultrasound device using the principle of the one-dimension transient elastography (TE) for the assessment of liver stiffness [9]. The first affection which was studied by using this method has been the chronic hepatitis of C viral etiology. Technically, the practice of Fibroscan is based on establishing some cut-off values of the liver stiffness for each stage of fibrosis. The aim of this paper is multi-fold: (a) to reduce the dimensionality of the dataset, reducing the biological parameters to the significant ones, (b) to compare the results obtained using naive Bayesian classification both for the complete database [5] and for the reduced dataset obtained by features selection, (c) to compare the results obtained using Probabilistic Neural Networks (PNN) both for the

Received: 10 November 2008.

complete and the reduced databases, and (d) to analyze the performances of the two machine learning (ML) models using the reduced database.

2. Naive Bayesian Classifier

The naive Bayesian classifier is considered to be one of the most powerful of the statistical classification models. In the case of the N-classes $\Omega_1, \Omega_2, ...\Omega_q$ classification problem, using training features vectors we can make an estimate of the probability of an unknown vector belonging to a particular class Ω_i. Each n-dimensional feature vector $x = (x_1, ..., x_n)$ represents a data point by depicting n-measurements made on the sample from n attributes. The unknown feature vector x is assigned to the class that has the highest conditional probability of the vector x belonging to it: $P(\Omega_i|x) > P(\Omega_j|x)$, $j = 1, 2, ..., q, j \neq i$ where $P(\Omega_i|x)$ is the posterior conditional probability that the vector's class membership is Ω_i given that the vector is x.

Bayes' law gives the equation of the probability of obtaining the vector x in each of the possible N classes. $P(\Omega_j|x) = \frac{P(x|\Omega j) \cdot P(\Omega_j)}{P(x)}$ where

- $P(x|\Omega j)$ is the class conditional probability that a vector is x, given that it belongs to the class Ω_j.

- $P(\Omega_j)$ is the probability that a vector belongs to class Ω_j regardless of the identity of the vector.

- $P(x)$ is the probability that a vector is x regardless of its class membership.

As $P(x)$ is constant for all classes, only $P(\Omega_j|x) = \frac{P(x|\Omega j) \cdot P(\Omega_j)}{P(x)}$ needs to be maximized.

In practical cases the probability that the vector x belongs to class Ω_j, $P(x|\Omega j)$ can be assumed to be Gaussian.

3. Probabilistic Neural Networks

Probabilistic Neural Networks (PNN) are basically classifiers. They determine the class membership of a n-dimensional feature vector $x = (x_1, ..., x_n)$ into one of q possible classes $\Omega_1, \Omega_2, ...\Omega_q$.

If we know

- the (multivariate) probability density functions (p.d.f.) $f_i(x)$ associated with the classes $\Omega_1, \Omega_2, ...\Omega_q$.

- the prior probabilities $P(\Omega_i)$ of occurrence of patterns of Ω_i.

- the loss (or cost) parameters l_i associated with all incorrect decisions given $\Omega = \Omega_i$ then according to the Bayes decision rule we classify x in the class Ω_i if:

$l_i \cdot P(\Omega_i) \cdot f_i(x) > l_j \cdot p(\Omega_j) \cdot f_j(x), i \neq j$

The key of this problem consists in the choose of the p.d.f.s $f_i(x)$.

We have considered the algorithm related to the Parzen - Cacoulos window classifiers [10],[11] using a sum of small multivariate Gaussian distributions, centered at each training sample, that is:

$$f_i(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot \sigma^n} \cdot \frac{1}{m_i} \cdot \sum_{i=1}^{m_i} \exp\left(-\frac{||x - x_j||}{2\sigma_i^2}\right), i = 1, ..., q$$

where:

- m_i is the total number of training patterns in Ω_i.

- x_j is the j-th training pattern from category Ω_i.

- n s the input space dimension.

- σ is an adjustable "smoothing " parameter using the training procedure.

We used a modified algorithm, using Genetic Algorithm procedure for searching the smoothing parameter [4]

4. Feature Selection

Feature selection plays an important role in classification problems. Basically, feature selection is the process of removing features from the dataset that are irrelevant or redundant with respect to the task that is to be performed. It can be extremely useful in reducing the dimensionality of the data to be processed by a certain classifier, reducing execution time or even improving predictive accuracy, since inclusion of irrelevant/redundant features may introduce noise into the data. Diverse feature selection techniques have been proposed in the machine learning literature, such as: Principal Component Analysis (PCA), correlation-based feature selection methods (with genetic search, greedy strategy etc.) [7], Support Vector Machine Feature Elimination [6] etc. In this study we selected a subset of attributes from the entire list of explanatory variables considering the linear relationship defined by the correlation between them and the outcome variable of interest (metavir F).

5. The Dataset

125 patients with chronic HCV infection examined at the 3rd Medical Clinic, University of Medicine and Pharmacy Cluj-Napoca, Romania, between May 2007 and August 2008 were prospectively included in this study. All of them had positive HCV-RNA in their serum and underwent percutaneous liver biopsy (LB) for grading and staging the diseases. All patients were referred to liver stiffness measurement (LSM) 1 day prior to LB. Besides the epidemiological, anthropometric and clinical parameters, the following biological parameters were determined for all patients on the same day as LSM: aspartate aminotransferase, alanine aminotransferase, gamma-glutamyl-transpeptidase, total bilirubin, alkaline phosphatases, platelet count (trombocytes), prothrombin time ratio, prothrombin index, sideremia, fasting blood glucose (glycemia), cholesterol and triglycerides. The study was approved by the local Ethical Committee of the University of Medicine and Pharmacy Cluj-Napoca. The nature of the study was explained to the patients, each of whom provided written informed consent before the beginning of the study, in accordance with the principles of the Declaration of Helsinki (revision of Edinburgh, 2000). Transient elastography (TE) was performed using FibroScan device (Echosens, Paris, France), which consists of a 5-MHz ultrasound transducer probe mounted on the axis of a vibrator. The vibrator generates a completely painless vibration (frequency 50 Hz and amplitude 2 mm) which is similar to a "flick", generating an elastic share wave that propagates through the skin and the subcutaneous tissue to the liver. Histological study consists in a liver biopsy examination, performed by using the TruCut technique with a 1.8 mm (14G) diameter automatic needle device - Biopty Gun (Bard GMBH, Karlsruhe, Germany). Summarizing, the medical dataset used in this study consists of 125 records, each of one containing 27 attributes (26 main biological parameters plus the Fibroscan data (i.e. the stiffness)) and a decision attribute (categorical data), consisting in five labels (F0 - F4), measuring the stages of the liver fibrosis. The aim is to classify a patient into one of the five categories depending on the explanatory attributes (the 26 main biological parameters combinrd with the stiffness-the Fibroscan output).

6. Results

The first step in our statistical multivariate analysis is to examine the relation between each potential explanatory variable (biological parameters and stiffness) and the outcome variable (metavir -Fibroscan output), that is carrying out linear regression analysis on each variable. Table 1 summarizes these analyses for the whole database. Thirteen of the twenty seven variables are significantly associated with metavir-F ($P < 0.05$).

Table 1. Results of separately regressing Fibroscan output (metavir-F) on each explanatory variable

Explanatory variable	**regression coefficient**	**p- level**
stiffness	***0.6774***	***0***
Age	***0.4515***	***0***
total bilirubin	***0.3548***	***0***
prothrombin index	***-0.5014***	***0***
INR (prothrombin time ratio)	***0.5287***	***0***
phosphatase (alkaline phosphatase)	***0.2772***	***0.002***
asat (aspartate aminotransferase)	***0.2819***	***0.001***
gamagt (gama glutamyl transferase)	***0.234***	***0.008***
glycemia	***0.2278***	***0.01***
APTT (prolonged activated partial thromboplastin time)	***0.2226***	***0.012***
hematia (red cells)	***-0.1952***	***0.029***
cholesterol	***-0.1922***	***0.031***
Sex	***-0.1794***	***0.044***
sideraemia	0.1586	0.076
chem (average concetration of hemoglobinin a red blood cell)	0.1586	0.076
hematocrit	-0.1412	0.115
hem (average eritrocitary hemoglobin)	0.1207	0.178
thrombocytes	0.1207	0.178
hemoglobin	-0.1082	0.228
BMI (body mass index)	0.0961	0.284
vem (medium eritrocity volume)	0.0872	0.331
leucocite	0.0872	0.331
tqs (tocopheryl quinones)	0.0724	0.42
triglycerides	-0.0452	0.615
alat (alanin aminotransferase)	0.0368	0.682
uree (urea)	0.043	0.633
creatinim	0.0154	0.864

We have performed a backward stepwise regression-like technique [1], based on the assumption that we have collected all these features because we believe them to be potentially important explanatory factors. Accordingly, we start by fitting the full model including all these potential explanatory features, and then remove unimportant variables, until all those remaining contribute significantly. Using the well-known statistical criterion based on the p-level of significance, we removed the variables with the smallest contribution to the model (that is with $P \geq 0.05$). In this way we obtained a reduced database, containing thirteen explanatory most significant features (stiffness, age, total bilirubin, prothrombin index, INR, glycemia, APTT, hematia, cholesterol and sex). Next, we applied the two machine learning models both to the whole database and the reduced database, in order to highlight possible significant differences in classification performance. Table 2 summarizes the relation between the number of features considered in the classification process and the corresponding diagnosis accuracies, illustrating the dimensionality reduction performance.

Table 2. Features selection performance

Model	Correlation Metavir-F vs. explanatory variables (p-level)	No. attributes	Accuracy (%)	Difference between accuracies (p-level)
Bayes	$p \geq 0.05$	27	70.08	
	$p < 0.05$	13	69.18	27 vs. 13 ($p = 0.87$)
	$p < 0.03$	12	46.08	27 vs. 12 **(p = 0.005)**
	$p < 0.02$	8	46.16	27 vs. 8 **(p = 0.003)**
PNN	$p \geq 0.05$	27	42.81	
	$p < 0.05$	13	40.95	27 vs. 13 ($p = 0.76$)
	$p < 0.03$	12	35.05	27 vs. 12 ($p = 0.24$)
	$p < 0.02$	8	35.32	27 vs. 8 ($p = 0.24$)

From the above table we draw the following conclusions: - The naive Bayes classification works much better than PNN (70% accuracy against only 43% for the entire database)

- For the naive Bayes classifier the optimum number of features is thirteen (drop in performance equaling 1% only). Beyond this significant data reduction (less than thirteen explanatory variables kept in model) the performance decrease dramatically (from 70% to 46%)

- For PNN classifier, even worse than naive Bayes), the same number of relevant features (thirteen) must be retained;

- Contrary to PNN, the naive Bayes is more sensitive to the change in the database dimensionality. Thus, accuracy decreases from 70% (initial database) to 46% (reduced database 8 attributes) in the case of naive Bayes, while for PNN the drop in accuracy ranges from 43% to 32%.

7. Conclusion and Future Work

The concept of dimensionality reduction by features selection is central both in classification and regression problems. In this paper the effectiveness of a correlation-based feature selection has been investigated. We applied this technique both to the naive Bayes classifier and probabilistic neural network, in order to highlight its efficaciousness in classification problems. Even the naive Bayes classifier performs better than the probabilistic neural network on this medical database, overall, this feature selection method proved useful for both models. Thus, in the case of our database, a dimensionality reduction by features selection is leading to an optimization in the management of liver diseases.

References

[1] Altman, D. G., 1991,Practical statistics for medical research, Chapman and Hall.

[2] Bellman, R. 1961, Adaptive Control Processes: A Guided Tour, Princeton, University Press.

[3] Bravo, A.A., Sheth, S.G., Chopra, S., 2001 Liver biopsy. N Engl J Med ; 344: 495-500.

[4] Gorunescu, F., Gorunescu, M., El-Darzi, E., Ene, M., Gorunescu, S., 2005, Statistical Comparison of a Probabilistic Neural Network Approach in Hepatic Cancer Diagnosis, Proceedings IEEE International Conference on ”Computer as a tool”- Eurocon 2005, Belgrade, Serbia, 237-240.

[5] Gorunescu, M., Belciug, S., Abdel Badeeh, S., Lupsor, M., Badea, R., Stefanescu, H., 2009, A machine learning-based diagnosis for liver diseases using the Fibroscan medical ultrasound technique, Proceedings 4th ACM International Conference on Intelligent Computing and Information Systems ICICIS09, Cairo (accepted).

[6] Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002, Gene selection for cancer classification using support vector machines, Machine Learning, Vol. 46, 389-422.

[7] Hall, M. A., Smith, L. A., 1998, Practical feature subset selection for machine learning, In Proceedings of the 21st Australian Computer Science Conference, 181-191.

[8] Lupsor, M., Badea, R., Stefanescu, H., Grigorescu, M., Sparchez, Z., Serban, A., Branda, H., Iancu, S., Maniu, A., 2008 Analysis of histopathological changes that influence liver stiffness in chronic hepatitis C. Results from a cohort of 324 patients. J Gastrointestin Liver Dis. ; 17(2): 155-163.

[9] Sandrin, L., Fourquet, B., Hasquenoph, J.M., et al., 2003, Transient elastography: a new noninvasive method for assessment of hepatic fibrosis. Ultrasound Med Biol ; 29: 1705-1713.

[10] Specht, D.F., 1988, Probabilistic neural networks for classification mapping or associative memory, Proc. IEEE International Conference on Neural Networks, vol1, 525-532.

[11] Specht, D.F., 1990, Probabilistic neural networks, Neural Networks, vol 3, 109-118

[12] Zaknich, A., 2003, Neural Networks for Intelligent Signal Processing, World Scientific.

(Smaranda Belciug) Department of Informatics, University of Craiova,
Al.I. Cuza Street, No. 13, Craiova RO-200585, Romania, Tel. & Fax: 40-251412673
E-mail address: `smaranda.belciug@inf.ucv.ro`

(Monica Lupsor) Department of Ultrasonography 3rd Medical Clinic University of
Medicine and Pharmacy Cluj - Napoca, Romania
E-mail address: `mmlupsor@yahoo.com`

(Radu Badea) Department of Ultrasonography 3rd Medical Clinic University of
Medicine and Pharmacy Cluj - Napoca, Romania
E-mail address: `rbadea@umfcluj.ro`