Operations and Quantitative Management

Volume 12 Number 1 March 2006

Process: An Application to Medical Diagnosis

Volume 12, *Number* 1
March 2006, *pp.* 73-83
Received: March 2005
Accepted: November 2005

Maria Teresinha Arns Steiner
UFPR – Departamento de Matemática, CP: 19081-CEP: 81531-990, Curitiba, PR, Brazil
(tere@mat.ufpr.br)

Nei Yoshihiro Soma
ITA – Divisão da Ciência da Computação, Pça. Mal. Eduardo Gomes, 50, Vl. das Acácias CEP: 12228-990, São José dos Campos, SP, Brazil
(nysoma@ita.br)

Tamio Shimizu
USP – Departamento de Engenharia de Produção, São Paulo, SP, Brazil
(tmshimiz@usp.br)

Júlio Cesar Nievola
PUC-PR – Programa de Pós-Graduação em Informática, Av. Imaculada Conceição, 1155, CEP 80215-901, Curitiba, PR, Brazil
(nievola@ppgia.pucpr.br)

Pedro José Steiner Neto
UFPR – Departamento de Administração, CP: 19081-CEP: 81531-990, Curitiba, PR, Brazil
(pedrosteiner@ufpr.br)

Knowledge Discovery in Databases – KDD – is a process that is made up of several steps starting with data collection for the problem under analysis and ending with the interpretation and evaluation of the final results. This work aims at showing the influence of exploratory data analysis over the performance of Data Mining techniques with respect to the classification of new patterns by means of its application in a medical problem. Through the analysis of the final results, one can conclude that if exploratory data analysis is properly carried out it may bring important improvements to the techniques' performance and become an important tool for final result optimization.

Keywords: Data Mining, KDD Process, Exploratory Data Analysis, Linear Programming, Neural Networks

1. Introduction

The purpose of an area called Knowledge Discovery in Databases (KDD) is to approach techniques and tools that try to transform stored data, whether from plants, banks, hospitals, or telecommunication, ecological, real estate data, into usable knowledge.

The KDD process is a set of continuous activities that share the knowledge discovered from databases. According to Fayyad et al. [7], this process is made up of

five steps: data selection; data pre-processing and cleaning; data transformation; Data Mining and results interpretation and evaluation. The three first ones make up to what is called "exploratory data analysis".

The KDD process refers to the whole process of discovering useful knowledge in data, whilst Data Mining refers to applying algorithms to retrieve models from data. According to Freitas [9], the knowledge to be discovered must be correct, understandable and useful. Also, the method of finding knowledge must be efficient, generic and flexible.

This work aims at showing, by means of a medical problem shown in Section 3, the importance of exploratory data analysis with respect to the classification within the KDD context, in Sections 4 and 5. In Section 2, we accomplish a revision of the available literature and in Section 6 we present our conclusions.

2. Related Work

With the ample use of advanced technologies for databases developed during the last decades it has not been difficult to store large amounts of data in computers, and to retrieve them whenever necessary. Although stored data are a valuable asset in any organization, many face the problem of being "data rich but knowledge poor" [13].

This large amount of data largely surpasses our ability to interpret them, thus creating the need for techniques that try to transform stored data into knowledge, which is the purpose of an area called KDD [7].

In general, Data Mining techniques accomplish data classification or data clustering or, also, discovering data association rules. Among the Data Mining methods capable of recognizing patterns (classification), we can mention decision trees, Support Vector Machines (SVM), statistical methods, Neural Networks (NN), Genetic Algorithms (GA) and the metaheuristics, in general. On the other hand, the three first KDD steps that refer to the exploratory data analysis and that make use of statistical tools have not received the same attention from researchers; this is exactly the aspect we intend to explore in this work: the influence of exploratory data analysis prior to the use of Data Mining techniques, through a medical diagnosis problem.

Among the numerous works that approach Data Mining techniques for classification, one can mention the following, which concentrate mainly on NN and GA. Lu et al. [13], and Lu et al. [14], present in their articles the algorithm *Neurorule*, which makes the rule extraction from a trained NN, obtaining rules of the IF-THEN type. Fidelis et al. [8] present a classification algorithm based on GA, which discovers IF-THEN-type comprehensible rules within the Data Mining context; it was evaluated in two public domain medical databases – dermatology and breast cancer – obtained at the *UCI (University of California at Irvine) - Machine Learning Repository.*

Santos et al. [19] point out that because NN are robust and have a good tolerance to noise, they are suitable for mining very noisy data. Their method uses a GA to define a good NN topology to be trained and the proposed system was evaluated on three public-domain data sets available at the *UCI's* Repository: *Iris*, *Wine* and *Monks*-2. In their work, Baesens et al. [2], indicate the high predictive accuracy rate

of NN and the authors show three methods for the NN rule extraction, comparatively: *Neurorule*; *Trepan* and *Nefclass*. Three real-life credit-risk data sets were used to compare the approached methods' performances: *German Credit* (*UCI's* Repository), *Bene 1* and *Bene2* (obtained from the two largest financial institutions in Benelux).

Olden & Jackson [17] describe some methods in the literature to "illuminate" the mechanisms of an artificial NN (NN Interpretation Diagram, Garson's Algorithm and Sensitivity Analysis). Also, they propose an additional method - *randomization approach* - to statistically understand the importance of connections (weights) in a NN, as well as the contribution of input variables. In real estate evaluation, one can mention the work of Nguyen & Cripps [16], which compare the predictive performance of NN with the Multiple Regression Analysis for family housing sales. In the work of Bond et al. [3], the authors examine the effect a view to a lake has over the price of a house.

3. Description of the Medical Problem

Jaundice or icterus (from the Greek ikteros = yellow), which is only a symptom that presents a yellowish color of the skin, can be caused by a huge number of illnesses. The physicist must initially separate these illnesses into two big groups: a) Cholestasis (chole = bile; stasis = halt): this is the case in which there is a partial or complete blockage of the liver's bile's components towards the intestines; b) other causes.

Only the first group – cholestasis – will be studied. To make this initial distinction, the physicist usually gets support from simple exams that define this syndrome with a fairly good surety. However, this is not enough and the separation into two more groups is needed: a_1) obstruction by cancer; a_2) obstruction by calculus.

This differential diagnosis is generally possible by means of the data already collected, associated with more elaborated exams. However, around 16% to 22% of patients cannot be classified and complementary exams of the main biliary duct area show errors ranging from 30% to 40%. Exams that are capable of establishing the real difference between cancer and calculus as the cause of obstruction, when used together with the previous ones, show a high precision, over 95%. However, they are generally invasive and present risk of serious or even lethal complications.

Considering the risks for the patients and the high costs involved in an accurate diagnosis, the use of a KDD process applied to this kind of problem is justified as a means to optimize the diagnosis process (minimizing risks and costs for the patients and, on the other hand, maximizing result effectiveness) [21].

For such, we need historical data from the patients that fit to the two previous cases. We used data from 118 patients from the *Hospital das Clínicas* (*HC*) from *Curitiba, PR, Brazil*, from which 35 had a confirmed cancer and 83 had calculus in the biliary duct. From each one of these patients we considered 14 attributes from clinical exams measurements, suggested by a specialist from this field. These 14 attributes were the following (with their corresponding abbreviations): 1. Age (ag); 2. Sex (sex); 3. Bilirubin Total (bt); 4. Direct Bilirubin (db); 5. Indirect Bilirubin (ib); 6. Alkaline Phosphatases (ap); 7. SGOT (sgot); 8. SGPT (sgpt); 9. Protrombine

activity time (pat); 10. Albumin (alb); 11. Amylase (amy); 12. Creatinine (cr); 13. Leukocytes (le); 14. Gv (gv); and the classes are: cancer (can) and calculus (cal).

In possession of these data (patterns), already classified by the physicist, as icteric with cancer or icteric with calculi, one can apply the exploratory data analysis and, after that, use the five Data Mining techniques approached here, training the techniques with the referred data, aiming at the classification of future patients as accurately as possible.

4. Techniques used in this Work

The five KDD process' steps listed in Section 1 in this work were approached here in three different phases: exploratory data analysis; Data Mining and result acquisition and analysis.

4.1Exploratory Data Analysis

The complexity of many problems demands from the researcher the collection of many observations, each one of them with many variables (attributes; inputs). The purpose of exploratory data analysis is to use statistical methods to capture information from these data and, due to the fact that data include simultaneous measurements of many variables, this methodology is called exploratory data analysis [10]. This ample area of study involves numerous statistical techniques and, for this specific work, this analysis was composed by the following techniques: Hotelling's T^2 test; attribute transformation; and discarding atypical data.

In these techniques, described below, we have that A and B are sets of data from the two sets to be distinguished (cholestatic with cancer, represented by "1", and cholestatic with calculi, represented by "0"), m (=35) is the amount of data in set A, k (=83) is the amount of data in set B and n (=14) is the number of attributes.

Initially, **Hotelling's T^2** test [10] was applied in order to verify if the average vectors of the two multivariate populations A and B were equal. In this test, where F is Snedecor's distribution, if:

$$\frac{T^2(m+k-n-1)}{(m+k-2)n} >> F_{n,m+k-n-1}(0.95)$$

we strongly reject, with a 95% probability, the hypothesis that populations A and B are centered in the same average vector.

In the search for a Multiple Linear Logistic Model, mentioned in section 4.2, to fit the dichotomic response variables ("1" and "0") with the attributes and/or co-attributes, the adequacy of the model is measured based on the deviation function (deviance), introduced by Nelder & Wedderburn [15]. If the deviation function's value was statistically significant (or not), the attribute/co-attribute was incorporated into the model (or not). In this work, some of the 14 original attributes were **transformed** in scale, deriving some co-attributes, as presented in Section 5, trying to better capture its information.

Deviation is a measure of the distance between the adjusted values and the observed ones, or equivalently, between the current model and the saturated one; in general, what is searched for are models with moderate deviations. The deviation function is defined by [1]:

$$s_p = -2\{L_p - L_{(m+k)}\}$$

where L_p is the maximum likelihood logarithm function for the model under investigation, with p parameters and $L_{(m+k)}$ is the maximum likelihood logarithm function for the saturated model with $(m+k)$ parameters. The freedom degrees associated to the deviation are defined by: $\nu = (m+k)-p$. The likelihood rate test can be used for decisions about the most adequate model, where the statistics test is given by:

$$s_p = -2\{L_p - L_{(p+1)}\} \sim \chi^2_\nu$$

and based on p-value corresponding to s_p,

$$p = P(\chi^2_\nu > s_p \mid \beta_{p+1} = 0)$$

the attribute or the co-attribute of the model is kept or removed. In this formula, the p-value corresponding to the statistics is the probability of the theoretical variable be greater than to the referred statistics [1].

When obtaining the Multiple Linear Logistic Model with the smallest possible deviance it is important to analyze the data residues in relation to the model in order to identify atypical points and, also, their causes. The procedure for **discarding atypical points** was made based on the Pearson's residue calculation, which is made for each one of the data through the following calculation [10], [5]: where Y_i is the value taken over by the datum i in the saturated model and θ_i is this value's estimate made by the model. A value $|e_i| > 1$ shows that the model is misclassifying datum i,

$$e_i = \frac{Y_i - \theta_i}{\sqrt{\theta_i(1-\theta_i)}}$$

i.e., observation i is "displaced" in relation to its set (A or B) and this characterizes it as atypical. In these cases, according to the justification for this occurrence, datum i can be discarded from the sample and, therefore, from the model. It must be noticed that in this case, estimates for the model must be recalculated.

4.2 Data Mining Techniques

Since Fisher's work [10] in 1936, numerous works have been developed with the purpose of presenting discriminating analysis techniques for the classification task. These techniques can be fit in Data Mining, in the KDD context. In this work, the purpose sticks to the classification task through the separation of patterns of clients with cancer or with calculi in the biliary ducts.

For this, we approached five Data Mining techniques that are capable of classifying (separating) the data belonging to sets A and B:

- one technique that makes use of Linear Programming was proposed by Bennett & Mangasarian [4]. They proposed the formulation of a single linear program which generates a plane that minimizes an average sum of misclassified data (points) belonging to sets A and B. It must be noticed that this is not an iterative method, i.e., the separating surface obtained through this model is unique for both sets of data, A and B.

- two statistical techniques: the 1st one is Fisher's Linear Discriminant Function (Fisher's LDF), where given two samples A and B, multivariate observations $X \in R^n$, Fisher's idea was to transform these multivariate observations into univariate observations $Y's$ in such way that they are as much separated as possible [10]. The 2nd one is Multiple Linear Logistic Model (MLLM) which consists in relating through a model the answer variable (patterns belonging to set A or set B) to the attributes that influence its occurrence [11]. The quality of adjustment is measured by the deviation function previously defined in Section 4.1.
- one uses NN: the Multiple Layer NN (MLNN), also called feed-forward networks, used in this work is the method mostly used and largely divulged by the scientific community and shows, in general, quite satisfactory results. The back-propagation algorithm used for its training is a supervised algorithm [12, 6, 20]. For the data presented and used in this work the NN needed approximately 1,000 iterations to converge in each one of the tests.
- the last one uses Decision Trees: Quinlan [18] developed the technique that allowed the use of knowledge representation by means of Decision Trees. His contribution consisted in the elaboration of an algorithm called *ID3*, which, together with its evolutions - ID4, ID6, C4.5, See 5, is a tool adapted for use in decision trees. The main advantages of Decision Trees are that they induce a choice in the decision process by considering the most relevant attributes and people understand them better [22]. There are other advantages, among which we can point out the following: they do not assume any particular distribution for the data; characteristics or attributes can be categorical or numerical; one can build models for any function, provided the number of training examples is large enough; and they have a high comprehension level.

5. Implementation of the Techniques to the Medical Problem

As described below, the computational tests were applied to data matrices: a first matrix, M_1, containing the original data (A and B sets), in which only Hotelling's T^2 test was applied, and a second one, M_2, in which data from A and B sets, besides being analyzed by Hotelling's T^2 test, had their attributes transformed and their atypical points discarded; this way, M_2 has "adjusted" or statistically "transformed" data.

In order to apply Hotelling's T^2 test a Visual Basic computational program was developed. As a result of applying this program we verified that the populations – patients with cancer and patients with calculi – are distinct at a 95% level of confidence. More specifically, Hotelling's T^2 test with $T^2(m+k-n-1)/(m+k-2)n$ statistics, when compared to $F_{n,m+k-n-1}$ (0.95) for each one of the data matrices M_1 (original data) and M_2 (adjusted data), supplied the following values:

- M_1: $4.84 > 1.78896 = F_{14,103}$ (0.95);
- M_2: $12.54 > 1.82239 = F_{13,97}$(0.95).

Consequently, the equality of the two average vectors is rejected, because the probability that they are equal is way under 5% in both cases, especially in M_2. Thus,

in the studied attributes and co-attributes, the population of cancerous icteric patients is distinct from that one of icteric patients with calculi.

The transformation of attributes and discarding atypical data in order to get matrix M_2, were done simultaneously with the use of the *GLIM* (*Generalized Linear Interactive Modeling*) statistical software [1] through analysis of deviations (s_i values) for each one of the models obtained and of analysis of residues ($|e_i|$ values) for each one of the data, respectively. In this procedure seven data that were considered atypical (presented $|e_i| > 1.5$) were discarded from the 118, i.e., 6% of the total. This hypothesis was assumed after discussion with the specialists in the field and causes were established.

We can say that these procedures (transformation of attributes and discarding atypical data) were almost "handcrafted", because obtaining each one of the Multiple Linear Logistic Models containing the attributes and their "variations" (obtained by adding, removing, dividing, combining and others, the original attributes, generating the co-attributes), was carried out manually, searching for some of the numerous possibilities, one at each time, of "working" with the 14 original attributes. At each "variation", we checked the model's deviation value, in order to minimize it, but not too much.

By executing this "handcraft" work, we obtained a Multiple Linear Logistic Model with 13 attributes. During the transformation of attributes, matrix M_2 was defined with the six original attributes and seven co-attributes (transformed from the original attributes). The six original attributes were the following (with their abbreviations): age (ag); bilirubin total (bt); direct bilirubin (db); amylase (amy); alkaline phosphatases (ap); globular volume (gv). The seven co-attributes are: (direct bilirubin)^2, i.e., (db)^2; ln(amylase), i.e., ln(amy); the division of the attributes sgot by sgpt, i.e., (sgot/sgpt), or also, st; (sgot/sgpt)^2, i.e., st^2; (alkaline phosphatases)^2/1,000 i.e., ap2n; (globular volume)^2, i.e., (gv)^2; (bilirubin total)^2, i.e., (bt)^2. It must be observed that the six attributes were discarded to obtain matrix M_2 (abbreviations): sex (sex); indirect bilirubin (ib); protrombine activity time (pat); albumin (alb); creatinine (cr); leukocytes (le).

When we obtain each one of the models, especially the last one with 13 attributes (six original attributes and seven co-attributes), the *GLIM* software automatically obtains the values of Pearson's residues. Considering the procedure described in section 4.1, seven data were discarded. It must be pointed out that matrix M_1 has the original data with no changes.

Having matrices M_1 (matrix with the original data, no adjustments, of order 118 x 14) and M_2 (matrix with the adjusted data, of order 111 x 13), we applied the five Data Mining methods discussed in this work.

For the 1st method (Linear Programming Model), we used the *LINGO* (*Language for Interactive General Optimizer*) commercial software to solve the Linear Programming model. For the 2nd one (Fisher's FDL) and 4th (MLNN) methods, we developed a computational implementation in Visual Basic; for the 3rd method (MLLM), we also used the *GLIM* statistical software and, finally, for the 5th and last method (Decision Tree J4.8) we used the *WEKA* (*Waikato Environment for Knowledge Analysis*, available at *www.cs.waikato.ac.nz/ml/weka*) free software.

It should be pointed out that we used a 3-layer NN (an input layer with the same number of input neurons as were the number of information – 14 or 13, for matrices M_1 and M_2, respectively), a hidden layer with a number of neurons varying from 1 to 20, complying to the methodology presented by Steiner et al. [20], and the output layer with 1 neuron (patients with cancer, "1", or patients with calculi, "0"). As for the formation of the decision tree, we used a tree classification algorithm J4.8 (C4.5 release 8).

In all five methods, the testing methodology consisted in applying the stratified holdout method [22], repeated 10 times (10 simulations) for matrix M_1 and 10 times for matrix M_2 and as the standard deviation among them was relatively small for the two matrices, we chose to consider only the three "better" simulations (with smallest error percentages) for matrix M_1 as well as for matrix M_2, for the average calculations presented in Table 1.

In order to execute the 10 simulations, we randomly divided the data contained in matrices M_1 and M_2 into 2 sub-sets: one of them, called "Training Set" was used to train each one of the five methods, while the other sub-set "Testing Set" was used to test the trained models; this procedure was repeated 10 times for each matrix, varying the two sub-sets. One must notice that the "Training Set" was also used to test the methods.

We have to emphasize that the standard deviation values for the 10 simulations, considering the Training and Testing Sets, for all five methods, were in the intervals (1.27%; 3.47%) and (1.78%; 5.12%) for matrix M_1, respectively; for matrix M_2, these values were in the intervals (0.17%; 2.09%) and (0.67%; 3.58%).

The five Data Mining techniques were separately applied to matrices M_1 (118 x 14) and M_2 (111 x 13), so we could compare their performances over the original and the adjusted data, as presented in Table 1.

For the Linear Programming Model (the two first columns), for example, we have in this table 83.65% and 100% hits for matrices M_1 and M_2, respectively, considering the Training Set, and 75.01% and 96.97% hits for matrices M_1 and M_2, respectively, considering the Testing Sets. This same way, we have the interpretation for the other four methods. As one can observe in this table, the 1st (Linear Programming) and the 4th (MLNN) methods presented the biggest hit percentages for the Testing Sets (96.97%), showing that these two methods present a satisfactory generalization capacity.

Table 1

Methods' hit percentage average (three simulations) for all five methods for the medical problem, with matrix M_1 (118 x 14) containing the original data and with the adjusted matrix M_2 (111 x 13) containing the data statistically explored

Methods	Linear Progr.		Fisher's LDF		Log. Model		Neural Net.		Dec. Trees	
Data	Train. Set	Test. Set	Train. Set	Test. Set	Train. Set	Test. Set	Train. Set	Test. Set	Train. Set	Test. Set
M_1	83.65	75.01	81.13	80.56	83.33	80.56	77.00	72.33	95.91	77.78
M_2	100.00	96.97	93.34	90.91	98.50	95.45	85.67	96.97	97.67	75.75

Following these two methods are the 3rd (MLLM), the 2nd (Fisher's FDL) and the 5th (J4.8 Decision Tree) methods, with 95.45%, 90.91% and 75.75% hits, respectively, for the Testing Sets.

6. Conclusions

This work analyzed the importance of the exploratory data analysis prior to the use of Data Mining techniques by means of the results obtained with their use with a real medical problem (data of 118 patients, each one with 14 attributes). The five Data Mining techniques (Linear Programming, Fisher's FDL, Logistic Model, Neural Networks and Decision Trees) were applied over the original data (matrix M_1 (118x14)) and over the data statistically explored data (matrix M_2 (111x13)). The results are in Table 1.

By analyzing these results, one can notice that all methods, with the sole exception of the Set for Decision Tree Testing, showed a significant improvement in their performances with the adoption of exploratory statistical data analysis prior to applying Data Mining techniques, as we can verify comparing the results obtained through matrices M_1 and M_2. This fact emphasizes the importance of having a reliable and consistent data set and, therefore, exploratory data statistically analyzed. From the methods approached in this work, Linear Programming and Neural Networks presented the best results for the medical problem studied, i.e., they presented the greatest hit percentages for the Testing Sets, thus demonstrating their satisfactory generalization capacities: almost a 97% hit.

It is worthwhile to point out that from all the Data Mining techniques mentioned here, only the Decision Tree makes it out clear to the user which are the attributes that are distinguishing the patterns (comprehensibility) and how (cutting points) this is happening, as can be seen in Figure 1, which exemplifies one of the tests carried out in this work. In this tree we have direct bilirubin (db) as the most important attribute or tree "root". If its value is smaller than or equal to 6.2, this means that the patient has calculus (calc); otherwise, the amylase (amy) attribute, the 2nd most important one, is analyzed. If its value is greater than 164, this means that the patient has also calculus; otherwise, we analyze, once more, the amylase attribute: if it is greater than 92, this means that the patient has cancer (can); otherwise the alkaline phosphatases attribute (ap) is analyzed and the rest of the tree is interpreted the same way.

The relatively high percentage of errors in this technique, when compared to those of the others, is compensated by this highly desirable characteristic, i.e., keeping clear the "discriminator" attributes with their respective cut points, in a prioritized format. We must emphasize that from a Decision Tree, we can generate classification rules; for the Figure 1 case, we have the following rules:

IF Db $\leq$ *6.2* $\Rightarrow$ *Class = calc*
IF Db $>$ *6.2 and Amy* $>$ *164* $\Rightarrow$ *Class = calc*
IF Db $>$ *6.2 and 92* $<$ *Amy* $\leq$ *164* $\Rightarrow$ *Class = can*
IF Db $>$ *6.2 and Amy* $\leq$ *92 and Ap* $>$ *337.7* $\Rightarrow$ *Class = can*
IF Db $>$ *6.2 and Amy* $\leq$ *92 and Ap* $\leq$ *337.7 and Gv* $>$ *42.3* $\Rightarrow$ *Class = calc*
IF Db $>$ *6.2 and Amy* $\leq$ *92 and Ap* $\leq$ *337.7 and Gv* $\leq$ *42.3 and Ib* $\leq$ *5.5* $\Rightarrow$ *Class = calc*

IF Db > 6.2 and Amy ≤ 92 and Ap ≤ 337.7 and Gv ≤ 42.3 and Ib > 5.5 ⇒ Class = can

Among the techniques approached in this work, the technique that involves NN could also become understandable. It would only be necessary to use some kind of rule extraction algorithm from the trained NN, as presented by Lu et al. [14], Santos et al. [19] and others. This same way, we could also think of building algorithms to extract rules from the other three methods.

The methods presented in this work can be used in the most varied real classification problems, as they were in the medical case studied here. Thus, specialists from many different fields could evaluate the results obtained with the methods approached (and/or other additional methods used for the classification task) and validate (or not) the plausibility of the predictions they make, thus having an auxiliary tool for their decision making.

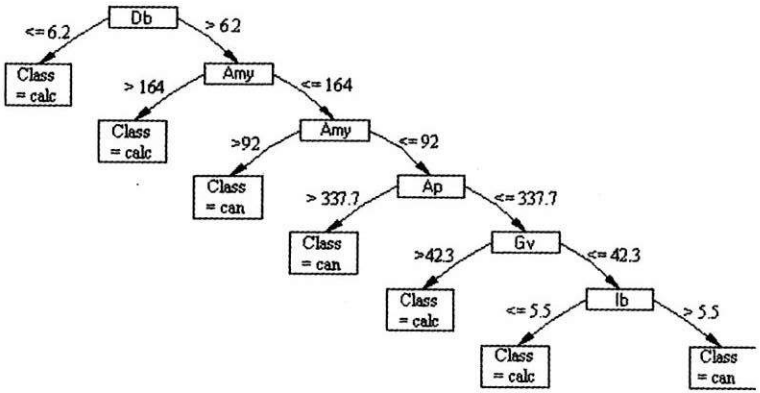

Figure 1 *An example of Decision Tree for the Medical Problem presented in this work, using the attributes' abbreviations, as presented in the text.*

7. References

1. Aitkin, M.; Anderson, D.; Francis, B.; Hinde, J., Statistical Modelling in GLIM, 1989, Oxford Statistical Science Series, Clarendon Press, New York.
2. Baesens, B.; Setiono, R.; Mues, C.; Vanthienen, J., "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation", *Management Science*, 49(3), 2003, 312-329.
3. Bond, M.T.; Seiler, V. L.; Seiler, M. J., "Residential Real Estate Prices: a Room with a View", *The Journal of Real Estate Research*, 23(1), 2002, 129-137.
4. Bennett, K.P.; Mangasarian, O.L., "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", *Optimization Methods and Software*, 1, 1992, 23-34.
5. Duda, R.O.; Hart, P.E.; Stork, D.G., Pattern Classification, 2001, John Wiley & Sons, inc., New York.

6. Fausett, L., Fundamentals of Neural Networks - Architectures, Algorithms, and Applications, 1995, Florida Institute of Technology, Prentice Hall, Upper Saddle River, New Jersey.
7. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., Advances in Knowledge Discovery & Data Mining, 1996, AAAI/MIT.
8. Fidelis, M.V.; LOPES, H.S.; FREITAS, A.A., "Um Algoritmo Genético para Descobrir Regras de Classificação em Data Mining", *Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação*, v.IV, 2000, 17-29.
9. Freitas, A.A., Uma Introdução a Data Mining, 2000, Informática Brasileira em Análise, CESAR - Centro de Estudos e Sistemas Avançados do Recife, Pernambuco.
10. Johnson, R.A.; Wichern, D.W., Applied Multivariate Statistical Analysis, 1998, Prentice-Hall, inc., New Jersey.
11. Hair JR, J.F.; Anderson, R.E.; Tatham, R.L.; Black, W.C., Análise Multivariada de Dados, 1998, Bookman, São Paulo.
12. Kröse, B.J.A.; Van Der Smagt, P.P., An Introduction to Neural Networks, 1993, University of Amsterdam, Amsterdam.
13. Lu, H.; Setiono, R.; Liu, H., "NeuroRule: A Connectionist Approach to Data Mining", *Proceedings of the 21st. VLDB Conference*, Switzerland, 1995, 478-489.
14. Lu, H.; Setiono, R.; Liu, H., "Effective Data Mining using Neural Networks", *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 1996, 957-961.
15. Nelder, J.A.; Wedderburn, R.W.M., "Generalized Linear Models", *J. R. Statistical Society*, A, no. 135, 1972, 370-384.
16. Nguyen, N.; Cripps, A., "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks", *The Journal of Real Estate Research*, 22(3), 2001, 313-336.
17. Olden, J.D.; Jackson, D.A., "Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks", *Ecological Modeling*, 154, 2002, 135-150.
18. Quinlan, J.C., C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo.
19. Santos, R.T.; Nievola, J.C.; Freitas, A.A., "Extracting Comprehensible Rules from Neural Networks via Genetic Algorithms", *IEEE*, 2000, 130-139.
20. Steiner, M.T.A.; Carnieri, C,; Kopittke, B.; Steiner Neto, P.J., "Probabilistic Expert Systems and Neural Networks in Bank Credit Analysis", *International Journal of Operations and Quantitative Management*, 6 (4), 2000, 235-249.
21. Steiner, M.T.A. ; Soma, N.Y.; Shimizu, T.; Nievola, J. C.; Steiner Neto, P.J., "The Influence of the Multivariate Data Analysis in the Knowledge Discovery in Databases", *IFORS 2005 (17 th. Triennial Conference of the International Federation of Operational Research Societies hosted by INFORMS)*, 2005, FB-20.
22. Witten, I. H.; Frank, E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2000, Morgan Kaufmann Publishers, San Francisco, California.