MDPI

Article

Data Mining for Primary Biliary Cirrhosis' Patients Survival Prediction

João Vilaça [1], Bárbara Cardoso [1], Cristiana Neto [2] and José Machado [2]

[1] Department of Information, University of Minho, Braga, Portugal
{a82339,a80453}@alunos.uminho.pt
[2] Algoritmi Research Center, Department of Information, University of Minho, Braga, Portugal
B8115@algoritmi.uminho.pt , jmac@di.uminho.pt

Version March 5, 2021 submitted to Journal Not Specified

Abstract: Primary biliary cirrhosis (PBC) is a rare, slow-progressing liver disease that can lead to fibrosis, liver failure, and culminate in liver transplantation or early mortality. Last year, in Portugal, 0.9 cases / 100,000 inhabitants were diagnosed and it is estimated that the prevalence of the disease is 15.6 cases / 100,000 inhabitants, of which 50% may be undiagnosed. For the DM process, the CRISP-DM methodology was followed and RapidMiner was the preferred tool used, being able to achieve some models with sensitivity bigger than 95% and accuracy bigger than 92%.

Keywords: Data mining; RapidMiner; Prediction; Health; Primary Biliary Cirrhosis

1. Introduction

The algorithms and methods that put together the Data Mining process, have existed for a long time. One of the essential basis in this area came up around 1973 when Thomas Bayes' work (where a theorem to relate current probability with prior probability was proposed, and later called Bayes' Theorem) was published. This theorem is fundamental to data mining and probability nowadays since it allows the understanding of complex realities based on estimated probabilities. Along with regression analysis, aiming at estimating the correlation between variables, first proposed by Adrien-Marie Legendre and Carl Friedrich Gauss in 1805, making up the statistic basis for most DM models.

Only from 1989 on, with the proposal of the first versions of Support Vector Machines - SVM (a supervised learning approach that analyses data and recognizes patterns for classification and regression analysis), the DM concept came up for the first time. From that point on, the theoretical evolution of algorithms and models progressed exponentially, but its application outside the academic world did not happen until much later. Only recently, with a substantial rise of the computational power we have been able to reach, the DM methods and techniques have been applied to organisations representing a big added value to them, since its capability to predict and anticipate, their processes patterns (negotiations or states) by analysing and processing large quantities of data, that only very recently companies became able to collect and store more efficiently and with a smaller associated cost.

One of the areas that invests the most in DM processes analysis, development, and application is definitely the healthcare area, that on top of having great research investment capability, it also has a lot of areas to put DM techniques to good use. These processes are used for many aspects since the prediction of the evolution of patients' states, in the interpretation and evaluation of medical analyzes, among others. In this particular context, it's important to adapt these DM processes to fit the particular case of patients suffering from Primary Biliary Cirrhosis to predict their survival probability, allowing the optimization of human and material resources and increasing the monitoring quality and subsequent treatment for these same patients.

2. Methodologies, Material and Methods

The data used to support this work refers to a Mayo Clinic trial on primary biliary cirrhosis (PBC) of the liver, carried out between 1974 and 1984. A total of 424 patients with PBC, referred to the Mayo Clinic during this ten-year interval.

To ensure ease of work and the best possible results all of the processes were based on the Cross Industry Standard Process for Data Mining (CRISP-DM) Methodology which provides a structured approach to planning a data mining project. It is a robust and proven methodology.

2.1. Business Understanding

In terms of business, it was set as an objective the prediction of survivability of primary biliary cirrhosis' patients accordingly to their medical data. The successfulness of this work will be measured through the accomplishment of some goals, namely the reduction of consultations and surgeries planning costs, the improvement of patient follow-up and medical suggestions, the forecast of organs (liver) needs for transplants, the efficient management of medical staff responsible for each case and the pre-preparation of possible symptoms and complications.

Concerning the data mining process, and being in a medical environment, it is essential to ensure a high accuracy (>90%), also a high sensitivity (>90%), and the minimal number of false negatives possible.

2.2. Data Understanding

The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The 112 additional cases did not participate in the clinical trial but allowed their basic measurements to be recorded. Six of these cases were lost to follow-up shortly after diagnosis, so the data presented here refer to a total of 418 patients and 19 important features in the identification and monitoring of Primary Biliary Cirrhosis: number of days between registration and the earlier of death, transplantation, or study analysis time; status; administered drug; age; sex; presence of ascites, hepato, spiders and edema; amount of bili, chol, albumin, copper, alk_phos, sgot, trig and platelet, prothrombin time in seconds and histologic stage of disease.

The target variable *Status* represents whether the patient was still alive at the end of the trial, if he needed a liver transplant or if he died, and can have the values 0, 1 or 2 respectively. In the next figure, it is possible to analyse the data distribution of this variable, where we can observe that 39% of the patients died during the trial.

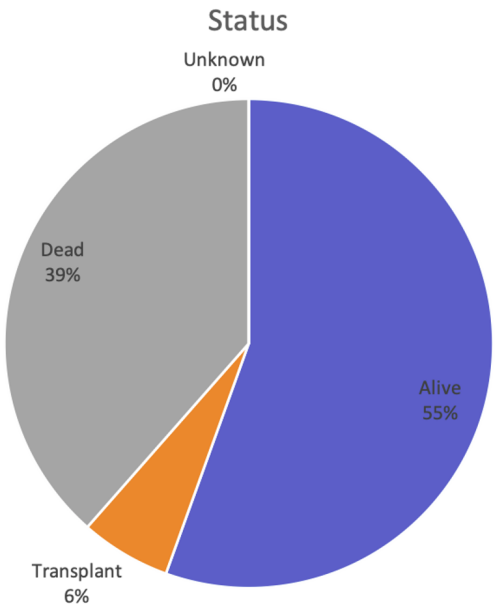

Figure 1. Data distribution of the target variable *Status*

2.3. Data Preparation

In this stage, a data cleansing swipe was performed. Firstly, because 112 cases did not participate in the clinical trial and their data is largely incomplete, resulting in a total of 1033 null values in fields like the presence of ascites or hepato and the amount of copper in the blood, there was a need to opt between the removal of those 112 entries or the removal of this attributes with null values. Considering the reasonable small size of the dataset, removing all those entries would result in only 312 remaining entries, which would be even harder to work with. After calculating the correlation matrix, it was possible to verify that those attributes did not have a strong correlation with the target variable, so not using them in the DM models was the best solution.

Besides that, to improve the performance of DM models, all values were converted to numerical types and normalized to ensure the best results, and in cases like the age, the values were discretized, in this case, into 5 bins. The target variable was simplified, keeping the only status where the patient survived or died, and later converted into a binomial type.

From an initial dataset with 424 entries, after the various operations, the result was a dataset with 448 entries, without duplicate values or with null attributes and optimized for most DM models.

2.4. Modeling

The first step of the modeling stage, and because the target variable is binomial, a receiver operating characteristic curve (ROC curve) was generated to illustrate the performance measurement for this classification problem achieve with several algorithms from different classes.

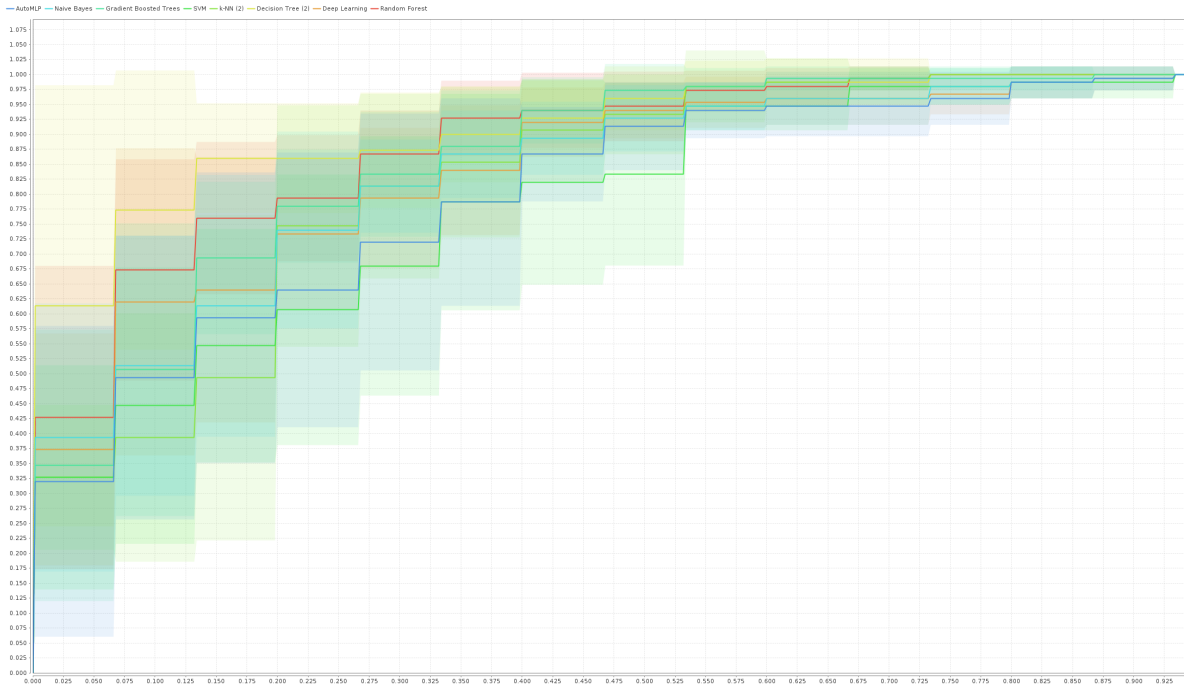

Figure 2. ROC Curve

Through the analysis of this curve it is possible to infer some conclusions about the performance of the many classes of techniques and their relevance to this work. From all the possibilities offered by RapidMiner, only 9 were used: *k-Nearest Neighbors* (KNN), *Naïve Bayes* (NB), *Support Vector Machines* (SVM), *Decision Tree* (DT), *Gradient Boosted Tree* (GBT), *Random Forest* (RF), *Deep Learning* (DL), *Neural Net* (NN) and *AutoMLP*.

In all of these techniques, the sampling method used was Cross-Validation, with 10 folds, stratified sampling, and in which all data was used for testing. Considering the fact that some more patients survived than died, it is important to analyze possible data approaches to balance the dataset and increase accuracy values. Both SMOTE upsampling and undersampling were tested, but in all models upsampling showed better results.

In terms of attributes, the starting point was the correlation matrix generated (presented below), from which the most relevant attributes to predict the status of the patient was used. The set of attributes used to model the system will now me referred as S.

Attrib...	id	futime	status	drug	age	sex	ascites	hepato	spiders	edema	bili	chol	albu...	copper	alk_p...	sgot	trig	plate...	protime	stage
id	1	-0.354	-0.282	-0.037	0.037	0.084	-0.116	-0.056	-0.097	-0.160	-0.062	0.033	-0.129	-0.099	-0.353	-0.012	-0.034	-0.077	-0.192	-0.034
futime	-0.354	1	-0.417	-0.008	-0.126	0.007	-0.307	-0.288	-0.266	-0.305	-0.404	-0.138	0.431	-0.365	0.149	-0.225	-0.153	0.151	-0.111	-0.366
status	-0.282	-0.417	1	-0.025	0.192	-0.118	0.317	0.335	0.232	0.303	0.430	0.187	-0.262	0.388	0.226	0.294	0.212	-0.084	0.339	0.320
drug	-0.037	-0.008	-0.025	1	-0.134	0.056	-0.044	0.103	0.008	-0.012	0.086	0.019	0.009	0.000	-0.018	0.042	0.009	0.034	0.073	0.066
age	0.037	-0.126	0.192	-0.134	1	-0.163	0.232	0.075	-0.050	0.197	0.002	-0.158	-0.182	0.062	-0.047	-0.150	0.022	-0.148	0.114	0.189
sex	0.084	0.007	-0.118	0.056	-0.163	1	-0.009	-0.051	0.141	-0.033	0.028	0.011	-0.030	-0.240	-0.032	0.004	-0.051	0.091	-0.070	-0.017
ascites	-0.116	-0.307	0.317	-0.044	0.232	-0.009	1	0.161	0.161	0.586	0.378	-0.058	-0.389	0.227	0.011	0.092	0.207	-0.216	0.320	0.250
hepato	-0.056	-0.288	0.335	0.103	0.075	-0.051	0.161	1	0.295	0.171	0.302	0.138	-0.299	0.234	0.110	0.136	0.160	-0.190	0.187	0.467
spiders	-0.097	-0.266	0.232	0.008	-0.050	0.141	0.161	0.295	1	0.272	0.288	0.062	-0.234	0.261	0.038	0.131	0.097	-0.161	0.246	0.292
edema	-0.160	-0.305	0.303	-0.012	0.197	-0.033	0.586	0.171	0.272	1	0.331	-0.108	-0.331	0.257	0.024	0.141	0.083	-0.204	0.332	0.243
bili	-0.062	-0.404	0.430	0.086	0.002	0.028	0.378	0.302	0.288	0.331	1	0.397	-0.314	0.457	0.117	0.442	0.437	-0.013	0.315	0.201
chol	0.033	-0.138	0.187	0.019	-0.158	0.011	-0.058	0.138	0.062	-0.108	0.397	1	-0.070	0.126	0.149	0.353	0.277	0.192	-0.031	0.011
albumin	-0.129	0.431	-0.262	0.009	-0.182	-0.030	-0.389	-0.299	-0.234	-0.331	-0.314	-0.070	1	-0.265	-0.101	-0.220	-0.103	0.159	-0.201	-0.305
copper	-0.099	-0.365	0.388	0.000	0.062	-0.240	0.227	0.234	0.261	0.257	0.457	0.126	-0.265	1	0.187	0.294	0.280	-0.064	0.218	0.269
alk_phos	-0.353	0.149	0.226	-0.018	-0.047	-0.032	0.011	0.110	0.038	0.024	0.117	0.149	-0.101	0.187	1	0.112	0.180	0.144	0.089	0.041
sgot	-0.012	-0.225	0.294	0.042	-0.150	0.004	0.092	0.136	0.131	0.141	0.442	0.353	-0.220	0.294	0.112	1	0.126	-0.120	0.112	0.165
trig	-0.034	-0.153	0.212	0.009	0.022	-0.051	0.207	0.160	0.097	0.083	0.437	0.277	-0.103	0.280	0.180	0.126	1	0.103	0.020	0.124
platelet	-0.077	0.151	-0.084	0.034	-0.148	0.091	-0.216	-0.190	-0.161	-0.204	-0.013	0.192	0.159	-0.064	0.144	-0.120	0.103	1	-0.167	-0.254
protime	-0.192	-0.111	0.339	0.073	0.114	-0.070	0.320	0.187	0.246	0.332	0.315	-0.031	-0.201	0.218	0.089	0.112	0.020	-0.167	1	0.208
stage	-0.034	-0.366	0.320	0.066	0.189	-0.017	0.250	0.467	0.292	0.243	0.201	0.011	-0.305	0.269	0.041	0.165	0.124	-0.254	0.208	1

Figure 3. Correlation Matrix

Starting from using only the most correlated attribute to status, S = bili, the amount serum bilirubin, the accuracies of the techniques were measured, and more attributes were incrementally added. An increase of performance was observed in almost all methods until S = {albumin; alk_phos; ascites; bili; copper; edema; futime; hepato; protime; sgot; spiders; stage; trig} after which an inflection point occurred and accuracies started to decrease. The only exception was *Naïve Bayes* where the inflection point was only observed before.

In terms of parameters used, in almost all the techniques the best performance was achieved using the standard configurations from RapidMiner, but in all of the neural net methods, the default were insufficient. The following changes were made:

Deep Learning: Improvement in Accuracy: $\approx$ 87.21 -> $\approx$ 92.5%
- epochs (10.0 -> 20.0)
- hidden layers (2 de 50 -> 6 de 75)

Neural Net: Improvement in Accuracy: 86.33 -> 94.33%
- training cycles (200 -> 650)
- learning rate (0.01 -> 0.02)

AutoMLP: Improvement in Accuracy: 81.67 -> 92.33%
- training cycles (10-> 20)
- number of generations (10 -> 15)
- number of esemble mlps (4 -> 8)

2.5. Evaluation and Deployment

The performance of each tested DM techniques was evaluated by analyzing the resulting confusion matrix, where we can observe the number of True Positives, False Positives, True Negatives and False Negatives from which we calculated the sensitivity, specificity and precision of this algorithms.

—— Indicators Models ——	Accuracy (%)	False negatives	Precision (%)	Sensivity (%)	Specificity (%)
KNN	82.67	12% (36)	78.82	87.69	78.82
NB	76.33	19.33% (58)	70.26	87.61	70.25
SVM	75.15	23.05% (77)	67.65	93.75	67.64
DT	83.33	16% (48)	75.51	98.07	75.51
GBT	93.67	4.67% (14)	91.19	96.45	91.19
RF	90.33	9% (27)	84.57	98.40	84.57
DL	92.37	3.44% (9)	90.35	90.35	93.91
NN	94.33	4% (12)	92.36	96.50	92.35
AutoMLP	92.33	5.33% (16)	89.94	95.03	89.93

As observed, the NN and GBT techniques achieved both the highest accuracy ($\approx$ 94%) and second highest sensitivity ($\approx$ 96%), being either one of these suited for the chosen technique.

3. Discussion

With the use of various techniques, it became possible to reach very good results in terms of sensitivity and accuracy. First of, and due to the facts that we only had a relatively small and unbalanced dataset with very little entries, the good results were only achieved using oversampling.

On the other hand, in terms of validation, it's important to highlight the chosen method, that allowed the reduction of the negative side effects caused by the shortage of the existent data. To ensure that, all the data was used for training with Cross Validation. Resorting to a large number of iterations, during which a partitioning of the data set in subsets mutually exclusive used for training is done in each one, being the remainder used for validation.

In terms of results, except for the case with the NB and SVM algorithms, all of the other algorithms gave relatively acceptable results (accuracy >80% and sensitivity >85%). There are although two

algorithms that stand out, showing an accuracy >93% and a sensitivity >96%, being them the GBT and the NN. Adding to that, a very important indicator in this area is the false negatives problematic (the prediction of the patients survival when it in fact doesn't occur), and both algorithms have good performance on that subject (GBT showed 14 false negatives - 4.67% of the predictions).

4. Conclusions and Future Work

With this work, and its subsequent results, it's proven that it is possible in an efficient and accurate way to predict the survival of patients with PBC, and this way contribute, not only to the improvement of their attendance but also on their condition in general, making it easier to adapt technical and human means to each of their needs. For the clinics, foundations and hospitals, this also represents a substantial improvement on their processes , allowing the cost reduction on the scheduling of consultations and surgeries, making it easier to manage the teams to respond adequately to each case, and supporting the need of organs for transplant purposes.

Despite all this, the basis of this project is a relatively small data set, that doesn't translate in having very trustworthy prediction results when applied to different environments, conditions and regions. So, it's imperative to guarantee the continuous improvement of the model, using more and more diverse data from different hospitals, clinics and regions improving, like so, the reliability of the obtained results, increasing the capability of the system to recognise patterns in a more global level.

Lastly, and on the path to improve the results' reliability, it's important that the work done on the created models is continuous, accompanying not only the growth of the data set that supports them, but also the technological evolution of the used tools and algorithms. DM is definitely a constantly growing study field, that day by day, is significantly improved in all aspects. Being this, a health care sector centered project, it's essential to always follow these improvements, keeping the state of it as rigorous as possible in technical and scientific terms.

Abbreviations

The following abbreviations are used in this manuscript:

DM	Data Mining
CRISP-DM	Cross Industry Standard Process for Data Mining
PBC	Primary Biliary Cirrhosis
KNN	k-Nearest Neighbors
NB	Naïve Bayes
SVM	Support Vector Machines
DT	Decision Tree
GBT	Gradient Boosted Tree
RF	Random Forest
DL	Deep Learning
NN	Neural Net

References

1. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. *Elsevier*; 2011.
2. Fayyad, U.M. ; Piatetsky - Shapiro, G. ; Smyth, P.(eds). Advances in knowledge discovery and data mining. *AAAI press*; 1996.
3. Forsyth, David, and Jean Ponce. Computer vision: a modern approach. *Upper Saddle River, NJ; London: Prentice Hall*; 2011.