APA- Practical Work 2017-2018

Albert Ribes Kerstin Winter

January 10, 2018

Contents

Todo list

Tener todos los datasets listos para revista, preparados para poder meterlos en un modelo

1 Introduction

1.1 Desciption of the work and its goals

The goal of this project is to build a classification model to predict whether a lung cancer patient will die within one year after surgery or not. To do so we will study a dataset with real lung cancer patients.

As this is very sensitive information, our priority will be to minimize the amount of false negatives, i.e., avoid predicting a patient will not die within one year when it certainly does.

The data is taken from `https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data#` [**zieba2013boosted**]

1.2 Desciption of available data

The data we are working with is about patients who underwent major lung resections for primary lung cancer in the years from 2007 to 2011. For each patient we are given information about his diagnosis and effects produced by the cancer.

The dataset is very limited in the number of instances available: it only has 470. In addition, the distribution of the predicted class isn't quite balanced, since only 70 of the patients died in one year period. This may become a problem in some of the prediction models due to the fact that the results will be biased towards the biggest class. However, we can suppose that the data has been collected uniformly and that this proportion is similar to the real one.

For each patient we have 16 different atributes. 3 of them are numerical, and the rest are categorical. From those, 10 are binary. The response atribute is also binary.

2 Related Previous Work

Hablar del paper

Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients

Que está el directorio del proyecto. Habla sobre cómo tratar dataset no balanceados, i.e., la proporción entre positivos y negativos no es nada parecida. Los métodos que propone son para las SVM, que todavía no hemos visto.

3 Data exploration process

3.1 Pre-processing

3.1.1 Treatment of missing values

Our dataset do not have missing values, so there is no need to treat them.

3.1.2 Treatment of anomalous values

The age of the patients is not well distributed. Most of them are over 60 years old, and only 4 of the patients are under 40. Due to this, it is very likely that the conclusions of our study will only be applicable to the elder people. However, we are reluctant to remove the younger patients.

3.1.3 Treatment of incoherent values

The variable FEV1 which is the Forced Expiratory Volume in 1 second, shows a few anomalously high values. Depending on factors like age and sex of a patient the average value of the FEV1 should be around 3-6 litres, whereas the dataset shows values up to 86. As most of the values are within 0 and 10 we assume that the dataset contains the FEV1 in litres. To decide which values are to be determined as outliers we calculate the FEV1/FVC ratio which gives the percentage of the lung volume exhaled in the first second over the whole exhaled volume. All patients having a unrealistic ratio higher than 100%, which are 22 patients, are determined to be outliers and eliminated. We chose not to apply any stricter constraints because the dataset does not include the sex of the patients which influences the normal values of the FEV1 much.

Source of knowledge about FEV1 and FVC: https://www.nuvoair.com/blogs/blog/do-you-know-how-to-interpret-the-results-of-your-spirometry-test

El enlace antiguo ha caido, ahora es este:
https://www.nuvoair.com/do-you-know-how-to-interpret-the-results-of-your-spirometry-test
Referenciar quizá el histograma con FEV1 Poner bien la referencia a la página

3.1.4 Coding of non-continuous or non-ordered variables

As most of the dataset variables are logical, we have coded them as logical in R. We have converted the variable “DGN” to many binary variables, each one saying wether the patient showed that diagnosis or not.

Originally, the variables “PERFORMANCE” and “SIZE” were categorical, but as they seem to have some kind of order, we have coded them as numeric. The “PERFORMANCE” can have values $1, 2, 3$ and the variable “SIZE” can have values $1, 2, 3, 4$. Both of them are then normalized.

“AGE” is also normalized in the range $[0, 100]$

3.1.5 Possible elimination of irrelevant variables

Some of the variables of our dataset are not well represented. In particular:

DGN	There is just one patient with DGN = 1 and just 8 have DGN = 8
PAD	Just 8 patients have PAD = True
ASHTMA	Just 2 patients have ASHTMA = True
MI	Just 2 patients have MI = True

As we have very few instances we will train and run the models on two different datasets. One of them, thoraric.original, will contain all of the original attributes, and the other , thoraric.removed, will contain all attributes but “DGN.1”, “DGN.8”,“PAD”,“ASHTMA” and “MI”.

3.1.6 Creation of new useful variables (Feature extraction)

We have not created any new variable as most of the ones we have are logical and it doesn’t seem to be any relation among them.

3.1.7 Normalization of the variables

We need to normalize only our numeric variables, which are the AGE, FVC and FEV1. As we have converted the variables “PERFORMANCE” and “SIZE” to numeric (as we will see in section 3.1.8) we also need to normalize them. To normalize the age we will only consider cases between 0 and 100 years old. For FVC and FEV1 the range will correspond to the maximum and minimum observed values with a margin of 10%.

3.1.8 Transformation of the variables

Acording to the paper we found the accetable range for skewness in a numeric variable is $(-2, +2)$. In the original dataset, the one which contained all the patients, the skewness of variables AGE, FVC and FEV were out of this range. But after eliminating 22 of the patients (in 3.1.3) all of them are inside the acceptable range, so there’s no need to transform them.

3.2 Clustering

hacer varios k-means con distintos valores de k (2,3,4,5,6) para ver si descubrimos algún cluster que nos permita crear una variable nueva

3.3 Visualization

Hacer MCA

4 Resampling protocol

After the pre-processing we have 2 datasets with 448 instances. *thoraric.original* has 23 attributes and *thoraric.removed* has 19. 0.85% of the patients have the target variable (DIED) equals FALSE.

The following lines explain the resampling protocol that we have used. All the numbers used can easilly be tuned in the scripts provided.

To test our models we split our data into two different datasets, one for training and one for testing. The testing dataset will contain $\frac{1}{3}$ of all the patients. As they are chosen randomly, it is expected that the proportion of patients in each of the classes is kept. We will only use the testing dataset in the end, to test our models.

In the training dataset we have 299 patients, and it is expected that over 254 of them will be DIED $=$ FALSE and 45 will be DIED $=$ TRUE. As most of the models are sensitive to non-balanced datasets, we need to do something to balance our training dataset. We use the *bagging* algorithm to generate 51 balanced datasets, and each one of them will be used to train 51 models of the same type. That number could be different, but it is good for it to be odd, to avoid ties. Then, to generate a prediction, each of the models will be used to make a *hard* vote, and the class predicted by the majority will be the answer. This way we have constructed what we have called a *super-model* built with simpler models.

Each of the *bags* will contain the same amount of instances. As we want them to be balanced, all of them will contain every TRUE instance in the training dataset, and a random sample of the same size of the FALSE instances. Thus, each bag will have over $2 \times 45 = 90$ instances.

5 Results obtained using linear/quadratic methods

5.1 Naive Bayes

Buscar la biblioteca que lo calcula y aplicarlo a nuestros datos

- Comentar las decisiones que se han tomado (cuidado con los valores por defecto). p.e: Por qué hemos usado distancia Manhatan en vez de Euclidea en knn
- Comentar los hiperparámetros que se han obtenido por crossvalidation. Como no vamos a hablar de cada uno de los modelos (51), hablar de la media de todos ellos, o algo así
- Comentar los resultados que se ha obtenido. Intentar ver por qué ha ido bien o mal, y cosas así
- Poner la confussion matrix

5.2 KNN

Si no hacemos nada para desbalancear los datos, hay que usar una k muy pequeña

Quizá es recomendable usar algún método para balancear los datos, y probar así otros valores de k

5.3 LDA

Suponiendo que las varianzas de cada una de las clases son la misma, se usa este algoritmo, (que simplifica QLA) para ver la probabilidad de pertenencia a una clase

5.4 QDA

5.5 RDA

5.6 Logistic Regression

Mirar el vecino más cercano para precedir

Si suponemos que las variables son independientes: -Haces naive Bayes para ver la probabilidad de que pertenezca a cada una de las clases (habría que estudiar si las variables son independientes) - Logistic regression

6 Results obtained using non-linear methods

7 Desciption and justification of the final model chosen

7.1 Estimation of the generalization error

8 Self-assessment of successes, failures and doubts

9 Scientific and personal conclusions

10 Possible extensions and known limitations

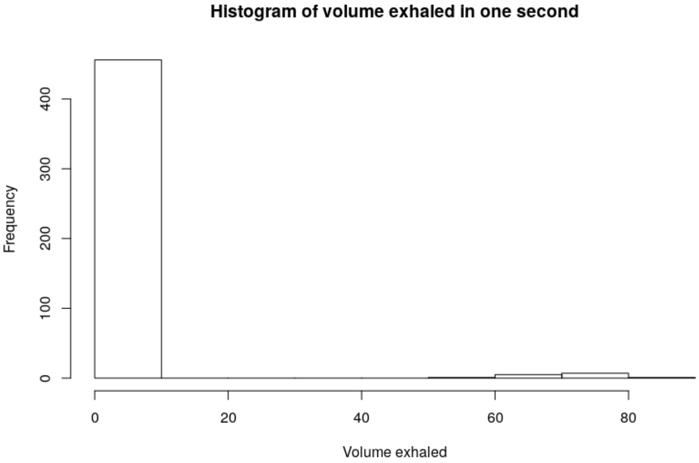

Figure 1: Show the ammount of people having each value