APA- Practical Work 2017-2018

Albert Ribes Kerstin Winter

December 13, 2017

Contents

Todo list

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

1 Pasos a seguir

- Dividir el dataset en un tercio de test y dos tercios de train
- Generar 10 datasets usando el dataset de train, de manera que cada uno de esos 10 datasets tenga la misma cantidad de Trues y de Falses. Los True serán siempre los mismos, los False serán samples. (bootstraping)
- Para cada modelo que conocemos
  - Lo entrenaremos con cada uno de esos 10 datasets. Si el modelo necesita parámetros, probaremos distintos parámetros, y cojeremos el que vaya mejor en la mayoría de esos 10 datasets.
  - Ahora tenemos 10 clasificadores. Generamos un clasificador nuevo, que pregunta a cada uno de esos 10 clasificadores y decide lo que decida la mayoría
- Ahora tenemos un clasificador para cada modelo que conocemos.
- Enfrentamos a cada uno de estos modelos con los datos de test
- Ponemos en el report los resultados

2 Introduction

2.1 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**]

2.2 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.

3 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.

4 Data exploration process

4.1 Pre-processing

4.1.1 Treatment of missing values

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

4.1.2 Treatment of anomalous values

Quizá hay que quitar algunas personas por ser demasiado jóvenes comparadas con el resto

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.

4.1.3 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 is 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.html

Referenciar quizá el histograma con FEV1 Poner bien la referencia a la página

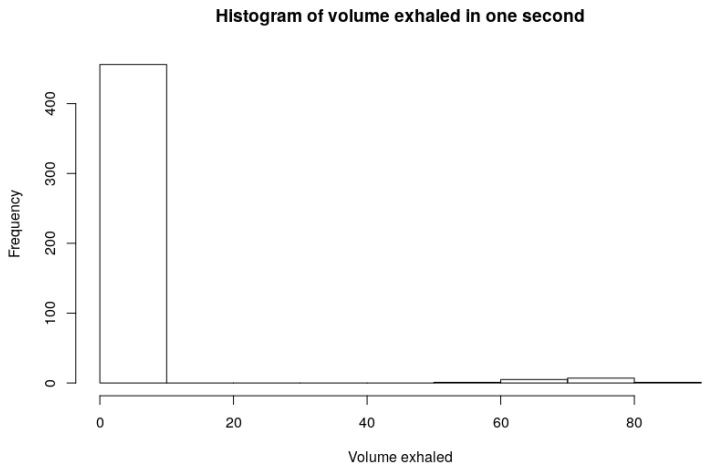

Figure 1: Show the ammount of people having each value

4.1.4 Coding of non-continuous or non-ordered variables

La mayoría de variables son binarias, y establecemos su tipo en "binary"

Las variables "PERFORMANCE" Y "SIZE" sí que tienen un orden, por lo tanto las definimos como numeric (integer)

AGE se queda como está, numeric

Tanto PERFORMANCE como SIZE como AGE se normalizarán

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 $[0, 100]$

4.1.5 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

Hence the normal thing to do would be to eliminate those columns in our dataset, since they do not provide good information. However, as we have very few instances we can afford to try out keeping and removing them. We will run our models with each of the 16 combinations of keeping DGN1, DGN8, PAD and ASHTMA and we will see which one works better.

Mirar si podemos agrupar varios DGN que se parezcan (preguntar a alguien que sepa)

Haremos 2 datasets, uno que sea el original sin quitar nada, otro en el que quitaremos cosas

Seguro que quitamos DGN1, DGN8, MI, y ASTHMA

4.1.6 Creation of new useful variables (Feature extraction)

Entender cómo funciona MCA, y ver si podemos sacar una variable nueva

Quizá es interesante añadir la variable FEV/FEV1

4.1.7 Normalization of the variables

We need to normalize only our numeric variables, which are the AGE, FVC and FEV1. 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%.

Si finalmente añadimos el atributo FEV1/FVC, indicar que no hace falta estandarizarlo puesto que ya lo está

Puesto que al final PERFORMANCE y SIZE las hemos puesto como numéricas, también las hemos normalizado. Explicarlo

4.1.8 Transformation of the variables

Acording to the paper we found the accetable range for skewness in a numeric variable is $(-2, +2)$. The skewness of our original variables AGE, FVC and FEV are:

AGE	-0.1899413
FVC	0.5417132
FEV1	5.597584

But we have to take into account that we've eliminated 22 patients, so the new values are:

Poner los nuevos valores

As the three variables are in the specified range, there is no need of transforming them.

Referenciar (y leer un poco...) el paper

4.2 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

4.3 Visualization

Hacer MCA

5 Resampling protocol

6 Results obtained using linear/quadratic methods

6.1 Naive Bayes

Buscar la biblioteca que lo calcula y aplicarlo a nuestros datos

6.2 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

6.3 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

6.4 QDA

6.5 RDA

6.6 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