Data science–Fundamental of Data Science
FABIO GALASSO

Pages 1-3
Dec 2021

Logistic Regression: Heart disease Prediction

Seyed Behdad Ahmadi Matr. 1985602, Saeed Zohoorianmoftakharkhodaparast Matr. 1955809, Mohammadmehdi Razavi Matr. 2023856, Seyed Mohammad Mousavi Nishabouri Matr. 1922872, Altynai Toiguliyeva Matr. 2027713
Fundamental of Data Science
Sapienza, Universit `a di Roma
26/12/2021

Keywords: Logistic regression, image classification, Heart disease

Abstract

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things. Various studies give only a glimpse into predicting heart disease with ML techniques. In this project we propose a method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with Logistic Regression techniques.

Introduction

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K -Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB). The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. In this work, numerous readings have been carried out to produce a prediction model using not only Logistic Regression but also image classification of heart beats to have a better consideration. The main objective of this project is to improve the performance accuracy of heart disease prediction.

Prepared Method Explained

In the part of working with a structured dataset we first cleaned our data by determining how many NaN values we have and plotting a histogram of missing values. After removing missing values, we decided to see our categorical variables distribution. After that we plot variables correlation to see which features are correlated with our target. We found out there are some features that are highly correlated with each other. We trimmed some of them to reduce the dimension of our features.

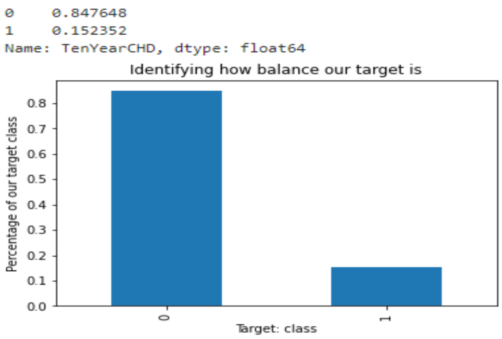
- “sysBP” and “diaBP” are highly correlated. (0.79)
- “prevalentHyp” and “sysBP” are highly correlated. (0.7)
- “currentSmoker” and “cigsPerday” are highly correlated. (0.77)

Then we check how each of these features are correlated with our target (“TenYearCHD”)

- “currentSmoker” = 0.019
- “diaBP” = 0.15
- “prevalentHyp” = 0.18

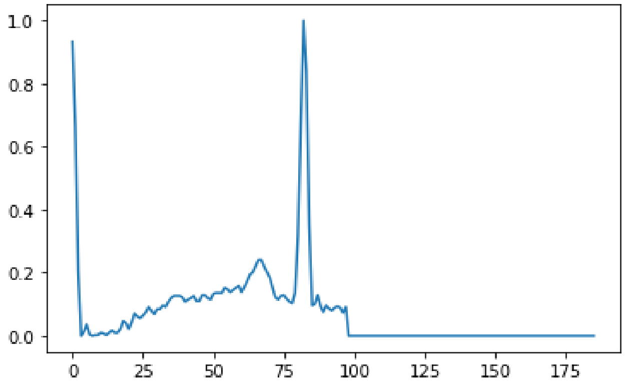
We found important features by performing a Logistic Regression and evaluating coefficients. Also we used RandomForestClassifier to find important features. We achieved some evident that showes the most important features are 1,9,10,11,12,13,14 ['age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']. Third method we used to make sure was the chi-squared (chi²) statistical test.

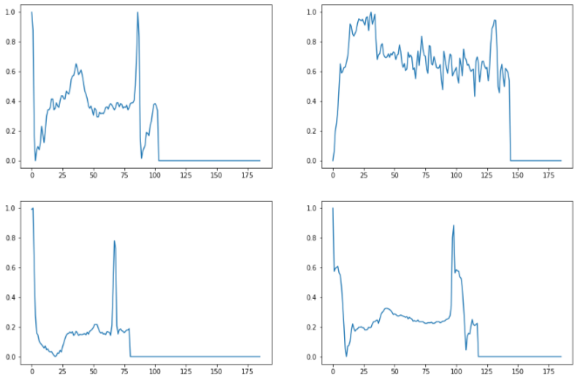
But before that we decided to see how our target is balanced to know which metrics we should use to evaluate our model. Our dataset is imbalanced so we used confustion_matrix and roc_auc_score. Also we over-sample our dataset to have a balanced dataset.



```
0    0.847648
1    0.152352
Name: TenYearCHD, dtype: float64
```

In image classification we split the dataset into 2part, one for train test which contains the ¾ of the main data, and the rest for test data. By looking at the data we understood that our data was not balance. Most of the samples are type ‘0’ which is a normal heartbeat. On the other hand, we have the least number of Fusion Beats. So, we resample the data next, and get a dataset including 100,000 samples where each category has 20,000 samples respectively. After that we visualized one sample from each category that you can see below. Then we plot one sample from 4 abnormal heartbeats.





Then we work on building our Neural Network. Since our data is image, and these images represent a signal we decided to use CNN instead of other methods like RNN. In our model we used three CNN layers. For CNN layers we used Relu activation function since these days this method works better than the others. After each CNN layer we used BatchNormalization which applies a transformation that maintains the mean output close to 0 and the output standard deviation close to 1. Following by Batch-normalization, using MaxPool is a good option. MaxPool is used for down sampling the input representation by taking the maximum value over a spatial window of pool size. After that, we flatten the output to be ready as input of next dense layers. The learned features are flattened to one long vector and pass through a fully connected three layers’ network before the output layer used to make a prediction. In the last dense layer, we used softmax activation function because our output is multiclass classification. For this model, we will use a standard configuration of 64 parallel feature maps and a kernel size of 3. The feature maps are the number of times the input is processed or interpreted, whereas the kernel size is the number of input time steps considered as the input sequence is read or processed onto the feature maps. Once the model is fit, it is evaluated on the test dataset and the accuracy of the fit model on the test dataset is returned.

Dataset and Benchmark

This dataset contains information concerning heart disease diagnosis. The data was collected from Kaggle, and it is available at Kaggle Repository. A data frame with 4239 rows,16 columns in which 7 string, 7 integers and 2 decimal observations on the following 16 parameters in table 1.

Table 1			
P1-Age	P2-Gender	P3-Education	P4-Current Smoker
P5-Cigs.Per day	P6 – BPMeds	P7-Prevalent stroke	P8-Prevalent HYP
P9-Diabetes	P10-TotChol	P11-sysBP	P12-DiaBP
P13-BMI	P14-Heart rate	P15-Glucose	P16-Ten Year CHD

Table 2
N: Non-ecotic beats (normal beat)
S: Supraventricular ectopic beats
V: Ventricular ectopic beats
F: Fusion Beats
Q: Unknown Beat

We also worked on the MIT-BIH Arrhythmia Database which contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. Twenty-three recordings were chosen at random from a set of 4000 24-hour ambulatory ECG recordings collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital; the remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias that would not be well-represented in a small random sample. In our dataset we have 109446 samples which fall into five categories. The frequency of sampling is 125Hz. The five class of our data is in table 2.

Experimental Result

In image classification part by using CNN we achieved **97.272 %** accuracy with batch size 8, **96.944%** with batch size 16 and **97.368 %** with batch size 32 which are efficient numbers. So, with this model we can predict five mentioned problems in ECG pictures. So, this can be very helpful in finding heart disease.

Since we worked on image classification and logistic regression separately below you can see information about our logistic regression part. Our selected features are: age, totChol, sysBP ,diaBP ,BMI, heartRate, glucose, cigsPerDay All of selected features are numerical. So we normalized them to improve our model. After we had a balanced dataset and you can see the results below.

	F1 Score	ROC AUC	Total
Stacking Classifier	**0.97**	**0.97**	**0.97**
Random Forest	**0.93**	**0.94**	**0.93**
K-nearest Neighbors	**0.88**	**0.87**	**0.87**
Support Vector Machine	**0.67**	**0.66**	**0.66**
Gaussian Naïve Bayes	**0.53**	**0.63**	**0.62**
Logistic Regression	**0.66**	**0.66**	**0.66**

Conclusion and Future Work

In this Project, some techniques were utilized to select significant attributes from Kaggle heart dataset to improve the performance of machine learning classifiers when predicting heart disease risk. A remarkable performance was achieved by the Stacking classifier compared with Logistic Regression. Eventually, we noticed that there was a significant improvement in the prediction performance with appropriate attribute selection and tuning the parameters like balancing using feature engineering. Although the performance of the classifiers looks satisfactory, 6 machine learning classifiers, and 2 feature selection methods were used in this research. There is a huge scope to explore various machine learning algorithms and feature selection techniques. In the future, it could be great to combine multiple datasets to obtain a higher number of observations and conduct more experiments by selecting appropriate attributes to improve the classifier's predictive performance and also our image classification.

References

1. https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease prediction

2. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists

3. https://www.python.org/psf/codeofconduct/

4. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

5. API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.

6. WHO. Available online: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 (accessed on 9 February 2021).

7. Ed-Daoudy, A.; Maalmi, K. Performance evaluation of machine learning based big data processing framework for prediction of

8. heart disease. In Proceedings of the International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, 26–27 December 2019; pp. 1–5

9. Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. Health Technol. 2020, 10, 1137–1144.

10. Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. Mater. Today Proc. 2021, 22, 660–670.

11. Gazelo ˘glu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. Prog. Nutr. 2020, 22, 660–670.