|  |
| --- |
| Data Science with R II (JEM220) – Charles University Prague, FSV |
| Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients |
| by Anchana Khemphila & Veera Boonjing |

|  |
| --- |
| Analysis and improvements by Erik Nemcik & Paul Mainka  17.5.2020 |

Overview

In their study Anchana Khemphila and Veera Boonjing compare the performance of logistic regression, decision trees and artificial neural networks in classification problems. For this purpose, the authors employ the Cleveland Heart Disease dataset on the development of heart disease among individuals.

In order to evaluate the performance of the three different methods, the authors use measurements of area under the curve (AUC), sensitivity, specificity, accuracy and the error rate. Since artificial neural networks have the lowest error rate and the highest accuracy in their estimations, they conclude that artificial neural networks is the most suitable of the three classification techniques assessed for this data set.

# Data

The data set used by Boonjing and Khemphila includes 303 individuals, 13 explanatory variables and the dependent variable whether the patient has developed a heart disease or not. The data set is quite balanced with 165 patients having developed a heart disease, which corresponds to approximately 54% of all the patients. Of the 13 explanatory variables, 8 are factor variables and 5 are continuous variables.

Description of the eight factor variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Name | Meaning | Different Levels | Frequency of occurrence |
| 1 | Sex | Patient gender | Female | 96 (32%) |
| Male | 207 (68%) |
| 2 | Chest Pain Type |  | Angina | 143 (47%) |
| Abnormal | 50 (17%) |
| Nonanginal pain | 87 (29%) |
| Asymptotic | 23 (8%) |
| 3 | Fasting Blood Sugar | Is fasting blood sugar less than 120? | True | 45 (15%) |
| False | 258 (85%) |
| 4 | Resting ECG | The electrocardiogram (ECG) measures the heart’s electrical activity, and a resting ECG is administered when the patient is at rest. | Normal | 147 (49%) |
| Abnormal | 152 (50%) |
| Left ventricular hypertrophy | 4 (1%) |
| 5 | Induced Angina | Does the patient experience angina as a result of exercise? | True | 99 (33%) |
| False | 204 (67%) |
| 6 | Slope | Slope of the peak exercise ST segment. | Up | 21 (7%) |
| Flat | 140 (46%) |
| Down | 142 (47%) |
| 7 | Number Coloured Vessels | Number of major vessels coloured by fluoroscopy. | 0 | 175 (58%) |
| 1 | 65 (21%) |
| 2 | 38 (13%) |
| 3 | 20 (7%) |
| 4 | 5 (2%) |
| 8 | Thal | Thalassemias, blood disorders characterized by decreased hemoglobin production | Normal |  |
| Fixed defect |  |
| Reversable defect |  |

Description of the 5 continuous variables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # | Name | Meaning | Min | Median | Mean | Max |
| 9 | Age | Age in years | 29 | 55 | 54.37 | 77 |
| 10 | Blood Pressure | Resting blood pressure upon hospital admission | 94 | 130 | 131.6 | 200 |
| 11 | Cholesterol | Serum cholesterol | 126 | 240 | 246.3 | 564 |
| 12 | Maximum Heart Rate | Maximum heart rate achieved. | 71 | 153 | 149.6 | 202 |
| 13 | Ole Peak | ST depression induced by exercise relative to rest. | 0 | 0.80 | 1.04 | 6.20 |

# Theory/Model

The authors take a data mining approach, implying that they do not make assumptions regarding the mode of influence of the various explanatory variables on the dependent variable. Instead, they let the data speak for itself. The aim of the paper is to determine which of the three presented techniques provides the best estimations when applied to new data. The three approaches used are Logistic regression, Artificial neural networks and Classification and Regression Trees. For this purpose, the authors of the study draw a random sample of 60% of the data set to train the models and then test their predictive power on the remaining 40%. By knowing the true outcomes of the test data set, they can then evaluate the accuracy of the prediction of the different models.

Improvements to the Study

We have a number of criticisms of the quality of the study and suggestions for improvement to increase the validity of the results.

First of all, we would like to criticize the form of the paper. The readability is partly limited, for instance due to missing spaces in the text and spelling mistakes.

More important, however, are technical comments, especially that Boonjing and Khemphila do not perform cross-validation. Our main contribution to improve the validity of the study is to utilize cross validation for the different methods (Logistic regression and neural networks) or to employ more sophisticated methods based on the same idea (bagged trees or random forest for decision trees). Without cross validation, the overall performance of the estimation methods rely heavily on the random training sample drawn. As we show in our study, the values of the quality measures area under the curve (AUC), sensitivity, specificity, accuracy and the error rate vary strongly with different splitting into training and test data set. A reliable statement about which data mining technique gives the best scores can therefore only be made by averaging the performance over many variations. Furthermore, we will use some more sophisticated methods, such as grid search to find the optimal parameters for our added models, to improve the robustness of the results.

More specifically, we will:

1. Replicate their estimations using Logistic regression, neural networks & decision trees
2. Expand their methods by introducing Random Forests, Bagged Decision Trees & Gradient Boosting for Decision Trees
3. Check the validity of the results of 1. and 2. by cross validation using all techniques
4. Employ a new technique called xgboost, which is an implementation of a gradient boosting technique.

Replication

To replicate the results of Boonjing and Khemphila we will pursue the same approach as them. We split the data set into a 60% training dataset and the remaining 40% we use as pseudo out of sample test data. Since it is a random split into test and training data, we cannot assume that we will obtain the same results as Boonjing and Khemphila. The decisive factor is whether we get a comparable ranking in the goodness of fit of the three different techniques. The authors of the study do not specify at which cutoff point they assign an estimated individual as result 1 (=has heart disease) or 0 (=has no heart disease). Since the aim of the study is to find the highest possible sensitivity, specificity and accuracy of the three classification techniques, we optimize the cutoff once for achieving maximum accuracy and once for finding the maximal point of a trade-off between sensitivity and specificity. We calculate the error rate at the optimal cutoff for accuracy as at this point the error should be minimized. As an additional measure of performance we will use the Area under the Curve value, which is threshold independent.

Comparison of performance in the training data

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Performance measure | Study results | Our results |
| Decision trees | Accuracy | 91.2 |  |
| Sensitivity | 90.1 |  |
| Specificity | 92.8 |  |
| Error rate | 0.082 |  |
| Area under the Curve |  |  |
| Neural networks | Accuracy | 92.3 |  |
| Sensitivity | 91.7 |  |
| Specificity | 92.9 |  |
| Error rate | 0.077 |  |
| Area under the Curve |  |  |
| Logistic regression | Accuracy | 91.7 |  |
| Sensitivity | 90.8 |  |
| Specificity | 92.8 |  |
| Error rate | 0.082 |  |
| Area under the Curve |  |  |

Comparison of performance in the test data

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Performance measure | Study results | Our results |
| Decision trees | Accuracy | 79.3 | 80.32 |
| Sensitivity | 81.7 | 92.75 |
| Specificity | 76 | 64.15 |
| Error rate | 0.21 | 0.2459 |
| Area under the Curve |  | 0.7798 |
| Neural networks | Accuracy | 80.2 | 66.39 |
| Sensitivity | 81.1 | 69.56 |
| Specificity | 78.7 | 54.7 |
| Error rate | 0.198 | 0.3442 |
| Area under the Curve |  | 0.624 |
| Logistic regression | Accuracy | 77.7 | 84.4 |
| Sensitivity | 81.2 | 82.6 |
| Specificity | 73.1 | 84.9 |
| Error rate | 0.22 | 0.1639 |
| Area under the Curve |  | 0.8928 |

Random Forests, Bagged Decision Trees & Gradient Boosting for Decision Trees

We use three more sophisticated methods of decision trees - Bagged Decision Trees, Random Forests, Gradient boosting - to improve prediction performance.

Bagging is a version of bootstrap aggregation, whereby the random samples with replacement are drawn from the data. From each sample a decision tree is constructed, and the final model is then the average of all the individual decision trees. We use 500 bootstrap samples in our specification.

The Random Forest Approach operates similarly to Bagged Decision Tress but in addition to bootstrapping, it seeks to uncorrelate the individual trees of each random sample. This is achieved by randomizing the set of explanatory variables that each tree can use. For each node of a growing tree, a random subset of m variables is drawn from the total of variables p and the best variable and the best split from this subset m is selected. For our random forest we use the standard specification of the Random Forest package, i.e. the minimum node size is 1, no maximum node size is defined, the number of trees is 500, m is the square root of p and we have no importance measurement. The reason we use the standard specifications is that our data set does not contain any unusual values or many meaningless explanatory variables and because there are no unambiguous rules for tuning the parameters of Random Forest.

Gradient boosting algorithm with application to decision trees produces an ensemble of weak prediction models and converts it into a strong learner. In practice, each newly produced tree is fit on a modified sample from the original data set. Each observation is then given a specific weight based on the model’s ability to classify it. Observations that are difficult to classify are given higher weights. On the adjusted data, a second tree is grown and prediction error is calculated on the ensemble of the two trees. A third tree is grown on the newly created residuals and the process repeats. The goal is to improve step by step the prediction ability of the final model. Therefore, the final model is a weighted average of the previously grown trees.

Hyperparameter Tuning

To further improve the predictive ability of the employed models, we have implemented a hyper parameter tuning algorithm for both xgboost and the artificial neural network. This way, we are able to optimize the parameters of the model while still maintaining the cross-validating function. For xgboost, we focus on the maximum depth of the grown trees, weights of the observations, subsampling and gamma (minimum loss reduction) and eta (adjusts learning rate to prevent overfitting) parameters. Because of the sheer number of parameters, we do not perform a grid search but rather perform 100 random combinations of the parameters to identify the best combination. For the neural network, we select the node size and decay (regularization parameter to avoid overfitting) parameters.

|  |  |  |
| --- | --- | --- |
| Method | Performance measure | Our results |
| Bagged Trees | Accuracy |  |
| Sensitivity |  |
| Specificity |  |
| Error rate |  |
| Area under the Curve |  |
| Random Forests | Accuracy |  |
| Sensitivity |  |
| Specificity |  |
| Error rate |  |
| Area under the Curve |  |
| xgboost (tuned) | Accuracy |  |
| Sensitivity |  |
| Specificity |  |
| Error rate |  |
| Area under the Curve |  |
| Neural Network (tuned) | Accuracy |  |
| Sensitivity |  |
| Specificity |  |
| Error rate |  |
| Area under the Curve |  |

Conclusion

In conclusion, the paper presented by Boonjing and Khemphila lacks basic scientific qualities. Our project therefore focuses on improvements to this paper to demonstrate proper techniques to be correctly employed. At first, we show that the results published by the paper are heavily dependent on the random sample generated from the original data. The resulting accuracy can vary anything between X and Y for logistic regression, X and Y for decision trees and X and Y for neural networks. We can clearly see that while the paper claims the superiority of the neural network approach, it seems that this might not be the case in reality.

To overcome this issue, we employ a cross-validating solution to randomly generate 500 samples from the dataset and train the models on these samples. In the end, we report the average accuracies (along other measures) to correctly show the actual predictive abilities. To further improve the results, we enhance the decision trees by utilizing bagged decision trees, random forests and gradient tree boosting. Furthermore, for our neural network and the gradient boosting algorithm (xgboost), we perform a parameter tuning. The resulting accuracies are X, Y, Z, W for bagged trees, random forests, xgboost (tuned) and neural network (tuned), respectively.

It can be clearly shown that the improved techniques dominate the original approach employed by the paper. The results are sample-robust and the run with close-to-optimal parameters. On average, we managed to improve the predictive ability (measured by accuracy) by X%. One might, therefore, believe that a large share of the information from the dataset is extracted by the models and further improvements would be only marginal.

# References

A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krackow, 2010, pp. 193-198, doi: 10.1109/CISIM.2010.5643666