International Journal of Scientific Engineering and Technology ISSN : 2277-1581
Volume No. 7, Issue No. 9, PP : 85-89 1 Sept. 2018

Comparison of Machine Learning Algorithms for prediction mortality in patients with Hepatocellular Carcinoma

Karolina AlicjaSala
Department Mathematics and Computer Science, Uniwersity of Warmia and Mazury in Olsztyn, Poland
karolinaalicjasala@gmail.com

Abstract : Hepatocellular cancer (HCC) is one of the most common cancers in the world. Early diagnosis is key to curing the patient. Unfortunately, the diagnosis of HCC is too often diagnosed with advanced disease. As a result, at this late stage there is no possibility of effective treatment. In this paper, compared the performance of five difference machine learning algorithms on the task of classifying patients with Hepatocellular Carcinoma. Performed multiple comparisons of classification methods, namely Logistic Regression, Decision Trees, Random Forests, Bagging, and the Boosting. Simulated training data were generated from dataset collected at a University Hospital in Portugal. The results showed that the Random Forests Classifier achieve the best accuracy for this issue.

Keywords— Logistic Regression, Decision Trees, Random Forests, Bagging, Boosting

I. Introduction

Hepatocellular carcinoma (HCC) is the most common histological type of liver cancer. It ranks first among adult liver cancer - 80-85%. In terms of mortality, it is in third place among all cancers. Most cases of HCC are caused by chronic HCV and HBV infection (Hepatitis B and C), advanced liver fibrosis or cirrhosis (alcoholism is the most common cause of liver cirrhosis). Other factors that can cause HCC include contraceptives, androgenic anabolic agents, and smoking. Adding risk factors increases the probability of cancer [1][2]. The prognosis of patients with HCC depends not only on the stage of the disease, but also on the liver function at the time of diagnosis. In 40% of cases, the cancer is asymptomatic for a long time, and at the same time almost 90% of cancer cases develop in the context of chronic liver disease. Early diagnosis of hepatocellular carcinoma is possible in 30-60% of cases [3].

The field of machine learning has been intensively expanded. A multiplicity of statistical, machine learning approaches and methods for solving the problem of classification and regression are presently available. In order to choose the algorithm appropriate for a given problem, it is worth comparing several of them on the same set of data. In this paper, studied the performance of all five models on the same data set of Hepatocellular Carcinoma. Focused on the achievable accuracy (classification accuracy) of different classification algorithms.

In the following section, the previous literature on the analysis of the presented problem was reviewed. In Section 3, described briefly characteristics of the several discussed algorithms, which allows to find most efficient model. Section 4 gives details about the data set. In Section 5, constructed a classifier from real data, and report results, which demonstrate accuracy of the classifiers. The final conclusion of the experiments are presented in the last section.

II. RelevantWork

A several of research work has been carried out in relation to HCC. The aim of the study presented by Ramanathan M. Seshadri [4] (and others) was to compare the survival in patients with stage I and II HCC undergoing liver transplantation (LT) or liver resection (LR). Performed statistical analysis. Calculated counts, percentages, means and standard deviations for patient-specific demographics. Survival was calculated in months from the date of diagnosis to the date of death or the date of the last contact. The Cox proportional hazard model was used univariate and multivariate analyses of time to date data as well as Cox's proportional risk models adapted to the procedure, stage, size of the tumor, age and gender. It turned out that the median survival for LT patients was significantly better in comparison with patients with LR in both stage I and II HCC.

In the paper [5], the authors studied the prognosis of hepatocellular carcinoma. Combines the Bayesian network (BN) with importance measures to identify key factors that have significant effects on survival time. The model's effectiveness was checked and statistical analysis was performed. As a result, it was identified as the most significant predictor of patient survival time with HCC.

In the research by Kumardeep Chaudhary[6], presented a model based on deep learning (DL) on HCC. Obtained the labels from K-means clustering, and built supervised classification model using SVM algorithm. Used the metrics closely reflect the accuracy of survival prediction in the subgroups identified. Compared the performances of approaches.

In the paper by Miriam Seoane Santos [7], A methodology based on a cluster-based oversampling approach was proposed. Two methods of classification of neural networks and logistic regression were applied. Three different measures were used: accuracy, AUC and F-Measure. The proposed methodology was assessed using a set of data that was also used in this study.

III.Dataset

A. Data description

The data set used for the experiment in this paper, is available in the Machine Learning Repository [8], which was collected at a University Hospital in Portugal for 1 year. It contains several demographic, risk factors, laboratory and overall survival features of 165 real patients diagnosed with HCC. It consists 49 different attributes with categorical, integer and real types. The data are heterogeneous with 23 quantitative variables, and 26 qualitative variables. Missing about 10.22% data on the whole dataset. As the target attribute selected a binary variable with values 0 (dies) and 1 (lives), that represent 63 cases labeled as "dies" and 102 as "lives". This small and unbalanced data sets that take into account the heterogeneity of patients with liver cancer.

The parameters are: Gender, Symptoms, Alcohol, Hepatitis B Surface Antigen, Hepatitis B e Antigen, Hepatitis B Core Antibody, Hepatitis C Virus Antibody, Cirrhosis, Endemic Countries, Smoking, Diabetes, Obesity, Hemochromatosis, Arterial Hypertension, Chronic Renal Insufficiency, Human Immunodeficiency Virus, Nonalcoholic Steatohepatitis, Esophageal Varices, Splenomegaly, Portal Hypertension, Portal Vein Thrombosis, Liver Metastasis, Radiological Hallmark, Age at diagnosis, Grams of Alcohol per day, Packs of cigarets per year, Performance Status, Encefalopathy degree, Ascites degree, International Normalised Ratio, Alpha-Fetoprotein (ng/mL), Haemoglobin (g/dL), Mean Corpuscular Volume (fl), Leukocytes(G/L), Platelets (G/L), Albumin (mg/dL), Total Bilirubin(mg/dL), Alanine transaminase (U/L), Aspartate transaminase (U/L), Gamma glutamyl transferase (U/L), Alkaline phosphatase (U/L), Total Proteins (g/dL), Creatinine (mg/dL), Number of Nodules, Major dimension of nodule (cm), Direct Bilirubin (mg/dL), Iron (mcg/dL), Oxygen Saturation (%), Ferritin (ng/mL), Class: nominal (1 if patient survives, 0 if patient died).

B. Data preparation

The main problem is the lack of data. Only 8 records have complete information. Therefore, abandoned the deletion of the row with missing data. This could cause unreliability of further calculations. Choose the imputation method for replacing absent fields. Imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing [9]. employed a supervised method to handle with missing data.

Existing k-NN imputation methods for dealing with missing data are designed according to Minkowski distance. This algorithm is be generally efficient for numerical data[10]. In this method, k neighbors are selected on the basis of a certain distance measure, and their average is used as an imputation estimation. The method requires selection of the number of nearest neighbors and distance metrics. k-NN can predict both discrete attributes and continuous attributes that occur in this data set. In k-NN classification, an object is classified by a majority vote of its neighbors. It takes a bunch of labelled points and uses them to learn how to label other points. The object is assigned to the class that is most common among its k, the number of neighbors [11]. To label a new point, it looks at the labelled points closest to that new point and has those neighbors vote so whichever label the most of the neighbors have is the label for the new point.

Calculated the correlation for attributes to find variability comparison between categories of variables. Attributes are highly correlated. Most of the input variables are continuous. The correlation matrix of variablesisshown in Figure 1.

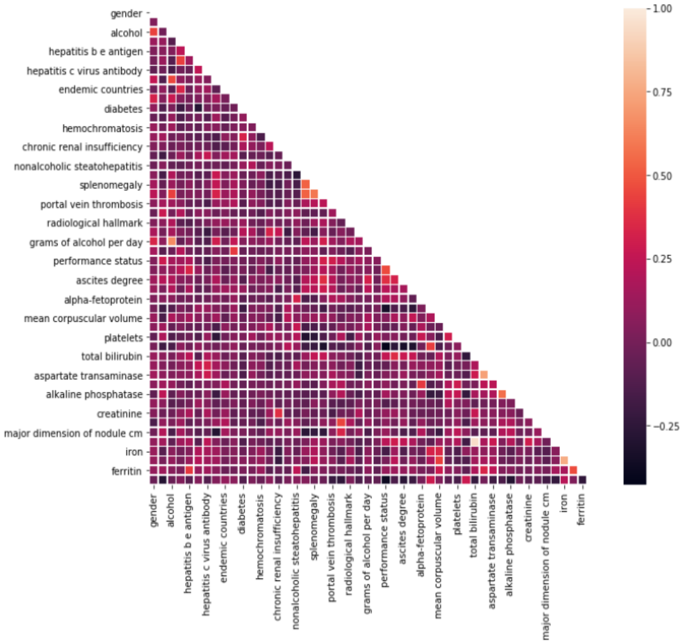


Fig. 1 Correlation matric

IV. Methods

It is significant to compare the performance of different machine learning algorithms consistently. In this section shown a briefly description of each algorithm. For this paper, used the following supervised algorithms:

Random forests or random decision forests makes a collection of decision trees at random chosen subset of training set. After that, aggregates the votes from distinct decision trees to choose the final class of the test object [15]. Random Forests are an improvement over bagged decision trees. It correct for decision trees' habit of over fitting to their training set. AdaBoost Classifier proposed by Freund and Schapire (1996) is another ensemble classifier, which made up of multiple classifier.

Bagging (Bootstrap Aggregating) proposed by Breiman (1996), creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set [14].

AdaBoost works in such a way that in subsequent iterations it trains and then measures the error of all available weak classifiers [16]. In each subsequent iteration, the validity of poorly qualified observations is increased, so that the classifiers pay more attention to them. Predicting with AdaBoost by weighting predictions from weak learners.

Logistic Regression was used in the biological sciences in early twentieth century. The goal of logistic regression is to studies the association between a categorical dependent variable and a set of independent (explanatory) variables. Is used for categorical target variable such as 0 and 1 and usually for the case when the dependent variable has many unique values [12]. Logistic regression allows calculating the probability of a certain event.

Decision trees can handle categorical and numerical data. The decision tree classifiers organized a series in a tree structure. Recursively breaks down a data set into smaller and smaller subsets and expanding the leaf nodes of the tree until the stopping criterion is met. The final result is a tree with decision nodes and leaf nodes[13].

V. Results and Analysis

In this section, presented the results of comparison of the five machine learning algorithms tested in this study. Compared the algorithms mentioned above to find the best performance for the dataset. For this research, used Python with scikit-learn library.

The dataset is randomly divided into subsets. The data set was split into training and test sets. The training sets contained 70% of the whole data. To determine the influence of different data set splits on the methods, each of the five algorithms presented above was run on the same splits of the data set into training and test data. Common problem is that the data is mixed, and contains both numeric and categorial variables. Numeric variables transformed to categorial type.

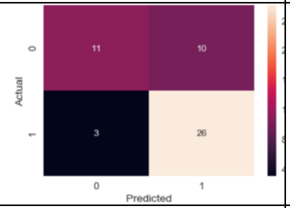
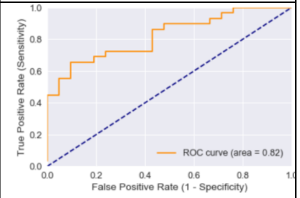
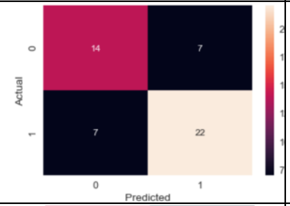
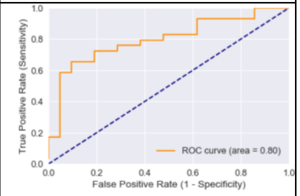
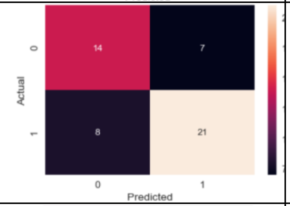
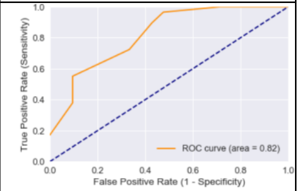
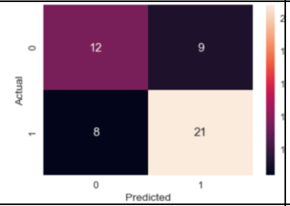
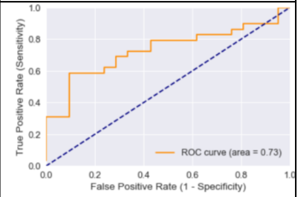
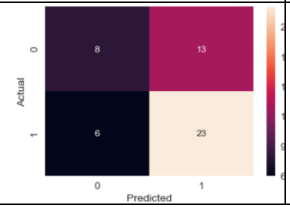
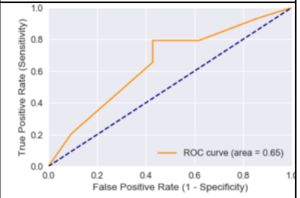
Used a different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies. Confusion matrices of the algorithms following the form of the change-detection error matrix proposed by Macleod and Conglaton (1998) are shown in Table 1, describes the performance of a classification model.

- True Positives (TP): correctly predicted that patient died
- True Negatives (TN): correctly predicted that patient lives
- False Positives (FP): incorrectly predicted that patient died (a "Type I error")
- False Negatives (FN): incorrectly predicted that patient lives (a "Type II error")

The ROC (Receiver Operating Characteristics) curve has been introduced to evaluate ranking performance of machine learning algorithms [17], has been proposed a measure for evaluating the predictive ability of learning algorithms. [18] The ROC curve compares the classifiers' performance across the entire range of class distributions and error costs. Confusion matrices and ROC curves for the best and worst methods are shown in Table 1.

TABLEII

Performance Comparison of namely Random Forests (RF),Boosting (Bo),Bagging (Ba), Logistic Regression (LR), and the Decision Trees (DT) algorithms

	Performance Comparison	
	Confusion Matrix	**ROC Curve**
RF		
Bo		
Ba		
LR		
DT		

A summary of the results of the methods for which analysis can be used is given in Table 2. Classification accuracy is a percentage of the percentage of correctly classified predictions. Classification Error, percentage of errors during classification. False Positive Error means how often is the prediction incorrect.Precision presents the number of class members classified correctly over the total number of instances classified as class members. Recall (or true positive rate) is the number of class members classified correctly over the total number of class members. AUC score is the percentage of the ROC plot that is underneath the curve.

TABLEII

Performance comparison measures of algorithms

	Algorithm				
	RF	**Bo**	**Ba**	**LR**	**DT**
Classification Accuracy	74%	72%	70%	66%	62%
Classification Error	0.26	0.28	0.30	0.34	0.38
False Positive Error	0.48	0.33	0.33	0.43	0.62
Precision	0.72	0.76	0.75	0.70	0.64
Recall	0.74	0.72	0.70	0.66	0.62
AUC Score	0.71	0.71	0.70	0.65	0.59

The studies have shown that a number of different algorithms are able to achieve high classification accuracy. Each model have different performance characteristics. Among the algorithms examined in this study, random forest has proven itself in many aspects evaluated, including prediction accuracy. Finally, used a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one to finalize.

VI. CONCLUSIONS

This paper evaluates performance of the application of supervised machine learning algorithms for prediction mortality in patients with Hepatocellular Carcinoma on dataset collected at a University Hospital in Portugal. We now briefly discuss the results of the different methods on the data sets. Additional analysis was necessary with graphical methods such us histograms.

Five machine learning algorithms were compared in terms of their performance in relation to the problem with supervised classification: Logistic Regression, Decision Trees, Random Forests, Bagging, and the Boosting. It represent five general machine learning strategies for automatic data-based reasoning. Performed multiple comparisons between those. Comparison refer to the combination of accuracy, precision and recall using the term classification accuracy. Results show that the mean error rates of many algorithms are sufficiently similar, however, there are significant differences for accuracy.

In conclusion, of the five methods investigated in this paper, the top three Random Forests, Bagging and AdaBoost give almost identical results. In most of the analyzed situations, the Random Forest proved to be the most appropriate method. Sometimes the best results were generated by the boosting algorithm. However, among the three methods, bagging was the worst in the analysis. Compared to others methods, Random Forests did better in terms of overall accuracy. This work identifies Random Forests as a good first choice inference algorithm for predicting mortality in patients with Hepatocellular Carcinoma. The random forest classification models are in this case straightforward to train, stable with range of model parameters values, and much more accurate than other presented algorithms. Methods for voting classification algorithms, such as Bagging and AdaBoost, have been shown to be very successful in improving the accuracy of certain classifiers for artificial and real-world datasets. The AdaBoost algorithm, generates the classifiers sequentially, while Bagging can generate them in parallel. [19] The logistic regression and decision tree paradigm is not well suited for this problem domain. This research demonstrated the value of comparison statistical methods to analyse and choose machine learning algorithms.It should be noted that the method of operation of a given method depends to a large extent on the parameters of the experiment.

ACKNOWLEDGMENT

I gratefully acknowledge to my thesis supervisorfor valuable guidance.

REFERENCES

i. *JF. Perz, GL. Armstrong, LA. Farrington, YJ. Hutin. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. „J Hepatol". 45 (4), s. 529-38, Oct 2006. DOI: 10.1016/j.jhep.2006.05.013. PMID: 16879891.*

ii. *HB. El-Serag, KL. Rudolph. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. „Gastroenterology". 132 (7), s. 2557-76, Jun 2007. DOI: 10.1053/j.gastro.2007.04.061. PMID: 17570226.*

iii. *AM. Di Bisceglie. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. „J Hepatol". 56 (4), s. 908-43, Apr 2012. DOI: 10.1016/j.jhep.2011.12.001. PMID: 22424438*

iv. *Ramanathan M. Seshadri, SiddeshBesur, David J. Niemeyer. Survival analysis of patients with stage I and II hepatocellular carcinoma after a liver transplantation or liver resection. HPB : the official journal of the International HepatoPancreato Biliary. ISSN: 1477-2574, Vol: 16, Issue: 12, Page: 1102-9, 2014*

v. *Matthew J.Cracknell, Anya M.Reading. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences Volume 63, February 2014, Pages 22-33*

vi. *Kumardeep Chaudhary, Olivier B. Poirion. Deep Learning based multi-omics integration robustly predicts survival in liver cancer. Mar. 8, 2017*

vii. *Miriam SeoaneSantosab, Pedro Henriques Abreu. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of Biomedical Informatics Volume 58, December 2015, Pages 49-59*

viii. *[8]http://archive.ics.uci.edu/ml/datasets/HCC+Survival*

ix. *MałgorzataMisztal. Imputation of missing data using R package. ACTA UNIVERITATIS LODZIENSIS FOLIA OECONOMICA 269, 2012*

x. *Shichao Zhang. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software Volume 85, Issue 11, November 2012, Pages 2541-2552*

xi. *Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.*

xii. *Logistic Regression. NCSS Statistical Software. Chapter 321.https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pdf*

xiii. *Classification and Decision Tree Classifier Introduction. http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html*

xiv. *Christoph Bussler. Artificial Intelligence: Methodology, Systems, and Applications: 11th International Conference, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004*

xv. *Savan Patel. Chapter 5: Random Forest Classifier.https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1*

xvi. *Robert E. Schapire. Explaining AdaBoost. http://rob.schapire.net/papers/explaining-adaboost.pdf*

xvii. *F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining , pages 43–48. AAAI Press, 1997.*

xviii. *Jin Huang, Charles X. Ling. Using AUC and Accuracy in Evaluating Learning Algorithms. December 2, 2003*

xix. *Eric Bauer, RonKohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105–139 (1999)*