Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230960

Heart failure prediction based on multiple machine learning algorithms

Lin Peng

Foshan Foreign Language school, Foshan, Guangdong, 528599, China

1811111120@mail.sit.edu.cn

Abstract. Heart failure is a complex medical condition that arises due to the heart's inability to adequately circulate blood throughout the body, which is challenging to predict. This research aims to investigate three distinct models, namely logistic regression, random forest and decision tree generation algorithms. Logistic regression is essentially a logistic function applied to linear regression, and the loss function associated with linear regression is similar to the mean square error-like loss. In contrast, the loss function for logistic regression follows cross-entropy loss. while cross-entropy loss is often used in practice, it differs from mean square error loss. The derivative of cross-entropy loss is a difference that updates rapidly when the error is significant and slowly when the error is small, which is a desirable trait for the purposes. Decision tree generation algorithms utilize tree structures in which internal nodes represent judgments on attributes, branches represent outputs of judgments, and leaf nodes represent classification results. Random forest is an integrated learning algorithm that employs decision trees as the base learner. In classification models, multiple decision trees are processed for voting, while multiple decision tree results are processed for averaging in regression models. Experimental results indicate that random forest outperforms the other two models, albeit with a marginal difference. Further studies should incorporate additional models to identify a more suitable model for predicting heart failure.

Keywords: heart failure prediction, machine learning, artificial intelligence.

1. Introduction

Heart failure is a heart disease in which the heart cannot pump blood as effectively as it normally does, resulting in the body's need for blood not being met. At present, the number of heart failure patients over the age of 25 in China has reached 12.05 million, with an increase of 2.97 million each year [1]. In developed countries, the prevalence of heart failure is about It is 2% to 3%, and the prevalence rate is higher in the elderly, and it is usually the terminal stage of the development of various heart diseases, such as hypertension, coronary heart disease, myocarditis, heart valve disease, etc., which means, both the prevalence and incidence of heart failure increase with age. Heart failure will affect the function of the whole-body organs: Insufficient renal vascular perfusion can lead to abnormal renal function, and long-term liver congestion and hypoxia can lead to cardiogenic cirrhosis. Early diagnosis, treatment, and management of this disease are therefore imperative. However, the current diagnostic method commonly used in hospitals is manual diagnosis, which is slow and misdiagnosed, and the labor cost is high.

Therefore, it is necessary to find a more scientific, faster and cost-effective method for the prediction of heart failure. Artificial intelligence is a technology that enables computers to perform various advanced

DOI: 10.54254/2755-2721/18/20230960

functions by simulating human intelligence, including the ability to view, understand, translate and analyze data, and make recommendations. Artificial intelligence is fast in calculation, search, statistics, and analysis. Its operation level, storage, and accuracy far exceed that of humans. Therefore, it can be considered to be combined with heart failure prediction to replace humans in predicting heart failure. It may be able to help clinical workers. A useful tool in the fight against heart failure in a variety of patients. Its development history can be traced back to the 1950s. The main stages include 1956-1974-symbolic reasoning stage, 1974-1980-knowledge base stage, 1980 -1987-knowledge representation and reasoning stage, 1987-present - subsymbol equipment phase. Artificial intelligence technology involves a wide range of fields, including machine learning, deep learning, natural language processing, computer vision, speech recognition, etc [2-7]. Until recent years, with the great advancement of computer technology, artificial intelligence has been widely used in the medical field, and its related research and applications include medical image recognition, health monitoring, drug research, medical decision support, etc.

Therefore, the data set based on kaggle in this paper uses deep forest, random forest, and logistic regression model based on heart failure text data to predict heart failure. First, the data set is preprocessed, and the data set is divided into 80% and 20%. out of 80 percent, 20 percent is used to train the model, and 20 percent is used to test the mortality.

2. Methodology

2.1. Dataset preparation

The source of the dataset is collected from the kaggle. The total amount of data is 299 items with one label. The features covered in the dataset include age, gender, the presence of anemia, the presence of smoking, the indicator of creatinine phosphokinase, the number of platelets, the value of serum creating, the follow-up period(day) of the time, the concentration of serum sodium, and the presence of diabetes, the presence of hypertension, and the labeling of whether the patient died (dichotomous). All features were based on the numerical format: the data were preprocessed in such away that 70% of the original data set was divided into a training set and 30% was used as a test set.

2.2. The introduction of the iogistic regression

This study employed three distinct models, namely logistic regression, decision tree, and random forest. Logistic regression is a machine learning technique utilized to tackle the binary classification problem, which involves estimating the likelihood of an event taking place, such as a user purchasing a particular product, a patient experiencing a certain disease, or an advertisement being clicked by a user [8]. It is important to note that the term "likelihood" is employed herein instead of "probability" in a mathematical context. Logistic regression outcomes do not equate to probability values in the mathematical sense, and as such, cannot be directly utilized as such. Instead, they are often combined with other eigenvalues through a weighted sum rather than direct multiplication.

2.3. The introduction of the decision tree

In the realm of decision analysis, the decision tree is a graphical approach employed to assess project risk and feasibility [9]. This technique entails the construction of a decision tree to determine the likelihood of achieving a net present value that is greater than or equal to zero based on the known probability of various potential scenarios. This approach represents an intuitive application of probability analysis and is aptly named the "decision tree" owing to the graphical representation of its branches, which resemble the branches of a tree. In the context of machine learning, the decision tree is a predictive model that establishes a mapping relationship between an object's attributes and its corresponding values. specifically, the decision tree is characterized by a tree structure where each internal node signifies a test on an attribute, each branch represents a corresponding test outcome, and each leaf node corresponds to a category.

DOI: 10.54254/2755-2721/18/20230960

2.4. The introduction of the random forest

Machine learning involves the use of random forest, which is a classifier consisting of multiple decision trees [10]. The output class of the random forest model is determined by the majority class prediction of the individual tree outputs. The algorithm for developing random forests was devised by Leo Breiman and Adele Cutler, and it is a registered trademark. The term "Random Forests" originated from Tin Kam Ho's proposal of random decision forests at Bell Labs in 1995. This methodology blends Breiman's concept of "Bootstrap aggregating" with Ho's "random subspace method" to generate an ensemble of decision trees.

2.5. ImpIementation detaiIs

In this paper, the sklearn framework was employed to first divide the dataset into x andy [11, 12], and then divide them into a training set and a test set and call the model. The training set is used to train the model, and the test set is used to test the model. The evaluation metric contains the accuracy rate, which is the percentage of all data with the correct prediction category.

3. Result and discussion

The Table 1 displays the performance of the training and testing for three distinct models. Logistic regression, a model which utilizes feature weights as its parameters to produce an interpretable probability value for each sample. while it is computationally efficient and requires low storage resources, it struggles with high feature covariance, leading to insufficient weights, convergence issues, and low accuracy, as evidenced by its testing accuracy of 0.7.

Decision trees, on the other hand, are less computationally demanding and can be readily translated into classification rules. They possess a degree of feature selection ability and can handle irrelevant features, but are susceptible to overfitting and disregard attribute interconnections in the dataset.

In comparison, random forests possess the ability to prevent overfitting and exhibit superior accuracy compared to individual algorithms. Additionally, each tree can be generated independently and simultaneously, allowing for simple parallelization, and it possesses some feature selection ability. while it can be overfit on some noisy classification or regression problems, the model's testing accuracy of 0.78 surpasses that of logistic regression and decision trees. Hence, if one were to choose a predictive model, the random forest would be preferable over the aforementioned alternatives.

Table 1. The performance of the model.

	Training accuracy	Testing accuracy
Logistic regression	0.80	0.70
Decision tree	1.00	0.74
Random forest	1.00	0.78

4. Conclusion

This study utilized three predictive models, namely logistic regression, decision tree, and random forest, to forecast the number of deaths in heart failure. The models were trained using a dataset, and their accuracy was assessed by comparing their training and testing values. Logistic regression exhibited the possibility of non-convergence, under-fitting, and limited accuracy, while decision tree was prone to over-fitting and disregarded attribute correlations within the dataset. In contrast, random forest demonstrated the ability to prevent over-fitting and had some feature selection capability, leading to a higher testing value of 0.78, compared to the 0.70 and 0.74 values obtained for the previous two models. Therefore, random forest was prioritized in this model selection.

It is noteworthy that there are multiple algorithms available to address any machine learning problem, and the principle of "no free lunch" in the field of machine learning implies that no one algorithm is universally suitable for all problems. The performance of a machine learning algorithm is heavily dependent on the structure and size of the data, and therefore the most effective means of assessing algorithmic

DOI: 10.54254/2755-2721/18/20230960

performance is to apply it to the data and compare the results. In summary, selecting an appropriate machine learning task entails matching the algorithm to the specific problem and its requirements. However, in many instances, good data is superior to good algorithms, and therefore designing effective features and performing feature engineering is crucial. Understanding the strengths and weaknesses of each machine learning algorithm is vital for selecting the appropriate feature engineering approach for different algorithms.

References

[1] CN healthcare 2022 https://www.cn-healthcare.com/articlewm/20221104/content-1461272.html
[2] He K Gkioxari G Dollar P et al. 2017 Mask r-cnn Proceedings of the IEEE international conference on computer vision 2961-2969
[3] Yu oYang Y Lin Z et al. 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of UsV China Communications 17(3): 46-57
[4] Yuan F Zhang Z Fang Z 2023 An effective CNN and Transformer complementary network for medical image segmentation Pattern Recognition 136: 109228
[5] Maaz M shaker A Cholakkal H et al. 2023 Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications Computer Vision-ECCV 2022 workshops: Tel Aviv Israel Proceedings Part VII Cham: springer Nature switzerland 2023 3-20
[6] Liu J Liu Y Li D et al. 2023 DsDCLA: Driving style detection via hybrid CNN-LsTM with multi- level attention fusion Applied Intelligence 1-18
[7] Chen J Chen x Chen s et al. 2023 shape-Former: Bridging CNN and Transformer via shapeConv for multimodal image matching Information Fusion 91: 445-457
[8] LaValley M P 2008 Logistic regression Circulation 117(18): 2395-2399
[9] Myles A J Feudale R N Liu Y et al. 2004 An introduction to decision tree modeling Journal of Chemometrics: A Journal of the Chemometrics society 18(6): 275-285
[10] Biau G scornet E 2016 A random forest guided tour Test 25: 197-227
[11] Feurer M Eggensperger K Falkner s et al. 2022 Auto-sklearn 2.0: Hands-free automl via meta-learning The Journal of Machine Learning Research 23(1): 11936-11996
[12] Komer B Bergstra J Eliasmith C 2014 Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn ICML workshop on AutoML Austin Tx: Citeseer