Heart Patient Triage Prediction of Clinical Outcomes Using Machine Learning Models

Baohua Jin[1, a], Qinghua He[1, b], Huaiguang Wu[1], Ming Cheng[2] and Pengjie Xie[1]

[1] School of Computer Communication and Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

[2] The First Affifiliated Hospital, Zhengzhou University, Zhengzhou, China

[a]450082368@qq.com, [b]874417718@qq.com

Keywords: Aided diagnostic; random search and grid search of a random forest algorithm; support vector machine grid search algorithm; Machine learning.

Abstract: In the medical field, the analysis of different clinical and pathological data by medical experts is a complicated process. Therefore, it is important to build a framework that can instantly and effectively identify the prevalence of heart disease in thousands of samples. Hence, the paper proposed a new method for diagnosing heart disease. The proposed machine learning method is based on a hybrid optimization algorithm of random search and grid search of a random forest and a parameter optimization algorithm of support vector machine grid search. The model, in predicting the presence or absence of heart disease in patients, has a maximum accuracy of 89%, which not only aided doctors in accurately predicting and diagnosing various diseases, but also helps patients with early diagnosis.

1. Introduction

Heart disease is the number one cause of death worldwide: more people die of heart disease each year than any other cause [1], [2]. At present, heart disease is still a significant disease that causes public health problems. Due to lack of health awareness and poor lifestyles, heart disease patients are still growing rapidly [3], [4]. Although the latest research in the medical field has been able to identify the risk factors that may lead to the development of heart disease, more research is needed to use this knowledge to reduce the incidence of heart disease, which can prevent the death rate caused by the disease and detect the heart early Disease to save the lives of many patients.

Machine learning technology has been widely used in the diagnosis of heart disease and other clinical diagnostic problems to predict and diagnose various diseases with good accuracy. These studies take different approaches to a given problem, Kemal Polat et al. [5] proposed the use of artificial immune recognition system and fuzzy weighted pretreatment for the diagnosis of heart disease, and the proposed method has good robustness. Latha and Subramanian proposed an intelligent cardiac prediction system using CANFIS and genetic algorithm, which has very low mean square error [6]. Amin S U et al. [7] proposed a technique of using genetic algorithm to predict heart disease by using major risk factors, and implemented a hybrid system that USES global optimization of genetic algorithm to initialize neural network weights to make its learning speed faster, more stable and more accurate. Radhimeenakshi S [8] used support vector machine (SVM) and artificial neural network (ANN) to analyze patients with heart disease, and compared the experimental results. Other researchers use data mining tools to analyze large amounts of data that can be obtained from medical diagnosis and extract useful information [9-12]. Sonawane et al. [13] proposed a heart disease prediction system using multi-layer perceptron neural network, and obtained a high degree of accuracy. Ashok et al. [14] proposed the performance evaluation of six different machine learning technologies for the prediction of heart disease, and obtained the highest classification accuracy up to 85%. As stated in many current studies, due to advances in machine learning and information technology, machine learning technology has the prospect of high classification accuracy related to

other data classification processes [15-17]. Achieving significant accuracy in predictions is critical, and doctors can help patients by predicting heart disease before it happens, and early detection of heart disease can save many patients' lives. The goal of this work is to build the workflow of classification methods based on machine learning models to assist doctors in the diagnosis of heart disease, so as to reduce the intensity of doctors' work.

There are many factors that cause disease in patients in the field of medical diagnosis. If doctors only diagnose new unknown cases based on information obtained from clinical data and clinical experience, it will make the doctor's job difficult. In order to make the diagnosis process easier, faster, more accurate and more effective, we proposed a new method for diagnosing heart disease in this study. The method adopts random search and mesh search parameter optimization methods. In the experiment, we adopt three methods (decision tree, random forest and support vector machine). In the random forest algorithm, we first search the range by random search, then search through the grid to find the optimal super-parameter, and then take the optimal super-parameter set in the search results to get the optimal model. In support vector machine (SVM), the kernel function parameters are determined by grid search algorithm, which is decoupled from each other, and the operation efficiency is improved. Through the experiment on the data in the machine learning database, we obtain higher classification accuracy. In our current study, our goal is to highlight comparisons with previous studies, improve algorithms, and use them in combination to be able to select the most appropriate prediction methods.

The remaining of the paper is organized as follows. In Section 2, we briefly introduced the experimental data, feature description, data preprocessing, and experimental methods used in the paper. We will introduce the evaluation criteria and the experimental results in three sections. Finally, we conclude this paper in Section 4 with future directions.

2. Methods

In this section, we first explain the heart disease database we used in our experiments. We then present the performance evaluation methods used to evaluate the proposed method. Finally, we give the experimental results and discuss our observations from the obtained results.

2.1 Hospital dataset

The database we used in our experiments are from a dataset published on kaggle, which was provided by the Cleveland Clinic Foundation. This database is part of the collection of databases at the University of California, Irvine collected by David Aha.The purpose of this data set is to determine whether a patient has a heart disease based on the results of various medical examinations performed on the patient.There are two classes: presence and absence (of heart-disease). This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Attributes of symptoms that are obtained from patient are listed in the next section.

2.2 Features description

We used 303 examples from this database, each containing 14 attributes, the selected properties include Age, Sex, CP, TRESTBPS, CHOL, FPS, RESTECH, THALACH, EXANG, OLDPEAK, SLOPE, CA, THAL, TARGET. Their attribute characteristics are described in table 1. This paper USES 13 input attributes to predict heart disease. In order to obtain more appropriate results, we preprocessed the input data and compared them with different algorithms.

Table 1. Clinical features and their description

Number	Features	Description	Values
1	Age	age in years	Continuous
2	Sex	Male or female	1 = male; 0 = female
3	CP	chest pain type	Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value4:asymptomatic
4	TRESTBPS	resting blood pressure (in mm Hg on admission to the hospital)	Continuous
5	CHOL	serum cholestoral in mg/dl	Continuous
6	FPS	fasting blood sugar > 120 mg/dl	1 = true; 0 = false
7	RESTECH	Resting electrocardiographic results	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8	THALACH	maximum heart rate achieved	Continuous
9	EXANG	Exercise induced angina	1 = yes; 0 = no
10	OLDPEAK	ST depression induced by exercise relative to rest)	Continuous
11	SLOPE	the slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 =downsloping
12	CA	number of major vessels	0-3
13	THAL	A blood disorder called thalassemia	3 = normal; 6 = fixed defect; 7 = reversable defect
14	TARGET (1 or 0)	TARGET (1 or 0)	0 = no, 1 = yes

2.3 Data Processing

Data preprocessing can improve the quality of data, thus contributing to improving the accuracy and performance of the model. In this paper, data are converted by extracting missing fields, extracting outliers and normalizing data, and missing attributes are processed by input means.

2.4 Training and Testing the Models

Classification algorithms play an important role in predicting heart disease. In this article, we have analyzed several different classification algorithms. The algorithm includes decision tree, random forest and support vector machine.

1) Decision Tree

A Decision Tree is a flow chart- like structure that includes a root node, branches, and leaf nodes. The dataset attributes are defined through the internal nodes. A branch is the result of each test for each node. The algorithm of decision tree learning is usually a process of recursively selecting the optimal feature， and then segmenting the training data according to the feature, so that each sub-data set has a best classification process. This process corresponds to the division of feature space and the construction of decision trees. The construction process is shown in Figure 1.

step 1:Start: construct the root node, place all the training data on the root node, choose an optimal feature, and divide the training data set into subsets according to this feature, so that each subset has the best classification under the current conditions .

step 2:If these subsets can be basically classified correctly, then construct leaf nodes, and divide these subsets into corresponding leaf nodes.

step 3: If there are other subsets that cannot be classified correctly, then select new optimal features for these subsets, continue to segment them, and construct corresponding nodes. If recursive, until all training data subsets are basically Correct classification, or no suitable features.

step 4: Each subset is divided into leaf nodes, that is, there are explicit classes, so a decision tree is generated

It is a popular classifier because it is simple, fast, and easy to interpret, explain and implement. It requires no domain knowledge or parameter setting.

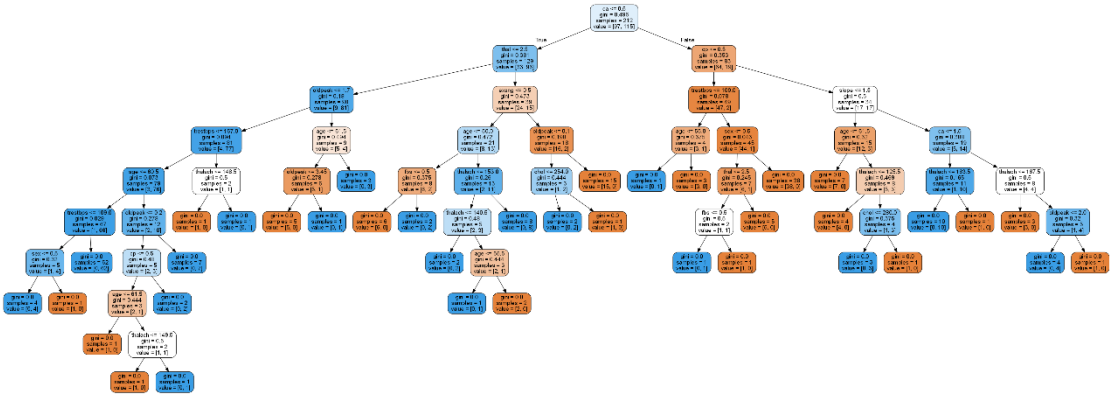

Fig 1: The decision tree construction process of this experiment

2) Optimization parameters of random forest algorithm

Random forest is an effective ensemble learning algorithm, which is widely used in the field of medical health, not only to predict the category of disease, but also to predict the risk of disease. However, the random forest algorithm uses the majority voting rule and cannot distinguish between strong and weak classifiers. In order to obtain higher prediction accuracy, parameters need to be optimized. In this paper, the algorithm combining random search and grid search is used to optimize the super parameters in the application scenarios of random forest with large data volume and super parameters, and the optimal value of super parameters is obtained. Traditional grid search will exhaust all the parameter combinations one by one, resulting in better model effect but longer time. Random search results are slightly worse than the former, but it takes less time. Therefore, we first search the range with random search, then search through the grid to find the optimal super-parameter, and take the optimal super-parameter set in the search results.

3) Grid Search of Support Vector Machine

Support vector machine is a supervised machine learning technique for classification and regression. When using SVM, first transforming data into high dimensional space may convert complex classification problems into simpler problems that can use linear discriminant functions. Secondly, SVM provides the most useful information for classification, finds the optimal classification hyperplane, and classifies unknown data. However, in practice, the classification performance of SVM is closely related to the selection of its parameters. In order to improve the classification accuracy of support vector machines, a grid search algorithm is used to determine the kernel function parameters to decouple them from each other and improve the operating efficiency. Here, SVM was used for binary classification having two categories Absence and Presence of heart disease.

3. Results and Discussion

In this section, we first explain the classification performance evaluation method, then give the experimental results, and discuss the experimental results.

3.1 Performance Evaluation Methods

In this paper, we apply 3 supervised machine learning techniques to classify heart disease samples. Out of total records 70% records are used for training and testing is done by using remaining 30% records. This experiment uses a confusion matrix to evaluate the classification model. Samples with presence of heart disease were considered as positive class, and samples with absence of heart disease were considered as negative class. Here, where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively. we use the following expressions.

True positive (TP): TP is the number of true positives, number of samples with presence of heart disease predicted as presence of heart disease.

True negative (TN): TN is the number of true negatives, number of samples with absence of heart disease predicted as absence of heart disease.

False positive (FP): FP is the number of false positives, number of samples with absence of heart disease predicted as presence of heart disease

False negative (FN): FN is the number of false negatives, number of samples with presence of heart disease predicted as absence of heart disease

Accuracy: Correct predictions as a percentage of total sample.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Precision: It refers to the prediction result, and its meaning is the probability of actually being a positive sample among all the samples predicted to be positive.

Precision=TP / (TP+FP)

Recall: It is for the original sample, and its meaning is the probability of being predicted as a positive sample in the actual positive sample.

Recall=TP / (TP+FN)

F1: The Measure of F1 is the harmonic average of accuracy and recall.

F1= (2*Precision*Recall) / (Precision + Recall)

3.2 Classification Experimental results

This section presents the experimental results of 5 methods (DT, RF, Random-grid Forest Search, SVM, Grid-SVM) used in the article to identify healthy and one patient with heart disease. To evaluate the effectiveness of our method, in this study, the confusion matrix is used to display the classification results of the system. Compare our results with those of traditional methods, Table 2 shows the classification accuracy of our method and previous methods. From these results, we can see that the method with improved parameters obtained the highest classification accuracy. The accuracy rate of traditional random forest is 79%, and the accuracy rate after changing parameters is 83%, the accuracy of the traditional SVM algorithm is only 56%, resulting in over-fitting. After changing the parameters, the over-fitting is avoided, and the accuracy is 89%.

From the above results, we conclude that the combination of random search and grid search parameter adjustment algorithm has achieved promising results in classifying the possible heart disease patients. We believe that the proposed system can be very helpful to the doctor's final decision on the patient, assist the doctor's diagnosis, and shorten the patient waiting time.

Table2: Classification performance measure indices for using machine learning techniques

Algorithm	Label	Accuracy	Precision	Recall	F1-Measure
Decision Tree	0	0.71	0.70	0.78	0.74
	1		0.80	0.72	0.76
Random Forest	0	0.79	0.82	0.78	0.80
	1		0.83	0.86	0.84
Random-grid Forest Search	0	0.83	0.89	0.76	0.82
	1		0.82	0.92	0.87
SVM	0	0.56	1.0	0.02	0.05
	1		0.56	1.00	0.71
Grid-SVM	0	0.89	0.85	0.80	0.83
	1		0.85	0.88	0.86

4. Conclusions and future work

There are different machine learning techniques that can be used for the identification and prevention of heart disease among patients. In this article, three different classification algorithms are used to predict heart disease. They are tuning algorithms for decision trees, random forest random search and grid search, and grid search algorithms for support vector machines. By analyzing the experimental results, we find that the grid search classification algorithm of support vector machines is the best classifier for predicting heart disease, and has the best performance in assisting the diagnosis of heart disease. Because it has higher accuracy and smaller error rate. In the future, we tend to improve performance efficiency by applying other deep learning techniques and optimization techniques. It also enhances auxiliary results by adding other attributes of the heart disease dataset.

References

[1] “Global atlas on cardiovascular disease prevention and control”, WHO,2011

[2] Dangare, Chaitrali & Apte, Sulabha. (2012). A Data Mining Approach for Prediction of Heart Disease Using Neural Networks. 3.

[3] World Health Organization. World Health statistics Annual,Geneva, Switzerland: World Health Organization (2006)

[4] Yanwei X, Wang J, Zhao Z, GaoY. Combination data mining models with new medical data to predict outcome of coronary heart disease. Proceedings International Conference on

[5] Convergence Information Technology; 2007. p. 868–72.

[6] Polat K, Salih Güne?, Sülayman Tosun. Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre- processing [J]. Pattern Recognition, 2006, 39(11):2186-2193.

[7] Latha Parthiban and R.Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, International Journal of Biological and Life Sciences, Vol.3, No.3, pp.157-160, 2007.

[8] Amin S U, Agarwal K, Beg R. Genetic neural network based data mining in prediction of heart disease using risk factors[C]// Information & Communication Technologies (ICT), 2013 IEEE Conference on. IEEE, 2013.

[9] Radhimeenakshi S. Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network[C]// International Conference on Computing for Sustainable Global Development. IEEE, 2016.

[10] N. K. S. Banu and S. Swamy, "Prediction of heart disease at early stage using data mining and big data analytics: A survey," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, 2016, pp. 256-261.

[11] Sultana M, Haider A, Uddin M S. Analysis of data mining techniques for heart disease prediction[C]// International Conference on Electrical Engineering & Information Communication Technology. IEEE, 2017.

[12] Bharti S, Singh S N . Analytical study of heart disease prediction comparing with different algorithms[C]// International Conference on Computing. IEEE, 2015.

[13] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng and E. J. Lin, "HDPS: Heart disease prediction system," 2011 Computing in Cardiology, Hangzhou, 2011, pp. 557-560.

[14] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, 2014, pp. 1-6.

[15] Ashok Kumar Dwivedi. Performance evaluation of different machine learning techniques for prediction of heart disease Neural Computing and Applications, 2018, Volume 29, Number 10, Page 685 Ashok Kumar Dwivedi

[16] H. Bouali and J. Akaichi, "Comparative Study of Different Classification Techniques: Heart Disease Use Case," 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, 2014, pp. 482-486.

[17] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207.

[18] M. A. jabbar, P. Chandra and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, 2012, pp. 628-634.