EasyChair Preprint

№ 3046

Heart Disease Prediction Model Based on Model Ensemble

Xu Wenxin

March 26, 2020

Heart Disease Prediction Model Based on Model Ensemble

Xu Wenxin
Automation Department
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: xwx34@sina.com

***Abstract*—With the increase of heart disease patients in the current society, early prevention of the heart health has become the top priority. The paper proposed a new heart disease prediction method based on model ensemble, which combined three independent models including SVM, decision tree and ANN. As for the data source, the public dataset published by UCI has been preprocessed and used into the combined model and three individual models. The prediction effect has been evaluated by accuracy, precision, recall, and f1 score indicators. As the result shown, compared with three independent models, the ensemble model has a better performance generally indicating a practical medical use.**

Keywords-heart disease prediction, model ensemble, decision tree

I. INTRODUCTION

With the rapid development of economy in current society, people's diet structure has been changed gradually. They are having a bad habit with lack of appropriate exercising, resulting in an unoptimistic health situation. According to a report called Chinese cardiovascular health published by the American Society of Cardiology, which is an authoritative medical journal, about three-quarters of Chinese have poor cardiovascular health, which is not the first time to be warned about the health problems of Chinese people[1].

The report is called cardiovascular health of Chinese adults, based on the cardiovascular data of 96000 Chinese people over the age of 20[2]. Through the observation of behavior health, body mass index, exercise frequency and diet health, and guided by the results of physical health indicators such as blood fat, blood pressure and blood sugar, the result shows that only 13.5% of Chinese people meet the health standard. Therefore, prevention and control of heart disease has gradually become a public concern. However, in the whole world, nearly one-third of the deaths are caused by heart disease, while the number of deaths due to heart disease in China has reached several hundred thousand[3].

With the help of data mining technology, such as extracting the index data of human health and analyzing the impact of different characteristics on heart disease, we can effectively prevent the occurrence of heart disease, so that many patients get proper treatments in time[4].

II. CLASSIFICATION PREDICTION ALGORITHM

A. Decsion Tree

Decision tree is based on information entropy, and it is one of commonly used algorithms nowadays. The principle is to find an optimal feature by building and pruning trees. When constructing the decision tree, the information gain corresponding to each node is selected to construct the decision tree recursively[5]. Firstly, the information gain of all possible features is calculated from the root node, and the feature with the maximum information gain is selected as the node feature[6]. The final decision tree model can be obtained when the information gain of all features is very small and no feature can be selected. The generated decision tree often classifies the training data very accurate, but the classification of the unknown test data is not, which leads to an over fitting phenomenon. Therefore, it is necessary to simplify the decision tree generated and carry out paper-cut processing[7]. Some subtrees are cut from the generated tree, and the root node is regarded as a new leaf node, in order to simplify the classification tree model and minimize the loss function of the whole decision tree.

B. Support Vector Machine

Support vector machines is a binary classification model, which is the linear classifier with the largest interval in the feature space[8]. Its learning strategy is interval maximization, which can be regarded as a problem of solving convex quadratic problem, and also equivalent to the problem of minimizing regularized loss function. The basic idea of SVM is to solve the separation hyperplane which can correctly divide the training data set and has the largest geometric interval. For the linear separable data set, there are infinite hyperplanes, but the separation hyperplane with the largest geometric interval is unique.

Given a series of data samples, each sample has a corresponding label, which can be shown as the following two-dimensional panel in Fig. 1.

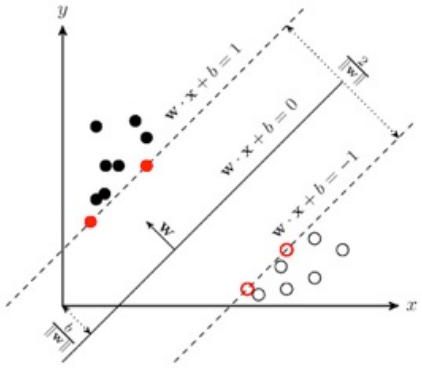

Figure 1. Example of two-dimensional panel.

To be noticed, x_i represents to the feature vector of number i, and y_i represents type notification. It is considered as positive class when it equals to one, and it is considered as negative class when it equals to minus one.

There are two different types of data, which can be represented by solid points and hollow points. A straight line can be discovered to separate the two types of data. The higher probability of data falling into the side of particular category, the more accurate of the final prediction can be. The equation of a hyperplane can be written in the following form[9].

$$w*x + b = 0 \tag{1}$$

When the hyperplane is determined, the support vectors can be conducted and the margin interval can be calculated. Each hyperplane corresponds to a margin, and the goal is aiming at find the hyperplane corresponding to the largest value in all margins. Therefore, the objective function of the optimization problem can be written as following form. Then the classification of the data can be determined by calculating the distance between the data points and the hyperplane.

$$\arg\max_{w,b}\{\min(\ y(w^T x+b))\cdot\frac{1}{||w||}\} \tag{2}$$

C. *Artificial Neural Network*

Artificial neural network is based on the basic principle of neural network in biology. After understanding and abstracting the structure of human brain and the response mechanism of external stimulus, it simulates the neural system of human brain to process complex information based on the theory of network topology[10]. Its intelligent adaptive learning ability and the ability to carry out complex logical operations have attracted the attention of various scientific fields.

Neural network is a kind of operation model, which consists of a large number of neuron nodes connected with each other. Each node represents a specific output function, which is called activation function. The connection between two nodes represents a weighted value for the signal passing through the connection[11]. The output of the network depends on the structure, connection mode, weight and activation function of the network. Artificial neural network combines the knowledge of biological neural network with mathematical statistical model, so that the neural network can have decision-making ability and simple judgment ability similar to human[12].

D. *Model Ensemble*

The concept of model ensemble is to merge multiple weak models into a strengthened model. Because make use of a single model will be over fitted, and multiple model fusion can improve generalization ability and prediction ability[13]. The key of the model ensemble is to ensure the diversity of weak classifiers. Each weak classifier will have more or more false judgments in the discrimination. However, the fusion of learning algorithms can significantly improve the performance[14] [14].

III. EXPRIMENT

UCI machine learning knowledge base is an open-source database for researches in various fields, which is used for empirical analysis of machine learning algorithms. The paper used the open-source dataset of heart disease provided by Cleveland Clinic Foundation, in order to explore the relationship between physical measurement data and the risk of having heart disease. The explanation and description of some important data items in the dataset is shown in Table I.

TABLE I. DESCRIPTTION OF DATASET

Name	Description	Value Scope
sex	sex type	male and female
age	age number	age value
cp	types of chest pain	typical angina atypical angina non-angina pain non-clinical response
trestbpss	resting blood pressure	blood pressure value
fbs	fasting blood glucose	blood glucose value
exang	exercise induced pain	true and false
slope	ECG slope degree	down flat up
thalach	maximal heart rate	heartbeat value
chol	cholesterol number	cholesterol value

A. *Data Preprocessing*

The procedure of data preprocessing can provide high quality data for subsequent experiments, especially for discontinuous numerical characteristics.

As shown in Table II, it presents a part of the original dataset.

TABLE II. PART OF ORIGINAL DATASET

Age	Sex	Cp	Trestbps	Chol	Fbs	Slope
63	male	angina	145	223	223	down
67	female	asympt	160	286	286	flat
57	male	asympt	120	229	229	flat
37	male	notang	131	254	254	down
41	female	abnang	129	204	204	up

After importing raw data, it appeared that a few fields have abnormal values, therefore they have been filled with mode value. In the original dataset, there are several items of data presented by text description, which is not convenient to process, so it has been transformed into the value of data characteristics. For example, the original value of sex field contains male and female, which can be recorded as one for men and 2 for women. As for the field of chest pain described by cp, the original value of angina can be recorded as one, atypical angina as two, non-angina as three and non-clinical symptoms as four.

Fbs describes the field of fasting blood glucose. It can establish a threshold value and compare the original value with it. Take 120 mg/dl as the boundary point. If it is higher than or equal to the threshold value, it will be recorded as one, and if it is lower than the threshold value, it will be recorded as two. As for the field of ECG slope, the original value of up can be recorded as one, flat as two, and down as three. After the above processing, the data turned to more convenient for experiments, as shown in Table III below.

TABLE III. DATASET AFTER PREPROCSSING

Age	Sex	Cp	Trestbps	Chol	Fbs	Slope
63	1	1	145	223	1	3
67	2	4	160	286	2	2
57	1	4	120	229	2	2
37	1	2	131	254	2	3
41	2	3	129	204	1	1

B. Experiment Steps

First of all, normalize the data of the 13 features in the dataset by StandardScaler method, and ensure the numerical range of each feature is divided into zero to one, which is mainly to operate the comparison and weighting of indicators between different unit level. Since normalization is a simplified calculation method, the dimensional expression is transformed into dimensionless expression and becomes pure quantity. This kind of normalized linear transformation has many good properties, which can improve the data performance and will not change the numerical ordering of the original data.

Install the sklearn dependency library and import pandas and tensorflow. Pandas is a python library that contains many useful utilities for loading and processing structured data. Use pandas to download the dataset from the URL and load it into the data frame. In order to avoid the problem of over fitting and lack fitting due to the limitation of algorithm, apply the cross validation in modeling which requires that the amount of training data should be sufficient, containing at least 50% of the total data. Therefore, the dataset has been split into training data with the proportion of 70% of total data and test data with the proportion of 30% of total data.

Next, wrap the data frame with tf.data, which will enable to use feature columns as a bridge to map from the columns in the pandas data frame to the features used to train the model.

In the process of model training, input training data features and results, while input only features in the testing procedure. As for the ensemble model, its voting mechanism obeys that the minority is subordinate to the majority.

Import VotingClassifier from sklearn.ensemble. Use VotingClassifer method and retrieve three parameters from three independent models' predication results, the ensemble model obtains the final prediction result by figuring out the majority value from three parameters.

Finally, the prediction results obtained by every classifier are compared with the original results, so as to evaluate and analyze the prediction effect of every model.

IV. RESULT ANALYSIS

After building models, it is important to evaluate performance on prediction effect. All the established models will be evaluated from four aspects, including accuracy, precision, recall and F1 score.

The accuracy is obtained by calculating the ratio of the number of samples correctly classified to the total number of samples in a given test data set, which can reflect whether a classifier is effective or not.

However, the measurement of accuracy cannot evaluate the performance of a classifier effectively, because the distribution of samples has a great impact on the evaluation results. For example, if the negative samples account for 95% and the positive samples only account for 5%. Then if the classifier mistakenly classifies all the positive samples as negative samples, it can achieve 95% accuracy, and the classifier with high accuracy does not have high-quality prediction effect.

Therefore, it is necessary to consider the evaluation of recall rate and F1 score for further evaluation, and these two evaluation indexes need to determine four classifications situation first. The task is to find all the patients with heart disease from the people in the area, so the positive category is the patients with heart disease, and the negative category is the healthy.

Based on two dimensions, including prediction and real situation, as shown in Fig. 2 below, the case of true positive is judged to be positive, that is to say, the case of true positive is defined as correctly determined to be a patient with heart disease. In the same way, the case of false positive is defined that the healthy person is wrongly judged as a patient with heart disease. The case of false negative is defined that the patient with heart disease is wrongly judged

as a healthy person. And lastly the case of true negative is defined that the healthy person is correctly judged to be healthy.

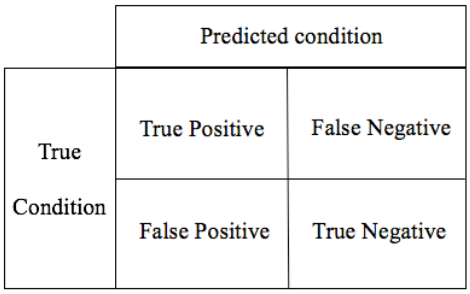

Figure 2. Four kinds of prediction situation .

Accuracy is calculated by the proportion of all correctly retrieved samples to all actually retrieved samples, that is, the proportion of true position to the sum of true position and false position. The calculation of recall rate is obtained by the proportion of all correctly retrieved samples in all samples that should be retrieved, that is, the proportion of true position in the sum of true position and false negative. And F1 score is the harmonic mean of accuracy and recall rate, which is regarded as a comprehensive evaluation index, and has important significance. Table 4 below shows the above performance evaluation indicators of three independent models and ensemble model.

TABLE IV. PERFORMANCE EVALUATION

	Accuracy	Precision	Recall	F1-score
Decsion Tree	0.831	0.787	0.846	0.809
SVM	0.853	0.813	0.912	0.853
ANN	0.861	0.820	0.875	0.851
Ensemble Model	0.873	0.828	0.908	0.871

As it shown vividly, from the accuracy and accuracy dimensions, ensemble model has a better performance, and from the recall dimension, SVM model has a better performance. Since accuracy represents the accuracy of determination, recall represents whether the model has a comprehensive correct determination, and f1-score is a general consideration of accuracy and recall, which has more important guiding significance. Therefore, the ensemble model with f1-score of 0.871 has a better performance than the other three individual classification models.

V. CONCLUSION

The paper proposed a heart disease prediction model based on model ensemble. It integrates decision tree, SVM and artificial neural network through cooperating voting, in order to reduce the influence of single model's misjudgment. Firstly, the original dataset has been preprocessed and normalized. Then the training dataset and test dataset were divided and used in three independent models and ensemble model in modeling project. After that, the prediction performance of them has been evaluated by accuracy, precision, recall and f1-score four indicators. The experiment results have shown that the ensemble model has an improved prediction performance in general. As a result, making use of ensemble model to predict heart disease can provide decision-making reference for doctors to make clinical diagnosis, which has practical clinical significance and application prospect.

REFERENCES

[1] Pedersen T R, Kjekshus J, Berg K, et al. Randomized trial of cholesterol-lowering in 4444 patients with coronary-heart-disease - the scandinavian simvastatin survival study (4s)[J]. 2017, 344(8934):1383-1389.

[2] Middleton E, Kandaswami C, Theoharides T C. The Effects of Plant Flavonoids on Mammalian Cells:Implications for Inflammation, Heart Disease, and Cancer[J]. 2018, 52(4):673-751.

[3] Valery T. Miller, John LaRosa, Vanessa Barnabei, 等. Effects of Estrogen or Estrogen/ Progestin Regimens on Heart Disease Risk Factors in Postmenopausal WomenThe Postmenopausal Estrogen/Progestin Interventions (PEPI) Trial[J]. Jama, 2016, 273(3):199-208.

[4] 林宇, 周慧、刘春霞. 成年人身体素质指数与体质健康指标的关联性研究[J]. 济宁医学院学报, 2018, 35(1):54-56.

[5] 薛薇. 数据挖掘中的决策树技术及其应用[J]. 统计与信息论坛, 2019, 17(2):4-10.

[6] 杨学兵、张俊、YANGXue-bing.等. 决策树算法及其核心技术[J]. 计算机技术与发展, 2017, 17(1):43-45.

[7] 苗夺谦, 王珏. 基于粗糙集的多变量决策树构造方法[J]. 软件学报(6): 2018, 26-32.

[8] Hao Zhang, Alexander C. Berg, Michael Maire, 等. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition[C] IEEE, 2019.

[9] Campbell W M, Sturim D E, Reynolds D A, et al. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation[C]// 2018.

[10] 刘惠文、姚海鹏、张培颖. 基于软件定义网络技术实现人工智能网络体系架构[J]. 信息技术与网络安全, v.37;No.490(02): 2019,11-14.

[11] 陈岭. 大数据时代人工智能在计算机网络技术中的应用[J]. 环球市场, 2017(32):34-34.

[12] 高冲. 浅谈人工智能在计算机网络技术中的应用[J]. 科技视界, 2019(10):210-211.

[13] Haber, Eldad, Holtzman Gazit, Michal. Model Fusion and Joint Inversion[J]. Surveys in Geophysics, 2018(5):675-695.

[14] O. G. Logutov, A. R. Robinson. Multi-model fusion and error parameter estimation[J]. Quarterly Journal of the Royal Meteorological Society, 2019(613):3397-3408.

[15] 许言路、张建森、吉星、et al. 基于多模型融合神经网络的短期负荷预测[J]. 控制工程, 2019(4):619-624.