Predicting Cardiovascular Events by Machine Learning

Caiwei Zhang¹, Junhao Qu², Weicheng Li³, Lehan Zheng⁴

¹Shanghai Starriver Bilingual School, 201108, China
²Qingdao Sixty-seventh Middle School, 266000, China
³Nanjing University of Information Science & Techno, 211800, China
⁴Beijing National Day School, 100039, China

Email of all the authors: zhangcaiweimaggie@163.com; 18765939557@sina.cn; 2313177240@qq.com; 1878971291@qq.com.
admissions@sparkchinaedu.com

Abstract: In the field of medical treatments, machine learning has become an essential technology to mine known data. People always build models to assist doctors to judge. In this paper, 14 features of heart disease patients in two cities, Cleveland and Switzerland, are analyzed by multiple types of neuronal networks and several classifiers after cleaning the data. The model based on the featured data from heart disease patients is designed to predict whether the patients have heart disease or not. Our experimental results on the prediction of cardiovascular events in two cities show that the logistic regression classifier outperforms other methods.

1. INTRODUCTION

1.1 Background and Importance

At present, cardiovascular disease is one of several diseases threatening people's health in the world, which has seriously affected world economic development. Among the top 10 causes of the death claimed by WHO (2017) cardiovascular disease remains one of the top leading causes for deaths. In the era of rapid information development, combining electronic science and technology with medical heart disease is now a hot research issue for researchers. With the continuous efforts of medical researchers in recent years, the research on the diagnosis of heart disease has achieved initial results, and a large amount of data has been accumulated for the reference of researchers.

In [1], the authors have developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using machine learning, namely, Decision Trees, Naive Bayes, and Neural Network. Results show that each technique has its distinctive strength in realizing the objectives of the defined mining goals.

In [2], machine learning is also applied, which encompasses common regression methods such as linear regression, logistic regression, and Cox regression, but also lesser-known models, including penalized regression, principal component regression, cluster analysis. The use of machine learning has seen an uptick because of recent advances in calculation capacity.

In [3], the performance of several data mining techniques is compared, based on accuracy. The results show that the accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. The analysis of this research shows that out of these three classification models Neural Networks predicts Heart disease with the highest accuracy.
In [4], to predict heart diseases, researchers consider a combination of these variables. The accuracy of selecting variables and combining them for evaluation becomes challenging. Data mining and machine learning can be used to identify and predict variables used to assess cardiovascular disease risk.

A great limitation of previous researches is that the capacity of dataset is limited. Most of them use the data from Kaggle or UCI. However, we choose the combination of those several data sources like Cleveland, Switzerland, etc. Therefore, compared to those previous researches, we have much amount of data to analyze. Besides, previous researches have tried several models to predict cardiovascular events like Naïve Bayes, Decision Tree. As a result, the creation of such a model for predicting heart disease has been only partly successful. Therefore, more complex models and multiple geographically diverse data sources are needed to improve the accuracy of predicting the early onset of disease.

We can get more comprehensive results and conclusions by including multi datasets and models. All in all, we focus on neural networks, so we have made some new attempts in this area.

2 METHODOLOGY

2.1 Data Set
The data we used in our study are gained from the Heart Disease Data Set of Center for Machine Learning and Intelligent Systems, UCI Machine Learning Repository [5]. The database contains 14 attributes that were usable to the prediction of cardiovascular events in machine learning.

The two processed datasets we gained are collected from two different cities: Cleveland, and Switzerland.

2.2 Cleaning Data
Since the datasets are incomplete, we need to clean data to continue to do the experiment. The following steps are the data cleaning procedure.

Firstly, we find valid form of the data. Then we input the raw data and reset them into the correct form.

Next, we put in that data and do some analysis. Shown in Fig. 1, we find that if the colors of variables are bright, the variables are more related to the target. We choose the variables whose correlations are greater than 0.5, shown in Fig. 2.
After that, we divide the chosen part into a training set and a testing set, modeling a random forest classifier. We just need to put each column data of those properties into the modeling, and then use the testing set to predict the accuracy.

Finally, we check whether the data is complete and save the completed data.

2.3 Machine Learning
We design our machine learning mainly focusing on two parts: the neuronal network, and the classification methods. We implement three models of the neuronal network, and eight models based on different classification methods. Finally, we compare the accuracy score of each method.

2.3.1 Neuronal networks
In this section, all the implementations are done by TensorFlow. For all measuring methods, we apply the metrics scoring method in sklearn.

A. Artificial Neuronal Network
The artificial neuronal network is with one input layer, two hidden layers, and one output layer.

B. Product-based Neural Networks
Since part of our data are categorical data, a typical representation is to transform those data into a high-dimensional sparse binary feature representation via one-hot encoding. We divide our features into categorical data and quantitative data, and then use the one-hot encoding to transfer all the categorical data. Then we use the neuronal networks to train the data with two hidden layers.

C. Wide & Deep Model
As we have found that the Wide & Deep model is expected to be more effectively compared with wide-only and deep-only models, we apply the algorithms to our data.

The wide component (left part in Figure 1) is a generalized linear model of the form \( y = w^T x + b \) (\( y \): the prediction; \( x = [x_1, x_2, x_3, ..., x_d] \), a vector of \( d \) features; \( w = [w_1, w_2, w_3, ..., w_d] \), the parameters of the model; \( b \): the bias)

The deep model is a feedforward neural network embedding the high-dimensional features into the vector that is low-dimension. The formula is \( a^{(l+1)} = f\left(W^{(l)}a^{(l)} + b^{(l)}\right) \), as \( f \) represent the activation functions.

We use the wide & deep model to train the data in separate two parts and weigh the two results together as the final prediction.

---

1 TensorFlow: https://www.tensorflow.org
2 sklearn: https://scikit-learn.org/stable/index.html
2.3.2 Classifiers
Besides three types of neuronal networks, we also use eight existing classifiers to train the data.

A. K-Nearest Neighbors (KNN)
   Classification is made by measuring the distance between different eigenvalues.

B. Support Vector Machines
   SVM training builds the model by assigning the training samples to one or other categories and makes a non-probabilistic binary linear classifier.

C. Logistic Regression
   Based on the training data, a regression formula is established for classification boundary. According to the regression formula, the training dataset is classified.

D. Decision Tree Classifier
   The algorithm of classification decision tree learning is usually a process of selecting the optimal feature recursively and dividing the training data according to the feature to have the best classification for each sub-data set.

E. Multilayer Perceptron Classifier (MLP)
   MLP builds the model of an artificial neuronal network with one hidden layer, and all multilayer perceptron layers are fully connected to each other.

F. Random Forest Classifier
   Random Forest Classifier constructs a multitude of decision trees while training, and it gives the output that is the mode of the classification of the individual trees.

G. Gradient Boosting Classifier
   Combining the weak learning models in the previous round together, Gradient Boosting Classifier corrects the errors by fitting the negative gradient in the next round of learning.

H. Light GBM Classifier
   It is also a model based on the decision trees. The loss function's negative gradient is used as the residual approximation of the current decision tree to fit the new decision tree.

3 EXPERIMENTAL DESIGN & RESULT

3.1 Experimental Design
In this experiment, we want to choose the best model by comparing the accuracy of 11 models.

Before the experiment, we speculate that artificial neuronal networks will be the best model for this dataset. Moreover, in this dataset, the target is the dependable value, and the other 13 features: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, are undependable values.

First, we divide the dataset into training data and testing data in order to test our models' accuracy. In this part, we split the data into two parts, set the test size as 0.2, and then put the training data in a model. Finally, we use test data to calculate the final results to evaluate the models' performance.

We all use binary cross-entropy for three neuronal networks as the loss function, use the sigmoid activation in the output layer, and use the adam optimizer. In an artificial neuronal network, we establish the sequential function and optimize the network with two hidden layers with the nodes of 80 and 40, respectively. In Product-based Neural Networks, according to the definitions of features, we divide six features (['sex', 'cp', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']) into categorical data and the other into quantitative data, and then use the one-hot encoding to transfer all the categorical data. All the data are then trained by the neuronal network with two hidden layers (with 80 and 40 nodes, respectively). In the Wide & Deep Model, we horizontally separate the data into a wide-only part and deep-only part. Two hidden layers implement the deep-only part with 22 and 11 nodes, respectively, and then we contact the wide and deep part to form the output layer.

In classifier, we put the train data and test data into the existing classifier model and adjust some parameters to get the best outputs.
3.2 Result
Overall, from the three designed neuronal networks, the artificial neuronal network shows the highest test accuracy. (show in below table)

| Table1.overall neuronal network performance |
|-------------------------------------------|
| Artificial Neuronal Network | Product-based Neural Networks | Wide & Deep Model |
| Test accuracy | 86.25% | 84.50% | 61.41% |

For the eight classifiers, we gain the F1 scores of the testing data, and logistic regression shows the best result. (show below in table)

| Table2.overall classifier performance |
|--------------------------------------|
| Classifier | Test Accuracy | F1 score | Precision Score | Recall Score |
| K-neighbors Classifier | 85.94% | 88.75% | 92.21% | 87.65% |
| Logistic Regression | 87.50% | 90.24% | 91.36% | 89.16% |
| SVC | 86.72% | 89.82% | 89.29% | 90.36% |
| Decision Tree Classifier | 71.88% | 76.00% | 85.07% | 68.67% |
| MLP Classifier | 62.50% | 76.92% | 64.00% | 96.39% |
| Random Forest Classifier | 86.72% | 89.57% | 91.25% | 87.95% |
| LGB Classifier | 79.69% | 83.54% | 88.00% | 79.52% |
| Gradient Boosting Classifier | 83.59% | 86.42% | 89.74% | 84.34% |

In table 2, we can observe all of the classifiers’ performances, so we can gain the statement that K-neighbors Classifier, Logistic Regression, SVC, Random Forest Classifiers are more accurate than others, but they have no significant different accuracy. Therefore, we cannot choose an ultimately better model from these four classifiers. Nevertheless, we can know that these classifiers sustain a better performance in these datasets.

4. CONCLUSION
Machine learning is an important data mining technology and has been greatly developed in the medical field. Using artificial intelligence to assist the diagnosis and prediction of heart disease has become an important topic. This study tests the accuracy of three neural networks and eight classifiers, respectively, to find a more effective model to assist the diagnosis of heart disease. Our research results suggest that more accurate medical diagnoses can be obtained by using the Artificial Neuronal Network in the Neuronal Network and the Logistic Regression in the classification methods. In previous researches on UCI heart disease data set, people mainly used the decision tree model, random forest model, and perceptron model, and theoretically analyzed the applicability of each neural network. Our experiment finds that the Logistic Regression has higher accuracy than the random forest model, which was considered to have the highest accuracy in the previous works. However, there are some defects in this study. Firstly, clean data processing was carried out on the part of the experimental data, and secondly, the total amount of data was not sufficient. In future experiments, professional institutions can be established to collect experimental data, and more models need to be tried to further improve the accuracy of predicting cardiovascular events.

REFERENCE
[1] Ordonez, C., 2020. Association Rule Discovery With The Train And Test Approach For Heart Disease Prediction. https://www.ncbi.nlm.nih.gov/pubmed/16617622.
[2] Quesada, J., Lopez-Pineda, A., Gil-Guillén, V., Durazo-Arvizu, R., Orozco-Beltrán, D., López-Domenech, A. and Carratalá-Munuera, C., 2020. Machine Learning To Predict
Cardiovascular Risk.

[3] S. Dangare, C. and S. Apte, S., 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications*, 47(10), pp. 44-48.

[4] 2020. A Review of Data Mining & Machine Learning approaches for identifying Risk Factor contributing to likelihood of Cardiovascular Diseases.

[5] Archive.ics.uci.edu. 2020. *UCI Machine Learning Repository: Heart Disease Data Set*. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/heart+disease> [Accessed 27 August 2020].

[6] Mitchell, T., 2017. *Machine Learning*. New York: McGraw Hill.

[7] Qu, Y., Cai, H., Ren, K., Zhang, W. and Yu, Y., 2020. Product-based Neural Networks for User Response Prediction. [online] 6 pages, 5 figures. Available at: <https://arxiv.org/pdf/1611.00144.pdf> [Accessed 27 August 2020].

[8] Cheng, H., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X. and Shah, H., 2016. *Wide & Deep Learning For Recommender Systems*. [online] Dl.acm.org. Available at: <https://dl.acm.org/doi/epdf/10.1145/2988450.2988454> [Accessed 27 August 2020].