Cardiology Prediction Based on Machine Learning

Yu Shi*, and Qiuli Qin
Economics and Management College, Beijing Jiaotong University, Beijing, China

Abstract. Heart disease is an important disease that endangers human health, with a high mortality rate. Machine learning assisted diagnosis of medical data is a hot topic, and it has made great contributions in predicting patient outcomes and reducing mortality. Therefore, based on the heart disease index data, this paper uses Decision tree model, Clustering model, and Naive Bayes model to predict whether or not having heart disease. The results show that the Naive Bayes algorithm has better prediction accuracy and can assist doctors in diagnosis and treatment.

1 Introduction

The heart is the most important organ in the human body. At present, heart disease has become an important disease that endangers human health. According to "China's Cardiovascular Disease Report 2018" [1], the number of patients with cardiovascular disease in China is currently about 290 million, with the highest mortality rate, accounting for more than 40% of residents' deaths due to disease. The key to the prevention and treatment of the heart is "early", but the causes of cardiovascular diseases are quite complicated, and the detection methods will also bring some pain to the patients, so the prediction and prevention of heart diseases are very important.

Medical data includes physical examination information of many patients during the diagnosis and treatment process. These data objectively reflect a person's physical status. If these data can be used to make predictions and analyse the impact of different indicators on heart disease, it will play a role in preventing heart disease.

With the development of information technology, the use of machine learning for assisted diagnosis [2] has become an important research direction, such as logistic regression [3], neural network [4], Decision tree [5], Naive Bayes [6], etc. It is widely used in disease prediction. Based on the heart disease index data, this paper uses three machine learning algorithms, decision tree, clustering and naive Bayes, to predict whether there may be heart disease, and compare and analyse the prediction results.

2 Theoretical basis

2.1 Decision tree model

The decision tree was first proposed by Hunt et al. In 1966. It is a supervised learning method for classification. On the basis of known probability, the data entropy value is
calculated first, that is, the degree of disorder of the data, and then selected standard, in machine learning, a decision tree is a predictive model that represents a mapping relationship between object attributes and object values, generally using algorithms ID3, C4.5, and CART. The essence of the ID3 algorithm is to find the attribute with the largest information gain. The operation process of the C4.5 algorithm is to select the features with higher information gain than the average value, and then select the features with high information gain. The CART algorithm is to calculate the Gini coefficient.

The formulas for entropy, information gain and Gini coefficient are as follows:

\[
H(Y) = - \sum_{i=1}^{n} P(y_i) \log P(y_i)
\]  
(1)

\[
Gain(X,Y) = H(Y) - H(Y|X)
\]  
(2)

\[
Gini(T) = - \sum_{i=1}^{n} \left( \frac{Tc_i}{T} \right)^2
\]  
(3)

The computational complexity of the decision tree is not high, the output result is easy to understand, and it is not sensitive to missing values. It can handle unrelated feature data, but it may also cause over-fitting problems.

2.2 Clustering model

The clustering algorithm used in this paper is the K-nearest neighbour algorithm, which means that the training set is first divided into different clusters. When a new sample needs to be classified, the k measurement samples closest to the sample are found based on the distance measurement method, and based on the k training sample predicts the category of the new sample, and selects the category with the largest number of k samples as the classification result of the new sample. The Euclidean distance is generally used to calculate the distance. The formula is as follows:

\[
d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}
\]  
(4)

Clustering is not sensitive to outliers. When new data is added, it is not necessary to retrain the model, but the model calculation is large, the calculation complexity is high, and the deviation of the prediction will increase when the samples are not balanced. At the same time, the model has a black box effect and cannot explain the inherent meaning of the data.

2.3 Naive Bayes Model

Naive Bayes is a classification method based on Bayes' theorem and the independent assumption of feature conditions. Assuming that the feature attributes are mutually independent, under the premise of known probability, through a given training set, the joint probability distribution is calculated, and then based on the learned model, input x to find the output c with the largest posterior probability. Calculated as follows:

\[
P(c|x) = \frac{P(c)p(x|c)}{P(x)}
\]  
(5)

Naive Bayes has a stable classification efficiency, which is still effective when there is little data, and can handle multi-class problems. However, the algorithm has independent
requirements for the data, and the classification effect is not ideal when the data is highly related.

3 Model experiment

3.1 Data sources

The data source of this article is the UCI open source data set heart disease. The data set includes a total of 303 data with 15 fields, including basic patient information, patient diagnosis and treatment information, and target fields. The data is highly complete and the fields have clear meanings. The specific explanation of each data field is shown in Table 1:

| Field Name | Chinese name                      | Type of data | Value                 |
|------------|-----------------------------------|--------------|-----------------------|
| age        | Age                               | Continuous   | Age value             |
| sex        | Gender                            | Discrete     | 0,1                   |
| cp         | Chest pain type                   | Discrete     | 0,1,2,3               |
| trestbps   | Resting blood pressure            | Continuous   | Resting blood pressure |
| chol       | Serum cholestasis value           | Continuous   | Serum cholestasis value |
| fbs        | Fasting blood glucose             | Continuous   | 0,1                   |
| restecg    | Resting ECG results               | Discrete     | 0,1,2                 |
| thalach    | Maximum heart rate                | Continuous   | Heart rate            |
| exang      | Exercise-induced angina           | Discrete     | 0,1                   |
| oldpeak    | ST segment pressure value during exercise relative rest | Continuous | Low pressure value |
| slope      | The slope of the ST segment of the ECG | Discrete | 0,1,2                 |
| ca         | Number of blood vessels detected by fluoroscopy | Discrete | 0,1,2,3,4             |
| thal       | Types of defects                  | Discrete     | 0,1,2,3               |
| target     | Target                            | Discrete     | 0,1                   |

3.2 Data pre-processing

The data set includes continuous data and discrete data. First, the data is pre-processed, including processing missing values, outliers, and data standardization. In addition, the data set lacks a primary key column that can uniquely identify each piece of data, so add an id column as the primary key in the data set, and assign values to consecutive integers starting at 1.

The data is divided into a training set and a test set to evaluate the accuracy of the prediction model. 80% of the source data is randomly selected for training the model, and the other 20% of the source data is used for testing the model. According to the analysis of data, it is a classification problem. This article uses three classification models: decision tree model, cluster analysis model, and Native Bayes model.
3.3 Evaluation index

In order to evaluate the application effect of the algorithm, this paper uses Accuracy, Precision, and Recall to evaluate the model prediction results [10]. Accuracy: the ratio of the number of correctly classified samples to all samples. Precision: the ratio of the number of samples that are actually classified as positive and whose prediction result is positive to the number of samples whose prediction result is positive. Recall: The ratio of the number of samples that are actually classified as positive and the prediction result is positive to the number of samples that are actually classified as positive. The confusion matrix of the prediction result is shown in Table 2:

| Actual | Actual-1 |
|--------|----------|
| Prediction-0 | TN | FN |
| Prediction-1 | FP | TP |

The formulas for accuracy, precision and recall are:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{7}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

3.4 Comparative analysis of results

Three models were used for fitting, and the scores of the models were 0.90 for decision trees, 0.91 for cluster analysis, and 0.92 for Native Bayes. Using three models to predict the test set separately, the prediction results of the Decision tree model, Cluster analysis model, and Native Bayes model are shown in Table 3. From the table, it can be seen that the Native Bayes model has the best comprehensive prediction effect. Among them, Native Bayes has the highest accuracy and precision, and the Cluster model has the highest recall.

| Model       | Accuracy | Precision | Recall |
|-------------|----------|-----------|--------|
| Decision Tree | 0.80     | 0.85      | 0.78   |
| Clustering  | 0.80     | 0.80      | 0.86   |
| Native Bays | 0.82     | 0.87      | 0.80   |

Although the prediction effect of the decision tree is not good, the index relationship can be obtained through the decision tree model results and the classification path. The node indexes from near to far from the root node are cp; ca, sex; thal, slope. The closer the distance to the root node is, the higher the correlation between the index and the predicted column. From the prediction results of the Native Bayes model, the correlations of the six fields sex, exang, ca, cp, slope, and thal are stronger than other fields. From the comparison of the results of the two models, the five indicators of cp, ca, sex, thal, and slope are important features of the prediction column.
4 Conclusion

This article is based on the UCI open source data set heart disease, using three models: Decision tree, Cluster, and Native Bayes to predict whether or not you have heart disease, and the prediction results of the three models are analysed to compare the accuracy of the three models for the prediction results accuracy, precision, and recall rate. The experimental results show that the Native Bayes model has higher accuracy in heart disease prediction in the heart disease data set, and machine learning has a good performance in disease prediction.

References

1. Hu Shengshou, Gao Runlin, Liu Lisheng, Zhu Manlu, Wang Wen, Wang Yongjun, Wu Zhaosu, Li Huijun, Gu Dongfeng, Yang Yuejin, Zheng Zhe, Chen Weiwei. Summary of "China Cardiovascular Disease Report 2018" [J]. Chinese Journal of Circulation, 2019, 34 (03): 209-220.
2. Liu Yu, Qiao Mu. Heart disease prediction based on clustering and XGboost algorithm [J]. Computer System Application, 2019, 28 (01): 228-232.
3. Wang Zhe, Wang Kai, Yang Ridong, Zhou Yi. Research on prognostic factors of breast cancer based on machine learning methods [J]. China Digital Medicine, 2019, 14 (01): 18-20 + 46.
4. Li Xiaomin. Research and implementation of cardiovascular disease prediction system based on BP neural network [D]. Ningxia University, 2018.
5. Fan Wenjing, Lou Dongjie, Lu Xin, Xian Yu Shuming. Research on predicting the risk of nosocomial infection in elderly patients in neurosurgery based on decision tree method and Logistic regression [J/OL]. Chinese Journal of Hospital Infection, 2020 (06): 878-883 [2020-04-12]. http://kns.cnki.net/kcms/detail/11.3456.r.20200327.0920.032.html.
6. Chen Zhiyi. Research on prediction of liver cirrhosis based on machine learning [D]. Hunan University, 2018.
7. Huang Xiuxia. Optimization and application of C4.5 decision tree algorithm [D]. Jiangnan University, 2017.
8. Wang Haorui. Cardiac online auxiliary diagnosis system based on k-nearest neighbor algorithm [J]. Electronic Production, 2018 (20): 49-51 + 76.
9. He Zhongyan. Research and development of cardio-cerebrovascular disease diagnosis system based on NB and IGA algorithm [D]. Zhejiang University of Technology, 2019.
10. Li Xiaoqian. Research on heart disease prediction method based on convolutional neural network [D]. Northeast Forestry University, 2019.