ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.173

Heart Disease Prediction Model Based on Logistic Regression Theory

Bai Wei [1+] and Teng Jinbao [2]

[1-2] Xi'an University of Posts and Telecommunications

Abstract. The establishment of accurate and effective disease prediction model is of great practical significance to the medical community. In this context, this paper proposes a heart disease prediction model based on logistic regression theory. Firstly, the acquired data sets are transformed according to the characteristics of logical regression theory; secondly, the data are normalized to eliminate the possible influence of different dimensions on the simulation results; finally, the simulation experiments are carried out based on UCI data set to analyze the characteristics that affect the prediction accuracy of the model. The experimental results show that the accuracy and recall rate of the model can reach 82.33% and 95.83% respectively, which can be used to assist doctors in diagnosis and treatment of heart disease.

Keywords: predictive model; logistic regression; classification; heart disease

1. Introduction

Today, about 18 million people die of heart disease every year on earth. In China, the number of deaths due to heart disease has increased by 90% compared with previous years. With the increasing number of heart disease, doctors often judge the condition of heart patients based on the past experience. In order to reduce the risk of misjudgment caused by doctors' lack of experience, the classification algorithm in machine learning can be applied to the auxiliary diagnosis of diseases.

Aiming at the problem of heart disease diagnosis and prediction, a heart disease analysis method for universal health monitoring was proposed in reference, and Bayesian algorithm was used to analyze and model the heart disease data. In reference, the application of decision tree method in medical diagnosis and prediction is studied, and ID3 algorithm, C4.5 algorithm and cart algorithm are combined to diagnose and predict heart disease. However, through the analysis of UCI heart disease data set, it is not difficult to find that if the decision tree algorithm is used to build the model, it may create a too complex tree and can not predict the data very well, that is, it is prone to over fitting, and the decision tree may be unstable, and a very small mutation may also produce a completely different tree. For Bayesian algorithm, a stable prior probability is needed, and it is assumed that all features are independent of each other. Combined with the characteristics of UCI heart disease data set, it is necessary to build a more suitable classification method for the diagnosis and prediction of the disease.

Logistic regression is a machine learning algorithm proposed on the basis of linear regression, which is mostly used for binary classification problems. The classification algorithm is mainly used for disease diagnosis, natural disaster research, and advertising click through rate problems, such as CT prediction model of main artery injury after blunt pelvic ring fracture, Multivariate logistic regression model was used to evaluate the major artery injury after pelvic ring fracture. The prediction method of strong earthquake surface rupture based on logistic regression analysis uses logistic regression algorithm to evaluate whether there is surface rupture caused by earthquake in the construction site. Overview of click through rate prediction model for display advertising, through the comparison of logistic regression prediction method and deep learning model method to predict the click through rate of advertising. The logistic regression model is simple and fast, and can roughly infer the features that have great influence on the experimental results according to the feature weight. In this context, combined with the characteristics of UCI heart disease data set, this paper proposes a model based on logistic regression classification, taking heart disease data set as the research object, focusing on the final prediction of whether suffering from heart disease, and

[+] Corresponding author. Tel.: 13384915530; fax: 85262731.
E-mail address: *geniusbai@163.com*.

designs a large number of experiments to demonstrate. The results show that the accuracy and recall rate of the model for heart disease diagnosis and prediction are relatively high, which can meet the needs of auxiliary medicine, help doctors better grasp the condition and improve efficiency, and has important practical significance for the medical community.

2. LOGISTIC REGRESSION MODEL

Logistic regression is a simple machine learning classification algorithm based on linear regression model. The core idea of the classification algorithm is to establish regression formula for the classification boundary line according to the existing data, and then classify the samples“ Regression "means the best fitting, but to do the best fitting, we must find the best fitting parameters. For linear regression model, assume that the expression of a single sample data set is $\{x_1, x_2, x_3, \cdots, x_n\}$,Where x_i is the value of the corresponding feature, and the linear regression model attempts to obtain a function to predict through the linear combination of attributes, that is:

$$f(x) = w_1 x_1 + w_2 x_2 +, \ldots, + w_n x_n \tag{1}$$

The vector form is as follows:

$$f(x) = w^T x + b \tag{2}$$

Obviously, after learning w and b in linear regression, the corresponding model can be determined.

On this basis, the output of linear regression is regarded as the input, and the output data value is mapped to an output close to 0 or 1 through a function. Generally, the logarithmic probability function can be selected as the conversion function:

$$y = \frac{1}{1 + e^{-z}} \tag{3}$$

Then the logistic regression model can be described as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

Among them, $\theta = [b, w_1, w_2, \ldots, w_n]$, $h_\theta(x)$ is the possibility of prediction, X is the eigenvector matrix.

The results show that the logistic regression model has the following advantages compared with other models in the prediction of heart disease diagnosis results:

(1) It can be explained. According to the parameters learned from the logistic regression, we can roughly analyze the influence of different characteristics on the results, which can help doctors focus on the characteristics that have a great influence on the results in diagnosis. Compared with Bayesian model, logistic regression does not need stable a priori probability and independent conditions between

features, so it is more suitable for the prediction of heart disease.

(2) The output value is between 0 and 1, and has probability significance.

In conclusion, the logistic regression model is expected to have a better performance for the prediction of heart disease.

3. Prediction Model Of Heart Disease

3.1. Establishment Of Logistic Regression Model

The most important step in logistic regression is to solvethe characteristic coefficients based on training samples:

$$p(y|x) = \left(h_\theta(x)\right)^y \left(1 - h_\theta(x)\right)^{1-y} \tag{5}$$

Where $p(y|x)$ is the probability that the sample belongs to y label. On this basis, the maximum likelihood estimation of logistic regression model is as follows:

$$L(\theta) = \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)}\right) = \prod_{i=1}^{m} \left(h_\theta(x)\right)^{y^{(i)}} \left(1 - h_\theta(x)\right)^{(1-y)^{(i)}} \tag{6}$$

m is the total number of samples, $y^{(i)}$ is the i category, and $x^{(i)}$ is the i sample. For the sake of simplicity, the logarithm method can be used to convert equation (6) to:

$$T(\theta) = log\, L(\theta) = \sum_{i=1}^{m} y^{(i)}\, log\, h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - h\left(x^{(i)}\right)\right) \tag{7}$$

It can be seen from the above formula that the greater the probability of each sample belonging to its real marker in UCI heart disease data set, the better the effect. Based on the above formula, the loss function of logistic regression is introduced:

$$J(\theta) = -\frac{1}{m} T(\theta) \tag{8}$$

Obviously, the minimization of loss function is equivalent to the maximization of log likelihood function. According to the gradient descent algorithm, the parameters are solved:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)} \tag{9}$$

Among α For the learning rate, it's a small positive number. After the parameters are solved by the above formulas, the logistic regression model is finally determined.

3.2. Model Evaluation Strategy

For classification tasks, the overall accuracy of model prediction is generally used to measure the performance of the model. The overall accuracy is the ratio of the predicted correct samples in all samples. However, the object of this experiment is the patients who may have heart disease, and the main goal of this model is to predict the real patients with heart disease more accurately and find the patients as much as possible. Therefore, the judgment basis for the performance of the model in this paper can not only be based on the overall accuracy of the prediction model.

For the binary classification problem, the samples are divided into four cases according to the combination of the real category and the model prediction category, namely, true case, false positive case, true counterexample and false counterexample, thus forming the confusion matrix of the classification results, as shown in the following table.

Table 1: Confusion matrix

	Prediction example	Counter example of prediction
True examples	TP	FN
True counterexample	FP	TN

This paper is to predict the diagnosis of heart disease. Based on the above concepts, two criteria for evaluating the performance of the model are given.

Overall accuracy: It represents the percentage of samples with correct classification in the total samples in the classification task.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{10}$$

Recall: It indicates how many samples in the positive example can be successfully predicted by the model, that is, the size of coverage.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

When the model predicts the samples, it can improve the recall rate as much as possible under the condition of acceptable prediction accuracy, so as to find patients as much as possible, so that doctors can treat patients as early as possible.

3.3. Algorithm Description

The corresponding algorithm flow of heart disease prediction model of logistic regression method is as follows:

Table 2: Algorithm flow

Learning stage
Input: heart disease data T used as training data, learning rate α Output: learned logistic regression model M Step: Step1: Initialize the coefficients and b Step2: The output result of prediction is calculated Step3: The loss function is calculated Step4: Calculate the gradient Step5: Update the weight coefficient and b according to formula (9), and repeat step 2-step 5 until the minimum cost function J(θ) is obtained Step 6: Return to the trained logistic regression model M
Forecast stage
Input: trained logistic regression model M, prediction sample X Step: Step1: The median Z of all samples was calculated according to the model Step2: According to the intermediate value Z, the category is calculated Step3: Returns the category of prediction sample X

4. Experimental Analysis

4.1. Data Preprocessing

Data preprocessing is also known as data cleaning, and its main work is to denoise, fill in missing values, and type transform the data before it enters the algorithm flow. The UCI data set used in this experiment contains 14 features and 1 target value. Because the eigenvalues in the original data contain special types of data such as symbols, it is necessary to transform the original data. For example, the female data and the male data in the feature sex are mapped to 0 and 1, respectively. Map the down data to 0, the flat data to 1, and the up data to 2.

Table 3: Attribute table of UCI heart disease data set

Attribute	Attribute abbreviation	Data type
age	age	numerical type
sex	sex	semiotic type
types of chest pain	cp	semiotic type
venous pressure	trestbps	numerical type
weight of serum per milliliter of blood	chol	numerical type
fasting blood glucose	fbs	semiotic type
ECG results	restecg	semiotic type
maximal heart rate	thalach	numerical type
have angina during exercise	exang	semiotic type
exercise relative rest ST depression	oldpeak	numerical type
the inclination of ST segment in ECG	slop	semiotic type
number of vessels seen by fluoroscopy	ca	numerical type
defect type	thal	semiotic type
status	status	semiotic type

Table 4: corresponding data results after symbolic attribute conversion

Attribute	Numerical results after conversion
sex	female: 0 male: 1
cp	typical: 0 atypical: 1 non-anginal: 2 asymptomatic: 3
fbs	false: 0 true: 1
restecg	norm: 0 hyp: 1
exang	false: 0 true: 1
slop	down: 0 flat: 1 up: 2
thal	norm: 0 fix: 1 rev: 2
status	false: 0 true: 1

After the above processing, all the symbolic data in the dataset are converted into numerical data.

4.2. Normalization Of Data By Feature Engineering

Feature engineering is related to the design of feature sets for machine learning applications, focusing on how to design data features that conform to the characteristics of the data itself and the situation, including sample feature selection, data dimensionality reduction, data discretization, data normalization, data standardization and other operations. Because the data values of different features differ greatly, for example, the value after sex mapping is 0 or 1, while the value after Chol type conversion is usually a hundred digits, in order to remove the influence of dimension on simulation results, feature engineering processing is needed for UCI heart disease data set. In this paper, the data are normalized according to the characteristics of the data, The corresponding normalization formula is as follows:

$$x = \frac{x - min}{max - min} \tag{12}$$

When equation (12) is applied to each column, max and min correspond to the maximum and minimum values in the column respectively. Table 5 and table 6 show the data of the first 5 columns and the first 10 rows of the dataset before and after processing.

Table 5: data before normalization

Age	Sex	Cp	Trestbps	Chol
63	1	3	145	233
37	1	2	130	250
41	0	1	130	204
56	1	1	120	236
57	0	0	120	354
57	1	0	140	192
56	0	1	140	294
44	1	1	120	263
52	1	2	172	199
57	1	2	150	168

Table 6: data after normalization

Age	Sex	Cp	Trestbps	Chol
1	1	1	0.4807	0.3494
0	1	0.6666	0.1923	0.4408
0.1538	0	0.3333	0.1923	0.1967
0.7307	1	0.3333	0	0.3655
0.7692	0	0	0	1
0.7692	1	0	0.3846	0.1290
0.7307	0	0.3333	0.3846	0.6774
0.2692	1	0.3333	0.1923	0.5107
0.5769	1	0.6666	1	0.1666
0.7692	1	0.6666	0.5769	0

Through the above processing, the influence of dimension on the experimental results was eliminated, and a total of 304 data were obtained, including 166 sick data and 138 non sick data.

4.3. Verification Of Correctness

The data set is divided into 10 uncrossed subsets by using the cross method, one of which is taken as the test set each time, and the remaining 9 subsets are taken as the training set to train the model, and 10 operations are repeated to get 10 models, and each model is tested on the corresponding test set. The average accuracy and recall rate of 10 results were taken as the final result. The experimental results are shown in Table 7.

Table 7: Accuracy and recall results

Times	Accuracy	Recall
1	86.84%	87.23%
2	89.47%	93.19%
3	84.21%	95.45%
4	80.26%	87.81%
5	81.58%	90.91%
6	78.95%	82.86%

7	82.89%	92.31%
8	81.58%	90.00%
9	85.52%	90.48%
10	86.84%	93.18%
average	83.81%	90.34%

It can be seen from the above table that the average accuracy rate and average recall rate of this experiment are 83.81% and 90.34% respectively. Taking the results of the 9th experiment, the corresponding characteristic coefficient of this experiment is obtained, as shown in Figure 1.

```
accuracy score= 0.8552631578947368
        features weight
   age     [[-0.52635457]]
   sex     [[-1.12561845]]
   cp     [[2.05080471]]
   trestbps     [[-0.63599086]]
   chol     [[-0.17809827]]
   fbs     [[0.04658469]]
   restecg     [[0.76419463]]
   thalach     [[1.19313464]]
   exang     [[-1.23686575]]
   oldpeak     [[-1.61356606]]
   slope     [[0.98178158]]
   ca     [[-2.04471052]]
   thal     [[-1.49985989]]
```

Fig. 1: Weight results

The larger the absolute value of the characteristic coefficient, the greater the influence on the final experimental results. If the characteristic coefficient is greater than 0, it means that it has a positive correlation with the final experimental result; if the characteristic coefficient is less than 0, it means that it has a negative correlation with the final result; if the characteristic coefficient is equal to 0, it means that the characteristic has no effect on the final result. It can be seen from the above theory and result chart that the values of CP, CA, oldpack and thal of patients have a great impact on the prediction results. Therefore, doctors can focus on considering and analyzing these data values of patients in order to improve the accuracy and efficiency of diagnosis.

4.4. Influence Of Training Set Size On Experimental Results

In the experiment, the training set is set as 10%, 25%, 50%, 75% and 90% of the total number of samples respectively, and 10 experiments are carried out respectively, and the average value of 10 experimental results is taken as the final result of the experiment, and the corresponding experimental results are shown in Figure 2.

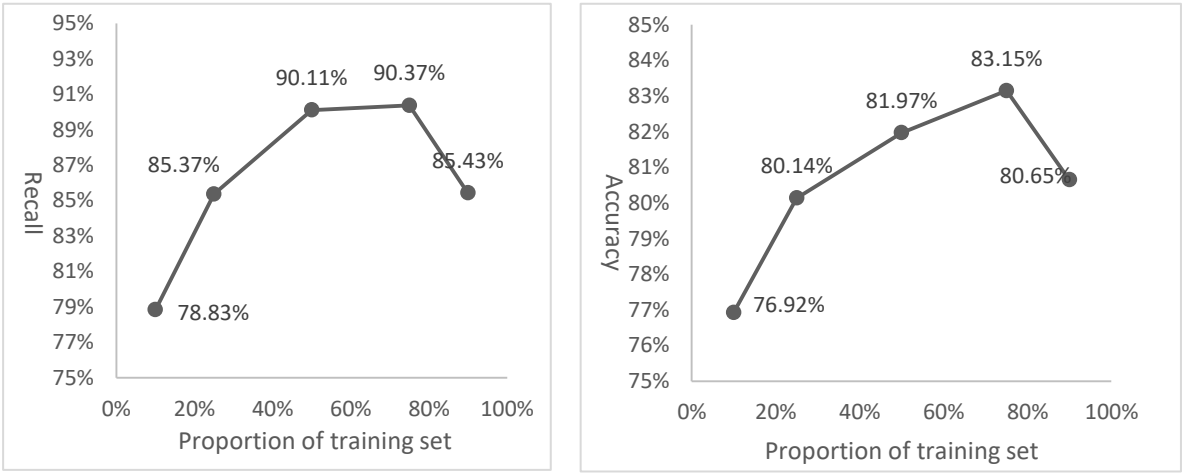

Fig. 2: Experimental results under different training sets

It can be seen from Figure 2 that the results of the corresponding model fluctuate greatly under different training sets. As the proportion of training set samples in the total samples increases, the accuracy and recall rate of the corresponding model also increase. When the proportion of training samples is too high, due to the lack of test samples, the accuracy of each test sample prediction will have a greater impact on the final results, and the accuracy and recall of the model will decline. Therefore, for this experiment, when the proportion of training samples is 70% ~ 80%, the effect will be better.

4.5. Influence Of Deleting FBS Feature On Experimental Results

According to the above experimental results, the absolute weight value of FBS (fasting blood glucose) is the smallest and close to 0. Therefore, we consider retraining the model after deleting FBS and conducting 10 comparative experiments. The training set of the experiment is selected as 75% of the total sample, and the average value of 10 experimental results is taken as the experimental result after deleting FBS feature, The statistical chart of the accuracy and recall rate of the corresponding results is shown in Figure 3.

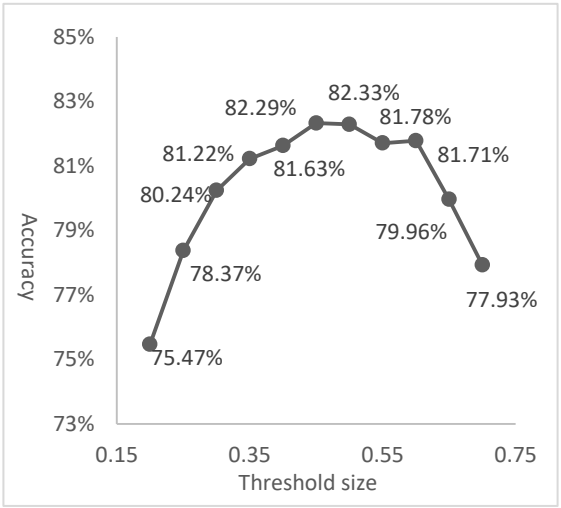

Fig. 3: Results before and after FBS deletion

It can be seen from the above figure that FBS feature deletion has little effect on the accuracy, but it slightly improves the recall rate. Therefore, considering the particularity of disease diagnosis and the efficiency of prediction model, in order to improve the recall rate of prediction and the efficiency of diagnosis, FBS feature can be deleted and the model can be trained again to improve the recall rate and prediction efficiency.

4.6. The Influence Of Classification Threshold On Experimental Results

Reviewing the principle of the above logistic regression algorithm, the output of the model is transformed into probability value through a logarithmic probability function, and the probability value is compared with the classification threshold. When the probability value is greater than the classification threshold, it is judged as a positive example, and when the probability value is less than the classification threshold, it is judged as a negative example. In general, the threshold of classification is 0.5.

But for the object of this study, we can consider setting the classification threshold less than 0.5 under the premise of acceptable accuracy to improve the recall rate, so as to find out the real patients with heart disease in the sample as much as possible. In this experiment, the threshold value of classification is set between 0.20 and 0.70, and other influence parameters of the experiment are the best values of the above experiment. The threshold step is 0.05, and 10 experiments are carried out, and the average value of 10 experiments is taken as the final result. The results are shown in figure4.

It can be seen that the threshold value has a great influence on the accuracy and recall rate. When the threshold value is between 0.45 and 0.50, the corresponding accuracy rate is the highest. This is because when the threshold value is too small, some samples will be misjudged as positive examples, while when the threshold value is too large, some samples will be misjudged as negative examples. Therefore, at the

beginning, the accuracy rate increases with the threshold value, When the threshold is large to a certain extent, it decreases with the increase of the threshold, and the accuracy is the highest in the range of 0.45 ~ 0.5. It can be seen from figure right that the recall rate is negatively correlated with the size of the threshold. This is because with the increasing of the threshold, the model will judge the samples that were originally judged as positive cases as negative cases with the increase of the threshold, so the recall rate will decrease with the increase of the threshold. Considering the accuracy rate and recall rate, it can be seen that the model has the best performance when the threshold value is 0.35 ~ 0.45, and the highest accuracy rate and recall rate are 82.33% and 95.83% respectively, which can reach the level of auxiliary medical treatment.

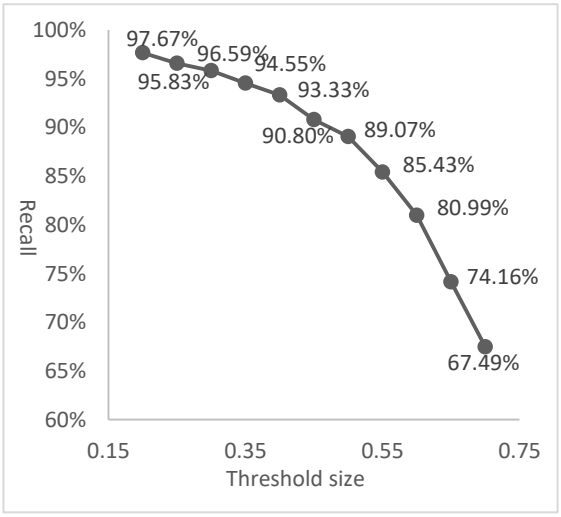

Fig. 4: Experimental results under different thresholds

5. Summary

Aiming at the problem of heart disease prediction, this paper proposes a heart disease prediction model based on logistic regression theory, and carries out a large number of simulation experiments. The experimental results show that the accuracy and recall rate of the model proposed in this paper are relatively high after a series of optimization, which can be used for auxiliary medical diagnosis. Moreover, this paper also analyzes the influence of different features on the disease diagnosis according to the experimental results. Therefore, doctors can focus on the features that have a great influence on the results according to the numerical value of feature weight, so as to further improve the diagnosis efficiency and accuracy.

6. References

[1] R. Dewri, and N. Chakraborti. Screening for Early-Stage Alzheimer's Disease Using Optimized Feature Sets and Machine Learning. *Pattern Recognition and Artificial Intelligence*. 2005, **13** (3): 173-183.

[2] R. Tin, and N. Wang. Simulating recrystallization through cellular automata and genetic algorithms. *Modelling Simul. Mater. Sci. Eng*. 2005, **15** (5): 177-188.

[3] R. Liu, and N. Zhang. Prognostic value of combined clinical and myocardial perfusion imaging data using machine learning. *Earthquake Engineering and Engineering Dynamics*. 2018, **33** (1): 177-189.

[4] R. Kleiman, and N. Liu. Prediction method of strong earthquake surface rupture based on logistic regression analysis. *Industrial Engineering Journal*. 2020, **13** (6): 117-129.

[5] R. Zhao, and N. Wei. Prediction model of rainfall induced landslide disaster based on feature aggregation decision tree. *Journal of Safety Science and Technology*. 2022, **1** (5): 100-120.