Intelligent Diagnosis of Heart Disease Based on Medical Feature Data

Feiyun Chen
High School Affiliated to Fudan University, Shanghai, China

Abstract: Heart disease has a very high incidence rate worldwide and is the "number one killer" that threatens human life safety. In the traditional medical industry, doctors can only diagnose patients' conditions based on their knowledge reserves and their own accumulated experience. In order to reduce the risk of misjudgment due to the lack of experience of doctors in small and medium-sized hospitals, the classification algorithm of machine learning can be used to assist doctors in judging the patient's condition. Doctors can use the classification results given by machine learning as a reference opinion and combine their own experience to make more accurate diagnosis and treatment judgments for patients. Using the massive data of the medical system, integrating artificial intelligence technologies such as machine learning, and building an auxiliary medical decision-making system is a key step in promoting smart medicine, which can bring innovation and change to the medical field. Therefore, this paper studies the diagnosis of heart disease by combining medical feature data with artificial intelligence related methods.

Keywords: Heart Disease; Artificial Intelligence; Data Analysis; Intelligent Diagnosis; Big Data

1. Introduction

Nowadays, with the continuous acceleration of the pace of life and the further pollution of the environment, people's health is seriously threatened. When seeking treatment in a hospital, it is very important to determine the type of disease the patient is suffering from. Misdiagnosis will lead to serious consequences, so there is an urgent need to reduce doctors' diagnostic errors. Traditional medical diagnosis is mainly based on the doctor's own experience, which results in sometimes large deviations in the diagnostic results of different doctors. As the amount of medical characteristic data in patients' examination reports increases, medical diagnosis becomes more complex, and it becomes more difficult for doctors to make accurate diagnoses based on large amounts of data. In order to improve the accuracy of doctors' diagnosis of patients' conditions when they have insufficient experience, machine learning algorithms are used to assist doctors in disease diagnosis and help doctors improve the accuracy of diagnosis.

Machine learning technology is a hot research topic today, and it analyzes the results of events by learning the intrinsic relationships between given samples. For many practical problems, the classification effect is average when using traditional classification methods. Using classification algorithms in machine learning can significantly improve classification capabilities. The machine learning method trains the corresponding classification model through the training set, and then brings the medical feature data into the trained classification model to determine whether there is a possibility of heart disease. For heart diseases, directly using a certain algorithm to classify may not necessarily achieve the desired results. There are relatively large differences in classification effects on different data sets. Prediagnosis of heart disease has very high requirements on accuracy, so it is necessary to select a more effective classification algorithm corresponding to it to improve the accuracy of prediction of heart disease.

The World Health Organization's summary report on human causes of death states that heart disease is one of the leading causes of death worldwide. Therefore, if we can predict people's probability of heart disease based on some medical characteristic data related to physical health, it can effectively assist doctors in diagnosis. Cardiovascular stenosis will cause insufficient blood supply to the myocardium, which in turn can lead to many heart disease problems such as angina pectoris and coronary heart disease. Therefore, analyzing the proportion of cardiovascular vessels with narrowed lumen diameter through classification algorithms can be used as an important criterion for judging whether there is a possibility of heart disease.

Therefore, this project aims at the above problems to research and develop several different classification algorithms to find a classification model that is relatively effective in predicting heart disease, improve the prediction accuracy on heart disease data sets, and assist doctors in intelligent diagnosis.

2. Research Methods

2.1 K-nearest neighbor

K-nearest neighbor is a supervised learning classification algorithm first proposed in 1968. Its principle is to first input unlabeled samples to be classified, find the K samples with the highest similarity to the newly input samples in the training set, and the category that appears most frequently is the category to which the new sample belongs. It can be simply considered as finding the nearest K neighbors. When K=1, the K-nearest neighbor algorithm becomes the nearest neighbor algorithm, that is, finding the nearest neighbor, and the category of the sample to be classified will be the same as that of the nearest neighbor. The core of the K-nearest neighbor algorithm is to find neighbors and measure the distance between neighbors to determine the degree of similarity between them. Generally, the commonly used distance measurement representations of the feature space of the K-nearest neighbor model include Euclidean distance, Manhattan distance, angle cosine, etc. The quality of the classifier in the K-nearest neighbor algorithm is mainly determined by the selection of K value. If the selected value is too small, if the neighbor happens to be a noise point, it will be overfitting. On the contrary, if the K value is too large, samples that are slightly farther away will also play a role, which will cause prediction errors. The K value generally tends to select a relatively small value, and the cross-validation method is used to select the optimal value. The calculation process of the K nearest neighbor classification model for the heart disease dataset is as follows in figure 1:

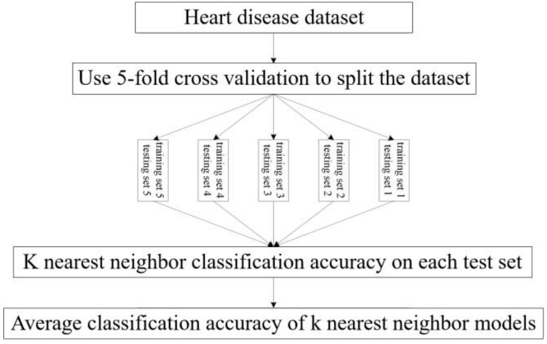

Heart disease dataset
Use 5-fold cross validation to split the dataset
training set 5
testing set 5
training set 4
testing set 4
training set 3
testing set 3
training set 2
testing set 2
training set 1
testing set 1
K nearest neighbor classification accuracy on each test set
Average classification accuracy of k nearest neighbor models

Figure 1. K-Nearest Neighbor Model.

2.2 Support Vector Machine

Support Vector Machine (SVM) is a classifier that has become the mainstream in machine learning technology due to its excellent performance in high-dimensional data operations and classification applications. SVM is a linear large-margin classifier in feature space, which maximizes the margin and is transformed into solving a convex quadratic optimization problem. Given two different types of data, x represents the data and y represents the category to which it belongs (y takes 1 or -1 to represent different categories), the learning task of the classifier used for data partitioning is to select a hyperplane in the n-dimensional feature space. The solid line shown in Figure 2 is the optimal hyperplane sought, which is located in the middle of the two dotted lines, and the distance from the two dotted lines is the geometric interval, and the support vector is located on the dotted line.

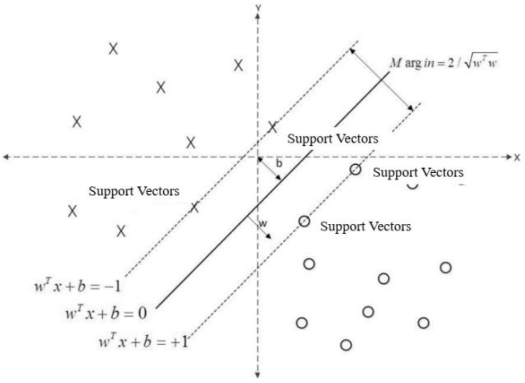


Figure 2. Support Vector Machine.

In most cases, data points are not linearly separable. At this time, SVM first completes the operation in the low-dimensional feature space, and the kernel function maps the input from the low-dimensional space to the high-dimensional space to find the best hyperplane, so that nonlinear data that is difficult to separate in the original space can be divided. As shown in Figure 3, since it is linearly inseparable in the two-dimensional space, it is mapped to the three-dimensional space for linear division.

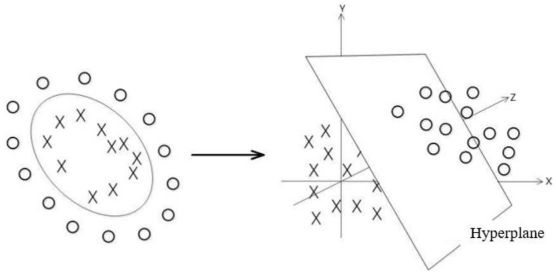


Figure 3. Space Mapping.

The kernel functions of support vector machines include radial basis function (RBF), linear, polynomial, etc. The calculation process of the average classification accuracy of the SVM classification model for the heart disease data set is shown in Figure 4:

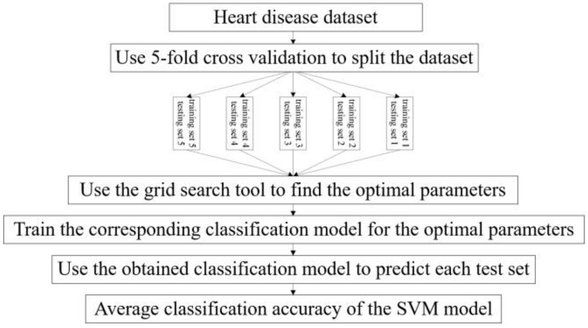

Heart disease dataset
Use 5-fold cross validation to split the dataset
training set 5
testing set 5
training set 4
testing set 4
training set 3
testing set 3
training set 2
testing set 2
training set 1
testing set 1
Use the grid search tool to find the optimal parameters
Train the corresponding classification model for the optimal parameters
Use the obtained classification model to predict each test set
Average classification accuracy of the SVM model

Figure 4. The Average Classification Accuracy Calculation Process of the SVM Classification Model On The Heart Disease Dataset.

2.3 Logistic Regression

Logistic regression, also known as "logarithmic probability regression", estimates the probability of something happening. It has been used to predict the probability, and by predicting the probability, the probability of diagnosing a certain disease can be achieved.

Logistic regression is to learn a binary classification model from features, and the independent variable is a linear combination of features. Since its independent variable values are between negative infinity and positive infinity, a function is used to map the independent variable to the interval (0, 1), and the mapped value is the probability of y=1. The function is shown in Figure 5. The calculation process of the average classification accuracy of the logistic regression classification model for the heart disease data set is shown in Figure 6.

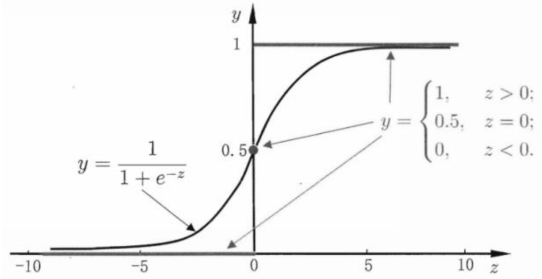


Figure 5. Logistic Regression.

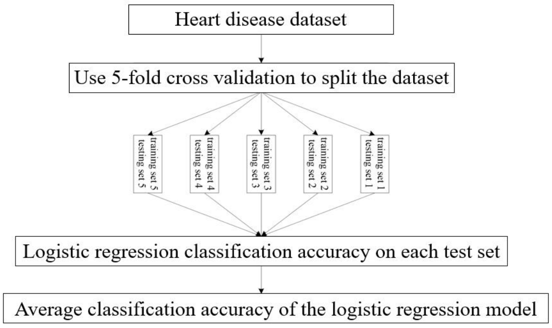
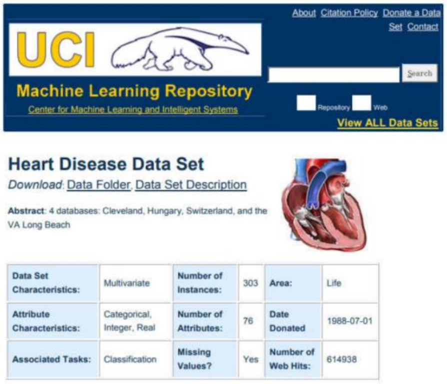

Heart disease dataset
Use 5-fold cross validation to split the dataset
training set 5
testing set 5
training set 4
testing set 4
training set 3
testing set 3
training set 2
testing set 2
training set 1
testing set 1
Logistic regression classification accuracy on each test set
Average classification accuracy of the logistic regression model

Figure 6. Calculation Process of Average Classification Accuracy of Logistic Regression Classification Model on Heart Disease Dataset.

3. Data Analysis and Experimental Results

3.1 Data Source

This article uses the UCI heart disease open source dataset in the study, which was created by Dr. Andras Janosi of the Hungarian Institute of Cardiology, Dr. William Steinbrunn of the University Hospital of Zurich, Switzerland, Dr. Matthias Pfisterer of the University Hospital of Basel, Switzerland, and Dr. Robert Detrano of the Long Beach Medical Center and Cleveland Clinic Foundation. As shown in Figure 7:



Heart Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	76	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	614938

Figure 7. Data Source

The heart disease dataset contains 76 attributes, 14 of which are mainly used. The heart disease dataset from the Cleveland Clinic is the most used dataset for heart disease research so far. The "target" field corresponding to the classification result in the attribute represents the classification of the possibility of heart disease, and its value ranges from 0 (no possibility of disease) to 4 (very likely to be diseased). The calculation of the heart disease dataset focuses on trying to distinguish between the impossible disease (value 0) and the possible disease (value 1, 2, 3, 4). In addition to using this heart disease dataset, this study also uses the dataset of the Hungarian Institute of Heart Disease. The heart disease dataset of the Cleveland Clinic has 303 instances, and the default value is represented by the symbol "?". Excluding the default data, there are 297 available instances. After processing, the dataset of the Hungarian Institute of Heart Disease has 294 available instances. The attributes used for classification are divided into 13 attributes for prediction and 1 attribute used to indicate the possibility of no existence (value 0) or the possibility of different degrees (value 1, 2, 3, 4).

3.2 K-Nearest Neighbor Experimental Results

The number of instances of misclassification of the heart disease dataset of the Cleveland Clinic by the k-nearest neighbor classification method and its corresponding average classification accuracy are shown in Table 1. When K=7, the average classification accuracy of the heart disease dataset of the Cleveland Clinic by the k-nearest neighbor classification model is the highest, which is 79.51%. Table 2 shows the number of instances of misclassification of the heart disease dataset of the Hungarian Institute of Heart Disease and its corresponding average classification accuracy when K is 3, 7, 11, and 15. When K=11, the average classification accuracy of the heart disease dataset of the Hungarian Institute of Heart Disease by the k-nearest neighbor classification model is the highest, which is 79.26%.

Table 1. Classification Results of K Nearest Neighbor Classifier on Cleveland Clinic Heart Disease Dataset.

testing set *K*	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5	accuracy
3	18	11	12	13	10	78.51%
7	17	9	11	12	12	79.51%
11	20	11	11	9	14	78.19%
15	20	13	11	7	12	78.87%

Table 2. Classification Results of K nearest Neighbor Classifier on the Heart Disease Dataset of Hungarian Institute of Cardiology.

testing set *K*	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5	accuracy
3	12	10	13	14	12	79.25%
7	11	13	16	14	12	77.56%
11	10	13	12	12	14	79.26%
15	9	13	14	13	13	78.93%

3.3 SVM Experimental Results

The prediction results of the SVM classification model on the heart disease data set of the Cleveland Clinic were predicted by the linear kernel SVM classifier and the RBF kernel SVM classifier. According to the number of misclassified instances in Tables 3 and 4, the average classification accuracy of the RBF kernel SVM classification model is 85.56%, and the average classification accuracy of the linear kernel SVM classification model is 84.88%.

Table 3. Classification Results of the Linear Kernel SVM Classifier on the Cleveland Clinic Heart Disease Dataset.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
C	2^3	2^{-1}	2	2	2
Number of errors	12	7	12	5	9

Table 4. Classification Results of RBF Kernel SVM Classifier on Cleveland Clinic Heart Disease Dataset.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
C	2^7	2^5	2	2^3	2
γ	2^{-7}	2^{-7}	2	2^{-3}	2^{-3}
Number of errors	12	7	11	5	8

The prediction results of the SVM classification model on the heart disease dataset of the Hungarian Institute of Cardiology were predicted by the linear kernel SVM classifier and the RBF kernel SVM classifier. According to the number of misclassified instances in Tables 5 and 6, the average classification accuracy of the RBF kernel SVM classification model is 87.78%, and the average classification accuracy of the linear kernel SVM classification model is 84.03%.

Table 5. Classification Results of the Linear Kernel SVM Classifier on the Heart Disease Dataset of the Hungarian Institute of Cardiology.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
C	2^3	2^{15}	2^3	2^{-3}	2^7
Number of errors	7	10	11	8	11

Table 6. Classification Results of RBF Kernel SVM Classifier on the Heart Disease Dataset of Hungarian Institute of Cardiology.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
C	2^9	2^{11}	2^9	2^{-1}	2^{13}
γ	2^{-3}	2^{-3}	2^{-3}	2^{-3}	2^{-7}
Number of errors	2	9	8	8	9

3.4 Logistic Regression Experiment Results

From the number of misclassified instances in Table 7, we can see that the average classification accuracy of the logistic regression classification model for the Cleveland Clinic heart disease data set is 83.87%.

Table 7. Classification Results of Logistic Regression Classifier on Cleveland Clinic Heart Disease Dataset.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
Number of errors	13	7	12	7	9

From the number of misclassified instances in Table 8, it can be seen that the average classification accuracy of the logistic regression classification model for the heart disease dataset of the Hungarian Institute of Cardiology is 81.32%.

Table 8. Classification Results of Logistic Regression Classifier on Cleveland Clinic Heart Disease Dataset.

	testing set 1	testing set 2	testing set 3	testing set 4	testing set 5
Number of errors	7	13	13	9	13

4. Conclusion and Discussion

By using K-nearest neighbor, linear kernel SVM, RBF kernel SVM, and logistic regression classification algorithms to build their corresponding classification models, the heart disease data sets of the Cleveland Clinic and the Hungarian Institute of Heart Disease were classified and predicted, and the average classification accuracy is shown in Table 9. It can be seen intuitively that the top three average classification accuracies on the heart disease data sets of the Cleveland Clinic and the Hungarian Institute of Heart Disease are: RBF kernel SVM, linear kernel SVM, and logistic regression. Among them, RBF kernel SVM has a higher average classification accuracy on the heart disease data set. This conclusion will serve as the basis for subsequent research to further improve the classification accuracy on the heart disease data set.

Table 9. Classification Results of Heart Disease Dataset.

Classification Model	Average classification accuracy (Cleveland Clinic)	Average classification accuracy (Hungarian Institute of Cardiology)
K-nearest neighbor	79.51%（K=7）	79.26%（K=11）
Linear Kernel SVM	84.88%	84.03%
RBF kernel SVM	85.56%	87.78%
Logistic Regression	83.87%	81.32%

References

[1]T.Tantimongcolwat, T.Naenna, C.Isarankura-Na-Ayudhya, et al. Identification of ischemic heart disease via machine learning analysis on magnetocardiograms[J]. Computers In Biology And Medicine, 2008, 38(7): 817-825.

[2]He Shiqi. Research on the identification rules of TCM syndrome of unstable angina pectoris in coronary heart disease based on data mining[D]. Beijing University of Chinese Medicine, 2012.

[3]Zhou Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 23.

[4]X. S. Yang. A novel improved accelerated particle swarm optimization algorithm for global numerical[J]. Emerald, 2014, 31(7): 1198-1220.

[5]Shi Qi, Wang Wei, Li Youlin, et al. Study on the identification model of blood stasis syndrome in patients with coronary heart disease based on metabolomics[J]. Journal of Integrated Traditional Chinese and Western Medicine for Cardiovascular and Cerebrovascular Diseases, 2014, (05): 513-516.