Early Heart Disease Diagnosis Method Based On KNN

Yuxuan Su [1, *] and Zhen Yang [2]

[1] School of Information Technology, Shanghai Ocean University, Shanghai, China

[2] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

* Corresponding Author Email: B22042419@njupt.edu.cn

Abstract. Heart disease is a health problem of wide concern around the world. This paper aims to realize an early prediction of heart disease based on k-nearest neighbors algorithm to warn the users of their potential risks of heart disease in an early stage so that measures can be taken to minimize the danger. The dataset introduced contains basic information of body that is possibly related to heart disease, with 14 feature dimensions and 303 samples in total. To realize the proposition of early prediction, three different strategies for feature dimension choosing are introduced to evaluate the predictions based on all the data, data that can be measured at home with household devices and data that do not require any extra measurements and devices. Several indexes are introduced to evaluate and compare the performance of the models that are trained and the changes caused by the decrease of feature dimensions are analyzed. Finally, possible future work for improving the models is discussed.

Keywords: Heart disease diagnosis; KNN; machine learning.

1. Introduction

The human body has many important organs, such as the heart, liver, lungs, and so on, and one of the most important is the heart [1,2]. Heart disease caused by arterial stenosis and decreased blood and oxygen in the heart [3] has now become one of the most common diseases in the world [4]. Heart disease claims the lives in the amount of 17.5 million people every year, which is 30% of all fatalities.The biggest problem in the treatment of heart disease now is that despite the best efforts of doctors, the mortality rate of heart disease is still high due to untimely detection of heart disease [5]. Therefore, exploring human health records can effectively discover their connection with heart disease [6.7], thereby determining in advance whether they have heart disease, which is of great help to patients and doctors [8.9] and can be effective. Previous research has also shown that machine learning has great help in predicting heart disease [10].

Considering these circumstances, this article explores the impact of 14 feature dimensions on heart disease and divides the feature dimensions into three groups: all dimensions, dimensions that require some household instrument measurement, and three groups that do not require instrument measurement. KNN classifiers are used for forecasting cardiac disease, and the models are each trained individually. On this basis, we further utilized methods such as changing the distance weight of the KNN classifier, comparing the impact of different distance calculation formulas on the model, and testing different K-classification coefficients to further optimize the KNN classifier model. Moreover, under three different diagnostic dimensions, the same model measured multiple data such as accuracy, recall, f1 score, precision, and auc, and evaluated the value of a model from multiple perspectives. On this basis, multiple data from several models were compared horizontally, and a columnar statistical chart was drawn to visualize the data more intuitively. Overall,this article attempts to improve the practicality of the model by reducing the difficulty of obtaining feature dimensions and sacrificing a portion of the model's accuracy. After all, a model that draws conclusions about having heart disease through a set of professional systemic examinations lacks practicality.

2. Method

2.1. Pipeline

Firstly, obtain a dataset for heart disease prediction on Kaggle, and then divide the feature dimensions into three groups: all dimensions, dimensions that require some household instrument measurement, and three groups that do not require instrument measurement. Train the KNN classifier model separately, and then optimize it to improve the model value. Perform a horizontal comparison between the last three models, draw a claim statistical chart to display, analyze the data, and draw the final conclusion.

2.2. KNN

Essentially, it is to calculate the distance between new data and data that has already been clearly classified, and then classify the points to be classified into the closest category according to the principle of minority obeying majority and shortest distance. Specifically, preprocess the data first, divide each data into N dimensional vectors, and then calculate the distance from all points to the points to be classified. Sort them in descending order of distance. Take the first K and take the classification with the highest proportion among them, which is the classification of the points to be tested.

And there is more than one distance calculation formula for KNN classifiers. Common distance calculation formulas include Euclidean formula and Manhattan distance formula. The Euclidean distance formula is shown in the following figure (1), where $x_n$and y_n represent the coordinates of the nth dimension, respectively [11].

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

The opposite is the Manhattan formula, as shown in the following figure (2). Similarly,x_nn and y_n represent the coordinates of the nth dimension, respectively.

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| \quad (2)$$

In this study, different classification methods will use different calculation formulas to achieve better model prediction performance, which will be explained in detail in subsequent sections.

The algorithm has a simple idea and is easy to understand and implement. There is no need to divide the data into regions, and future data can be directly classified based on existing datasets more intuitively. The required amount of data is smaller compared to other machine learning algorithms, making it more suitable for smaller datasets.

But KNN classifiers still have some drawbacks. Difficulty in handling imbalanced samples can easily lead to misclassification. For example, in an unbalanced dataset, there are only k/2-1 A-class data in total. Therefore, under the KNN classification algorithm, all new data points will be classified as B-class, which is obviously absurd. Another scenario is that the N song points closest to the point X to be classified belong to Class A (N is less than K/2-1). Normally, point X should be classified as Class A. However, due to the imbalanced dataset, there are too few Classes A feature points. Although X is very close to Class A, it will still be classified into other categories, which clearly leads to distortion in KNN classification.

Although the core algorithm of KNN classification is simple, the good news is that the dataset in this study is relatively balanced, and through reinforcement learning? Further optimization of the model has achieved good results.

3. Experiment

3.1. Dataset Introduction

This study is based on the Heart Disease Dataset dataset of the Kaggle data platform, which includes 303 sets of data covering a total of 13 characteristic dimensions such as gender, cholesterol content, maximum heart rate, etc. (specific content is shown in Table 1), including 202 males and 96 females, with a gender ratio close to 2:1, which is relatively balanced. From the perspective of illness, this dataset includes 165 patients and 138 healthy individuals, which is relatively average. The average age of all respondents is about 54 years old, and the dataset is relatively complete with no missing values.

Table 1 Characteristic Dimension 1.

Number	Name	Description	Type
1	age	age	Continuous
2	sex	sex	Discrete
3	cp	History of angina pectoris	Discrete
4	trtbps	Resting blood pressure	Continuous
5	chol	Cholesterol content	Continuous
6	fbs	Is there a high blood sugar level on an empty stomach	Discrete
7	restecg	Resting electrocardiogram characteristics	Discrete
8	thal	Maximum Heart Rate	Continuous
9	exng	Does exercise cause angina	Discrete
10	oldpeak	Old peak of electrocardiogram ST-T wave	Continuous
11	slp	Peak slope of electrocardiogram ST-T wave	Discrete
12	caa	Number of large blood vessels around the heart	Discrete
13	thall	History of Mediterranean anemia	Discrete
14	output	Whether have heart disease	Discrete

3.2. Feature selection

Based on this dataset, this study divided all feature dimensions into three groups, as shown in the following figure.

Table 2 Corresponding Special Diagnosis Dimensions for Three Classification Methods2

Group 1	Group 2	Group 3
age	age	age
sex	sex	sex
cp	cp	cp
thal	thal	thal
exng	exng	exng
thall	thall	thall
output	output	output
trtbps	trtbps	
chol	chol	
fbs	fbs	
restecg		
oldpeak		
slp		
caa		

Among them, the first group contains all feature dimensions of the dataset. The second group selected data that can be measured at home using household instruments. The third group is data that

does not require any instrument measurement at all. The three filtering methods have a progressively decreasing difficulty in obtaining data. The practicality of the corresponding model increases step by step.

3.3. Model training and testing

On the basis of the KNN classifier, the following optimizations were carried out in this study. Firstly, change the weight corresponding to each point to the reciprocal of its distance from the point to be classified. This can effectively enhance the impact of nearby points on the points to be classified and reduce classification errors caused by small data sizes. Moreover, after investigation, it was found that using the Manhattan distance calculation formula would have better results for classification three. For classification one and classification two, using the Euclidean calculation formula will yield better results. For classification three, its feature dimensions are relatively small, so adjusting the value of n_neighbors to 3 will result in better experimental results.

4. Result and Discussion

4.1. Index for evaluation

First, since it can reflect the correct classification ability of the models, accuracy is the primary index to evaluate the models. Besides, as the models are mainly trained to achieve a rough prediction of the possible risk of heart disease for users before they can get professional medical assistance, omitting real patients can bring higher risks than wrongly judging healthy people, so recall is considered more important than precision. F1-score is also an important index which is the weighted harmonic average of recall and precision.

4.2. Analysis of results

The three groups of feature dimensions are respectively used to train and evaluate different models. Table 3 shows the indexes of the three groups.

Table 3 Performance for Three Models

	Classification 1	Classification 2	Classification 3
Accuracy	83.5%	80.2%	81.3%
Recall	91.11%	92%	86.67%
Precision	78.8%	73.7%	78%
F1-score	84.5%	82.3%	82.1%
AUC	83.6%	80.4%	81.4%

For classification 1 which contains all the feature dimensions, the value of n_neighbors is set to be 5 and euclidean distance is proved to work better. Changes of other values have less influence on the result and can hardly further enhance the accuracy which finally reaches approximately 84%. Other indexes are also outstanding among the three groups as the recall reaches 91.11%, precision teaches 78.8%, F1-score teaches 84.5% and the AUC reaches 83.6%.

For classification 2 which drops data that can hardly be measured without professional medical devices, the value of n_neighbors is decreased to 3 and Manhattan distance is used. The accuracy dropped to about 80.2% which is still close to the accuracy of classification 1. It is also worth noticing that the recall of classification 2 is the only index that is even higher than the one of classification 1 and the precision is far lower than the recall. Decreases can also be witnessed among other indexes, but the changes are limited to around 2%.

For classification 3 only data that can be provided without any extra devices or measurements, which is mainly based on personal feelings or medical histories are included. The value of n_neighbors is set to be 5 and euclidean distance is chosen. Surprisingly, the accuracy of classification 3 is higher than the one of classification 2, which is 81.3%. This is possibly related to the low

correlation of the feature dimensions dropped. However, the recall of classification 3 is significantly lower comparing to classification 1 and 2, reaching 86.67%, c though the precision is close to the one of classification1. It is obviously that the recall is most influenced in this case.

Despite the differences among precision and recall, the F1-scores and AUC of classification 2 and 3 are similar, maintaining at approximately 82% and 80% respectively. They are also close to the values of classification 1.

Figure 1 shows the comparison of several indexes of the 3 groups. It can be seen that the decrease is not too noticeable and the performance of models with less data is still reliable. Since one of the most important problems with heart disease is that patients are left untreated and do not get medical assistance in time, sacrificing part of the performance and looking for an earlier and easier heart disease prediction should be meaningful.

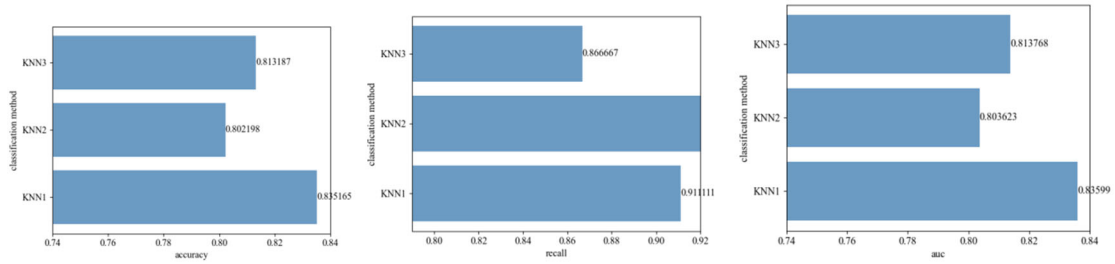

Figure 1 Indexes for Three Classification Models(Accuracy、Recall、AUC)

5. Conclusion

In this article, the k-nearest neighbors' algorithm is used to realize early prediction of heart disease. The model training is based on a dataset that contains 13 different basic body indexes that may be related to heart disease. In order to make the prediction easy to be conducted at home so that users can be warned in advance and get professional assistance, three kinds of strategies are introduced to process the dataset. It can be learned that although the decrease of feature dimensions can inevitably influence the performances of models, most indexes for model evaluation are not significantly affected and a rough prediction as well as an early warning of the risk of heart disease is possible.

Various future work can be done. Since the dataset is not large enough and the feature dimensions are also limited, introducing other datasets containing more statistics and feature dimensions should help to enhance the performance of models. The correlations of different feature dimensions also deserve further discussion. Other machine learning algorithm can be tested to select the most reasonable strategy. Combination of different algorithm is also an option.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Xinning Han. 2023. Heart Disease Type Prediction Model Based on SVM-ANN. In Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering (EITCE '22). Association for Computing Machinery, New York, NY, USA, 422–426.

[2] Asif Rahman Snigdha, Syeda Nishat Tasnim, Kamran Rafsan Miah, and Tohedul Islam. 2022. Early Prediction of Heart Attack using Machine Learning Algorithms. In Proceedings of the 2nd International Conference on Computing Advancements (ICCA '22). Association for Computing Machinery, New York, NY, USA, 344–348.

[3] Adel, Lahsasna, R. N., Ainon, Roziati, Zainuddin, & Awang, Bulgiba. 2012. Designof a fuzzy-based decision support system for coronary heart disease diagnosis.Journal of medical systems. 36, 5 (Jan. 2012), 3293-3306.

[4] Siyue Song, Tianhua Chen, and Grigoris Antoniou. 2021. ANFIS Models for Heart Disease Prediction. In Proceedings of the 2021 5th International Conference on Innovation in Artificial Intelligence (ICIAI '21). Association for Computing Machinery, New York, NY, USA, 32–35.

[5] Ching-seh Mike Wu, Mustafa Badshah, and Vishwa Bhagwat. 2019. Heart Disease Prediction Using Data Mining Techniques. In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology (DSIT 2019). Association for Computing Machinery, New York, NY, USA, 7–11.

[6] Chapman B, DeVore AD, Mentz RJ, Metra M. Clinical profiles in acute heartfailure: an urgent need for a new approach. Eur Soc Cardiol (ESC) Heart Fail.2019; 6(3):464–74.

[7] Poffo MR, Assis AVd, Fracasso M, Londero Filho OM, Alves SMdM, Bald AP,Schmitt CB, Alves Filho NR. Profile of patients hospitalized for heart failure intertiary care hospital. Int J Cardiovasc Sci. 2017; 30:189–98.

[8] Pandey AC, Topol EJ. Dispense with supplements for improving heart outcomes.Ann Intern Med. 2019; 171:216–7.

[9] Khan SU, Khan MU, Riaz H, Valavoor S, Zhao D, Vaughan L, Okunrintemi V, RiazIB, Khan MS, Kaluski E, Murad MH, Blaha MJ, Guallar E, Michos ED. Effects ofnutritional supplements and dietary interventions on cardiovascular outcomes:an umbrella review and evidence map. Ann Intern Med. 2019; 171:190–8.

[10] Pronab Ghosh, Sami Azam, Asif Karim, Mirjam Jonkman, and MD. Zahid Hasan. 2021. Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases. In 2021 the 5th International Conference on Information System and Data Mining (ICISDM 2021). Association for Computing Machinery, New York, NY, USA, 14–20.

[11] Huanian Zhang and Fanliang Bu. 2019. Weighted KNN Algorithm Based on Random Forests. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 231–235.

[12] Alexey Markin and Oliver Eulenstein. 2016. Manhattan Path-Difference Median Trees. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '16). Association for Computing Machinery, New York, NY, USA, 404–413.