International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 3, Issue.4, July-August 2022, page no. 01-07

To cite this article: Lin MEI, Ning GENG and Jie WANG (2022). PREDICTION OF HEART DISEASE CLASSIFICATION BY RANDOM FOREST MODEL BASED ON GREY WOLF OPTIMIZATION ALGORITHM, International Journal of Applied Science and Engineering Review (IJASER) 3 (4): 01-07

PREDICTION OF HEART DISEASE CLASSIFICATION BY RANDOM FOREST MODEL BASED ON GREY WOLF OPTIMIZATION ALGORITHM

Lin MEI, Ning GENG and Jie WANG

Department of Basic Education, Shandong University of Engineering and Vocational Technology, Jinan, 250200, China.

DOI: http://dx.doi.org/10.52267/IJASER.2022.3401

ABSTRACT

This paper proposes a random forest model based on the gray wolf optimization algorithm to study the classification of heart disease, and compares it with the prediction results of decision tree and basic random forest. The results show that the random forest using the optimization algorithm has better prediction results, its accuracy is 5.2% higher than that of ordinary random forest, which provides a certain reference for the prediction and prevention of heart disease in the future.

KEYWORDS: heart disease classification research; random forest; grey wolf optimization algorithm

1. INTRODUCTION

In 2018, cardiovascular death was the first cause of death among urban and rural residents in my country [1]. In recent years, with the development of science and technology, providing residents with a convenient life has also brought bad habits such as unhealthy diet and lack of exercise, which has caused the incidence of diseases such as hypertension and diabetes in my country to increase year by year, which has further affected our country. Cardiovascular morbidity and mortality.

From 2005 to 2020, the mortality rate of cardiovascular disease among Chinese residents increased by 48.06% [2]. It can be seen that the prevention and treatment of cardiovascular disease is imminent. In order to implement the national strategy of "focusing on prevention and focusing on the grassroots", how to detect, identify and treat early has become an important discussion point for the prevention and treatment of heart disease.

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

Based on this, Liu et al. clustered the data features, and applied XGboost for classification prediction, and obtained better prediction results [3]. Sun et al. used a variety of machine learning methods such as logistic regression, SVM, and KNN to model 13 influencing factors of heart disease, and compared various methods to reflect the prediction effect of different algorithms [4]. Zhao et al. used K-nearest neighbors to preprocess the initial data in the process of heart disease prediction, and determined the optimal parameters in the random forest through cross-checking and grid search, and the results were better compared with other machine learning models [5].

To sum up, scholars have used various machine learning methods to predict and classify heart disease. On this basis, this paper optimizes the random forest algorithm through the gray wolf optimization algorithm, in order to obtain a better classification effect, which will serve as a basis for future heart disease. reference for prevention.

2. MODEL APPROACH

2.1 Decision tree

Decision tree is a commonly used classification algorithm. The classification rules it generates are simple and clear, which is convenient for display and understanding. However, it also has shortcomings such as unstable results and easy overfitting. Therefore, the random forest algorithm appears.

2.2 Random Forest

Random forest is an ensemble learning method that uses bagging strategy, not only bootstrap (with replacement sampling) is used when generating the sample set selection of each classifier, but the unselected data (OOB) is used as the test set of the classifier, and a random extraction mechanism is also added to the feature selection. This process of generating classifiers is repeated to generate a "forest", and finally the majority is selected as the result through voting by each classifier. Random forest takes into account the influence of each feature in the running process. Compared with decision tree, its ability to resist overfitting is stronger, and it also has the advantages of fast training speed when running in parallel, but there are still parameters such as the number of trees in the forest. It is impossible to choose the optimal and other problems, so how to find the optimal parameters has become one of the problems to be studied by the random forest algorithm.

2.3 Grey Wolf Optimization Algorithm

Grey Wolf Optimizer (GWO) is a swarm intelligence optimization algorithm proposed by Mirjalili et al., a scholar at Griffith University in Australia in 2014[6]. The algorithm is an optimization search method inspired by the prey activity of gray wolves. It simulates the social hierarchy relationship

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

followed by gray wolves. When this relationship is applied to the algorithm process, it has strong convergence performance and fewer parameters, easy to implement and so on. The whole algorithm includes the processes of social hierarchy, encircling the prey and attacking the prey.

2.3.1 The social hierarchy of wolves

The gray wolf is a very special group animal, and there is a strict hierarchy between the groups. Its main body is pyramid-shaped as shown in Figure 1.

α is the first level of the pyramid, that is, the leader in the wolf pack, who plays the role of management and decision-making in the population.

β is the second layer of the pyramid, that is, the leader's think tank, which is mainly responsible for assisting the leader in making decisions, and when the leader's position is vacant, β will temporarily assume the identity of α.

δ is the third layer of the pyramid, obeying α and β, and at the same time dominates the ω of the fourth layer of the pyramid, and the main function of ω is to obey orders and obey the orders of other levels of wolves.

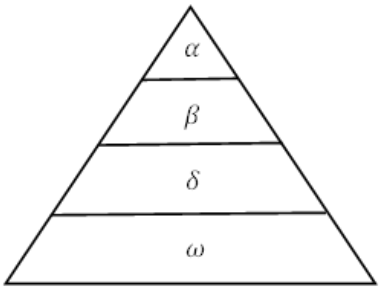

Figure 1 Wolves social class

2.3.2 Steps of Grey Wolf Optimization Algorithm

(1) Surrounding the prey

When the gray wolf finds the prey, the wolves will surround the prey, and let the distance between it and the prey be D, then there are:

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

$$D = |C \cdot X_P(t) - X(t)| \quad (1.)$$
$$X(t+1) = X_P(t) - A \cdot D \quad (2.)$$

Among them: X_p and X represent the position vectors of the prey and the gray wolf respectively; A and C are the coefficient vectors, and the formulas of A and C are as follows:

$$A = 2a \cdot r_1 - a \quad (3.)$$
$$C = 2r_2 \quad (4.)$$

Among them, $r_1$and $r_2$are random numbers between [0,1], $a = 2 - 2\frac{t}{t_{max}}$ is the convergence factor, and t_{max} is the maximum number of iterations.

(2) Hunting prey

After the gray wolf recognizes and surrounds the prey, the wolves will approach the prey and hunt it under the leadership of α, β, and δ. The specific formula is as follows:

$$D_\alpha = |C_1 X_\alpha(t) - X(t)| \quad (5.)$$
$$D_\beta = |C_2 X_\beta(t) - X(t)| \quad (6.)$$
$$D_\delta = |C_3 X_\delta(t) - X(t)| \quad (7.)$$
$$X_1(t+1) = X_\alpha(t) - A_1 D_\alpha \quad (8.)$$
$$X_2(t+1) = X_\beta(t) - A_2 D_\beta \quad (9.)$$
$$X_3(t+1) = X_\delta(t) - A_3 D_\delta \quad (10.)$$
$$X(t+1) = X_1(t+1) + X_2(t+1) + X_3(t+1). \quad (11.)$$

Among them: D_α、D_β and D_δ represent the distance between α、β and δ and other individuals respectively, X_α、X_β and X_δ represent the current position of α、β and δ respectively, A_i and C_i represent the coefficient vector.

(3) Attacking prey

When the prey is no longer moving, the gray wolf will attack the prey. When simulating the process of approaching the prey, the value of a gradually decreases, so the fluctuation range of A also decreases, and the value of A will affect the selection of the wolves: when |A|<1, the wolves move towards the prey Attack, when |A|>1, the gray wolf will separate from the current prey and search for the best prey again.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

(4) Searching for prey

At this stage, not only A will influence wolf pack selection, but C will also influence the search process. until the optimal solution is determined and the optimal solution is output.

3. Empirical Analysis

3.1 Data description

This paper uses a total of 303 sets of data. The data in this paper comes from the public data set of the kaggle platform. The variable information contained in it is shown in Table 1. The target is the output value of the study, that is, whether it is diseased. If it is diseased, the output value is 1, otherwise is 0.

Table 1 Summary of variable information

Variable name	Variable description	Variable type
age	age	Integer value
sex	gender (1=male, 0=female)	Classification
cp	Type of chest pain (0,1,2)	Classification
trestbps	resting blood pressure	Integer value
chol	cholesterol measurement	Integer value
fbs	Fasting blood sugar (whether >120mg/dl, 1=true, 0=false)	Classification
restecg	Resting ECG measurements (0=normal, 1=abnormal, 2=LVH)	Classification
thalach	maximum heart rate	Integer value
exang	Provoked angina (1=yes, 0=no)	Classification
oldpeak	Sports ST pressure	Integer value
slope	ECG tilt	floating point value
ca	Number of aortas	Classification
thal	Thalassemia syndrome	Classification
target	Output value (1 = diseased, 0 = not diseased)	Classification

3.2 Model Construction

The 303 sets of complete data are divided according to 3:1, that is, the first 226 sets of data are used as training sets, and the last 77 sets of data are used as test sets for model construction.

3.3 Analysis of model results

Use the random forest model optimized by gray wolf to train and fit the data, and compare it with the prediction results of decision tree and random forest through the following indicators commonly used

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

in binary classification problems. That is, the correct rate of the classification result is 1, so the patient population is set as the positive class, and the specific indicators are as follows:

Accuracy: It estimates the ratio of correctly grouped samples to the total samples. The closer the accuracy rate is to 1, the better the model effect is, and it is an indicator to measure the overall effect of the model prediction. Its formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12.)$$

Among them, TP is the number of the diseased population predicted to be the diseased population, FN is the number of the diseased population predicted to be the disease-free population, FP is the disease-free population predicted to be the number of the diseased population; TN is predicted to be the disease-free population as Number of people without disease.

Precision: It estimates the proportion of the correct number of results predicted as positive. The closer the accuracy is to 1, the better the model is. Its formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (13.)$$

F1-Score: It is the harmonic mean of precision and recall. The closer the F1-Score is to 1, the better the model predicts, which is defined as:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14.)$$

Where Recall is the recall rate, and its formula is: $Recall = \frac{TP}{TP + FN}$.

The performance of the above indicators in the three models is shown in Table 2. It can be seen that among the three models, the decision tree has the worst effect, while the random forest optimized by gray wolf has the best performance, and its accuracy is 5.2% higher than that of the random forest, which reflects its good overall prediction ability, and the Precision reaches 0.9231, which also shows its ability to distinguish the diseased population.

ISSN: 2582-6271

Vol. 3, Issue.4, July-Aug 2022, page no. 01-07

Table 2 Model Results Comparison

Model	Accuracy	Precision	F1-Score
Decision tree	0.8182	0.8095	0.8293
Random forest	0.8571	0.8919	0.8571
GWO-Random Forest	0.9091	0.9231	0.9114

4. CONCLUSION

This paper proposes a random forest model based on the gray wolf optimization algorithm to study the classification of heart disease, and compares it with the prediction results of decision tree and basic random forest. The results show that the random forest using the optimization algorithm has better prediction results. The accuracy rate is 5.2% higher than that of ordinary random forest, which provides a certain reference for the prediction and prevention of heart disease in the future.

REFERENCES

[1] Summary of China Cardiovascular Health and Disease Report 2020 [J]. Chinese Journal of Circulation, 2021, 36(06): 521-545.

[2] Zhu, Jia. The number of deaths from cardiovascular disease in my country has increased by 48% in 15 years [N]. Physician Journal, 2022-02-24(B01). DOI: 10.44211/n.cnki.nysbz.2022.000115.

[3] Liu, Qiao. Heart disease prediction based on clustering and XGboost algorithm [J]. Computer System Application, 2019, 28(01): 228-232. DOI: 10.15888/j.cnki.csa.006729.

[4] Sun, Yu. Research on classification and prediction of cardiac cases based on machine learning [J]. Computer Knowledge and Technology, 2021, 17(26): 96-97+104. DOI: 10.14004/j.cnki.ckt. 2021.2607.

[5] Zhao, Li, Wang, Zhang. Prediction Algorithm for Random Forest Heart Disease Based on Optimization [J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2021, 42(02): 112-118. DOI: 10.16351 /j.1672-6987.2021.02.016.

[6] Mirjalili S , Mirjalili S M , Lewis A . Grey Wolf Optimizer[J]. Advances in Engineering Software, 2014, 69(3):46–61.