Multilevel Risk Prediction of Cardiovascular Disease based on Adaboost+RF Ensemble Learning

Runchuan Li1,2 and Shengya Shen3 and Gang Chen1,2,4 and Tiantian Xie1 and Shasha Ji1 and Bing Zhou1,2 and Zongmin Wang1,2,a

1Industrial Technology Research Institute, Zhengzhou University, Zhengzhou Henan, 450000, China.
2Cooperative Innovation Center of Internet Healthcare, Zhengzhou University, Zhengzhou Henan, 450000, China.
3School of Foreign Languages, Zhengzhou University, Zhengzhou Henan, 450000, China.
4School of Distance Learning, Zhengzhou University, Zhengzhou Henan, 450000, China.

gchen@ha.edu.cn or zmwang@ha.edu.cn

Abstract. Background: In the field of diagnostic CVD, the predecessors used a large amount of data with no missing two-category data, and obtained good results. However, in the process of electronic input of historical data, a large number of data attribute values are missing, and there are multiple levels of disease risk. Goal: On the data set of imbalance and a large number of missing values, this paper focuses on the five levels of cardiovascular disease. Methods: A new prediction model of Adaboost+RF is constructed by using the information gain ratio to analyze the feature contribution degree of the data set. The performance of this model is evaluated with Precision, Recall, F-measure and ROC Area values. Results: The results show that the four key indicators of the Adaboost+RF model on five-categories unbalanced datasets in Precision, Recall, F1 and AUC values, which are 40.9%, 49.3%, 41.4% and 71.6%. Conclusion: The experiment results demonstrate that the four key indicators of the Adaboost+RF model on five-category unbalanced missing datasets are better than other machine learning algorithms.

1 Introduction

With the improvement of individual’s living standards, changes in living habits, the deepening of the aging society and the risk factors for cardiovascular disease (CVD) are becoming increasingly. At present, cardiovascular disease (CVD) has become one of the ruthless killers in individual’s daily life. According to the data released by China Cardiovascular Disease Report 2016 in June 2017, cardiovascular death is the main cause, 45.01% in rural areas and 42.61% in urban areas, and the mortality of CVD is increasing [1-2]. It is estimated that by 2030, due to the increase of hypertension, obesity, diabetes and an aging population, the number of deaths caused by cardiovascular diseases will increase to 23.6 million. Early diagnosis and early intervention of cardiovascular disease is recognized as the most effective way to reduce cardiovascular morbidity and mortality [3-4]. The American Society of Cardiology/American Heart Association (ACC/AHA) proposed concepts related to
cardiovascular disease risk assessment and early warning, pointing out that it can be based on established risk factors to predict future CVD risks, and the importance of these risk factors for CVD has been confirmed in many CVD risk prediction tools [5]. According to the identified risk factors, the disease can be detected early. The management of the disease in stages and the corresponding preventive interventions is also crucial for the diagnosis of the disease, which can be diagnosed and treated early to reduce the incidence and disability of CVD.

Machine learning [6] is widely used in medical diagnostics. With the deepening of the application of machine learning in the medical field, more and more machine learning applications have emerged in CVD diagnostics. For example, Shameer K et al. [7] examine the potential limitations and challenges of implementing machine learning method in the context of cardiovascular medicine. El-Saadawy et al. [8] used Probabilistic Neural Network (PNN), Support Vector Machine (SVM) and SoftMax regression classifiers to diagnose heart disease. Ade M R R et al. [9] proposed a cardiac classifier and explained the application of SVM and Naive Bayes (NB) in the classification of heart diseases. Li R et al. [10-11] proposed using support vector machines and random forests to establish disease risk prediction models to improve the accuracy of traditional predictive models on processed data set. Papers [8] [9] [10] [11] all use SVM classifiers. However, the SVM classifier has no general solution to nonlinear problems, and the SVM is only good at handling the two-class problem. Paper [12] used logistic regression (LR) analysis to determine risk factors for cardiovascular disease. However, the LR classifier is easy to under-fitting, and it can only deal with linear and two classification problems. Saško Ristov et al. [13] used the international CVD risk factor analysis machine learning methods, and proposed a new method for prevention and early detection of cardiovascular diseases in sample data. Jabbar M A et al. [14] proposed a new effective associative classification algorithm combining k-NearestNeighbor (KNN) and Genetic Algorithms (GA). The programming implementation of KNN and GA algorithm is more complicated and computationally intensive. In addition, they are sensitive to outliers and they are vulnerable to unbalanced data. Dominic V et al.[15] proposed how heart disease datasets and machine learning techniques can help understand the level of risk associated with heart disease. Das R et al. [16] used SAS software to build a diagnostic method for heart disease based on neural network integration. Chen M et al. [17] proposed a new multi-model disease risk prediction algorithm based on convolutional neural network (CNN) to effectively predict the occurrence of chronic diseases. Paper [16] and [17] are both based on neural networks. Neural networks is suitable for data sets with large data volumes. The computational cost is high, and its biggest disadvantage is the "black box" nature, in other words, there is no way to explain reasoning process basis. This paper uses Adaboost+RF algorithm to classify, the classification accuracy is high, and the generalization error is low.

Most of the above prediction models used the two-category data set with high standards without missing values to continuously explore and research in the field of CVD and obtained high accuracy results. However, due to the incompleteness of data information, the possibility of data loss is often caused during the electronic entry of the history form. Therefore, in response to this problem, we combined random forest and adaboost algorithm from the perspective of a basic tree-based unit to construct a new prediction model and apply it to the CVD field. In this paper, we establish a CVD risk prediction model based on Adaboost+RF ensemble learning and analyze the feature contribution degree of the data set using the information gain ratio. The innovations in this paper are as follows:

On the one hand, this paper uses the information gain ratio to analyze the feature contribution of existing CVD sample data to mine disease risk factors and improve the accuracy of identifying CVD, instead of establishing risk warnings based on the identified possible risk factors through experiments model.

On the other hand, due to the importance of prevention of disease grading, this paper selects random forests as weak classifiers, combined with adaboost algorithm to build a new prediction model, using multiple unbalanced sample with a large number of missing values for experimental studies to improve the accuracy of multilevel prediction of CVD risk.
The outline of this paper is as follows: In the second part, the pretreatment and detailed description of the original sample data and the analysis of feature contribution using information gain ratio are introduced. The third part introduces Adaboost+RF ensemble learning in detail. The fourth part analyzes and evaluates the experiment, including the selection of evaluation index, the results of different classification of Adaboost+RF ensemble learning classifier on five-categories unbalanced sample and the influence of different random factors on the classification accuracy of Adaboost+RF ensemble learning classifier. Finally, the fifth part summarizes this paper and discusses the future work.

2. DATA PROCESSING AND ANALYSIS

2.1 Data processing

This experiment selects the data set disclosed in the UCI Machine Learning Library: Heart Disease Datasets. The total number of Heart Disease Datasets are 920 [18-19]. It is a five-categories data set: no heart disease is represented by 0, and the severity of cardiovascular disease is represented by 1-4 different levels. We need to transform the attribute types of the data set. WEKA 3.8 is a very useful software for data analysis modeling, but sometimes the algorithm modeling used requires very strict data requirements. Numeric data is used to establish a regression model, and it is used to establish a classification model. Therefore, before the establishment of the disease prediction model, the attribute format of the data set is converted from numeric to nominal. Figure 1 shows the process of classification using the Adaboost+RF algorithm.

![Adaboost+RF algorithm flowchart](image)

**Figure 1.** Adaboost+RF algorithm flowchart.

2.2 Contribution analysis of the features

For ease of description, the definition of entropy and conditional entropy is given first, and entropy is a measure of the uncertainty of a random variable. Let $X$ be a discrete random variable taking a finite number of values with a probability distribution of

$$P(X = x_i) = p_i, \ i = 1, 2, ..., n$$

Then the entropy of random variable $X$ is defined as

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

(2)

With random variables $(X, Y)$, the joint probability distribution is

$$P(X = x_i, Y = y_j) = p_{ij}, \ i = 1, 2, ..., n, \ j = 1, 2, ..., m$$

(3)
Under the condition given by the random variable $x$, the conditional entropy of the random variable $y$ is defined as the mathematical expectation of the entropy of the conditional probability distribution of $y$ under $x$ given conditions for $x$:

$$H(Y|X) = \sum_{x_i} p(x_i) H(Y|x_i)$$  \hspace{1cm} (4)

The information gain represents the degree of uncertainty of the class $y$ information by knowing the information of the feature $x$. The information gain $g(D, F)$ of the feature $F$ to the training data set $D$ is defined as the difference between the empirical entropy $H(D)$ of the set $D$ and the empirical conditional entropy $H(D|F)$ of the $D$ under the given condition of the feature $F$.

$$g(D, F) = H(D) - H(D|F)$$  \hspace{1cm} (5)

Information gain ratio is based on the information gain multiplied by a penalty parameter. When the number of features is large, the penalty is small; when the number of features is small, the penalty is larger.

$$R_S(D, F) = g(D, F) / H(D)$$  \hspace{1cm} (6)

The penalty parameter indicates the data set $D$ with feature $F$ as the inverse of the entropy of the random variable. The feature information gain ratio is calculated by Eq. (6), and the feature contribution degree analysis is performed according to the feature information gain ratio. The ranking results obtained are shown in Table 1. It is further illustrated whether cardiovascular disease is related to information such as exercise-induced angina, level of chest pain, serum steroid content, and gender. Table 1 shows contribution analysis of the features. At the same time, we also analyze the six top-ranked feature subsets $Heart_{sec 1 2 6}$ and all the feature subsets $Heart_{all}$ on the Adaboost+RF algorithm respectively. The experimental results are shown in Table 2. As can be seen from the data in Table 2, the indexes of Adaboost+RF classifier are 39.5%, 46.4%, 41.7% and 68.6% respectively in the Heart$_{sec 1 2 6}$ section. The three indicators in the feature subset Heart$_{all}$ are higher than the Heart$_{sec 1 2 6}$. According to the results, it can be seen that the ranking of the feature items behind is also crucial for improving the accuracy of cardiovascular disease prediction.

**Table 1.** Contribution analysis of the features.

| ID | Attribute | Description |
|----|-----------|-------------|
| F1 | Exang     | exercise induced angina |
| F2 | Cp        | chest pain type |
| F3 | Chol      | serum cholesterol in mg/dl |
| F4 | Sex       | sex |
| F5 | Oldpeak   | ST depression induced by exercise relative to rest |
| F6 | Thalach   | maximum heart rate achieved |
| F7 | Trestbps  | resting blood pressure |
| F8 | Age       | age in years |
| F9 | Slope     | the slope of the peak exercise ST segment |
| F10| Thal      | normal; fixed defect; reversible defect |
| F11| Fbs       | fasting blood sugar > 120 mg/dl |
| F12| Ca        | number of major vessels (0-3) colored by fluoroscopy |
| F13| Restecg   | resting electrocardiographic results |

**Table 2.** The influence of different feature subsets on the results of Adaboost+RF algorithm.

| Data       | Precision | Recall | F1 | AUC  |
|------------|-----------|--------|----|------|
| Heart$_{sec 1 2 6}$ | 0.395     | 0.464  | 0.417 | 0.686 |
| Heart$_{all}$     | 0.401     | 0.490  | 0.411 | 0.710 |
3. METHODS

After the above analysis of the sample data, the existing machine learning algorithms can be used to establish five-categories CVD risk prediction models. First of all, we use WEKA 3.8 data analysis modeling software to convert the data set's attribute type, from numeric to nominal. Second, we use the information gain ratio to analyze the feature contribution degree of the data set. Finally, a 10-fold cross validation method is used to evaluate the performance of the predictive model. As an important part of disease risk prediction, the following briefly describes the principles of these two algorithms. Figure 2 shows the structure of the Adaboost+RF algorithm. The training samples of m samples are trained and weighted. Then, combined with random forest, Adaboost+RF algorithm is used for risk prediction.

![Figure 2. Adaboost+RF algorithm structure diagram.](image)

3.1 Adaboost+RF ensemble learning

Random Forest (RF) [20] is an integrated classification algorithm proposed by L. Breiman in 2001. The algorithm uses a decision tree model as a basic classifier. It adopts bootstrap resampling method to extract multiple samples from the original sample, performs separate decision tree modeling for each bootstrap sample, and then combines the predictions of multiple decision trees to determine the final category of the sample. Then, all decision trees are allowed to vote and the final forecast results are obtained by voting. This paper uses a C4.5 decision tree to form a base classifier for a random forest.

**Algorithm:** Adaptive Boosting

**Input:** $S$: A set of training tuples of class markers;

$K$: The number of rounds (the number of rounds is the number of classifiers (classification schemes). The $K$ round classification is equivalent to the classification of the $K$ classifier, which is equivalent to the $K$-classification scheme of the training set); A basic classification algorithm;

**Output:** A combined classifier model: contains $K$ classifiers, and the weight of the voting rights of the classifier;

1. Initialize the weight of each tuple in $S$ to $1/n$;
2. for $k = 1$ to $K$ do
3.   According to the weight of the tuple, the sampling is put back from $S$, and $S_k$ ($K$ training set) is obtained;
4.   Deriving (training out) the classifier $M_k$ ($K$ classifier) according to $S_k$;
5.   Calculate the error rate of $k$ according to formula (8);
6.   if $e_k > 0.5$ then
7.     Go to step 3;
8.   end if
9.   Calculate the weight of the voting rights of the classifier according to formula (9);
10. for each of $S_k$'s correctly classified tuples do
11.     Update the weight of the correctly classified tuple;
12. Normalize the weight of each tuple;
13. end for
14. end for

The Adaboost algorithm [21] is a lifting algorithm proposed by Freund and Schapire in 1995. It is the most representative algorithm in the lifting algorithm. Like the random forest algorithm, Adaboost algorithm is only
an algorithm framework that must have a basic learning algorithm as a support. The algorithm starts from the base classification algorithm, and iteratively obtains a series of base classifiers, and then combines these base classifiers to build a final strong classifier. During each iteration, the distribution of the weights of the training samples is constantly changed so that the next iteration is more concerned with the samples that are wrongly assigned, that is, giving higher weights to the misclassified samples. According to the classification effect of each base classifier, it is given its weight, so that the base classifier with a small classification error rate has a higher weight, and the error rate has a lower weight. Assuming a training data set $D = \{(x_i, y_i), ..., (x_N, y_N)\}$ for five-classification, in which sample instance $x_i \in \mathbb{R}^d$ and class label $y_i \in \{0, 1, 2, 3, 4\}$ are categorized using the Adaboost algorithm, the steps are as follows:

**Step 1:** initialize training data sample weight distribution: $W_t = (w_{t,1}, ..., w_{t,i}, ..., w_{t,N})$, where, $w_{t,i} = 1/N$, $i = 1, 2, ..., N$.

**Step 2:** Perform $k$ iteration, where $k = 1, 2, ..., K$.

During each iteration, training is performed on the training data set with weight distribution $w_t$ to obtain a base classifier:

$$G_t(x) : x \rightarrow \{0, 1, 2, 3, 4\}$$

(7)

Calculate the classification error rate of this base classifier $G_t(x)$ in this iteration:

$$e_t = \frac{1}{N} \sum_{i=1}^{N} w_{t,i} I(G_t(x) \neq y_i)$$

(8)

From Eq. (8), we can see that the error rate $e_t$ of $G_t(x)$ on the training dataset is the sum of the weights of $G_t(x)$ misclassified samples. Calculate the weight of the base classifier during this iteration according to Eq. (8).

$$\delta_t = \frac{1}{2} \log \left( \frac{1}{e_t} \right)$$

(9)

From Eq. (9), we can see that when $e_t = \frac{1}{2}$, $\delta_t \approx 0$, and $\delta_t$ increases with the decrease of $e_t$, it means that the smaller the classification error rate, the greater the effect of the basic classifier in the final classifier. Then update the weight distribution of the training data set: $W_{t+1} = (w_{t+1,1}, w_{t+1,2}, ..., w_{t+1,i}, ..., w_{t+1,N})$, where $G_t, G_{t-1}, ..., G_1, i = 1, 2, ..., N$.

**Step 3:** After $K$ iterations, $K$ base classifiers $G_1, G_2, ..., G_{K-1}, G_K$ can be obtained, then the $K$ base classifiers can be linearly combined to obtain the final integrated classifier $G(x) = arg\left( \sum_{k=1}^{K} \delta_k G_k(x) \right)$.

**4. EXPERIMENT**

**4.1 Experiment and result analysis**

In this experiment, we select six typical classification algorithms, establish a risk prediction model in a given training set, and verify the efficiency of the Adaboost+RF algorithm on five-categories data set through analysis and comparison of experimental results. We compare Support Vector Machine (SVM), Logistic Regression (LR), Random Tree (RT), Decision Tree (DT), Random Forest (RF), and Adaboost+RF differences in the performance of the classifier over five-categories unbalanced sample. We use the SVM algorithm in paper [11] and the LR algorithm in paper [13] to compare with the experiments in this paper. The SVM algorithm is marked as Baseline-1, and the LR algorithm is marked as Baseline-2. The experimental results with various classifiers are shown in Table 3, and the evaluation index values of different classifiers are compared as shown in Figure 3. The five-categories data set is divided into Value 0, Value 1, Value 2, Value 3, and Value 4. According to all the index values, it can be seen that the Value 0 has a higher sensitivity and the largest area. Since Value 0 has a large amount of data, there are 411 samples and 337 are correctly classified, so Value 0 has the best performance. The number of samples 1, 2, 3, and 4 are 196, 135,
Due to the large number of missing values in the sample data, their correct classification rate is lower than the value 0.

| Classifier   | Precision | Recall  | F1   | AUC   |
|--------------|-----------|---------|------|-------|
| Baseline-1   | 0.286     | 0.449   | 0.281| 0.503 |
| Baseline-2   | 0.390     | 0.332   | 0.352| 0.605 |
| RT           | 0.364     | 0.418   | 0.385| 0.607 |
| DT           | 0.384     | 0.488   | 0.416| 0.681 |
| RF           | 0.389     | 0.482   | 0.401| 0.706 |
| Adaboost+RF  | 0.409     | 0.493   | 0.414| 0.716 |

From the data in Table 2 and Table 3, it can be seen that when dealing with a five-categories unbalanced sample with a large number of missing values, the baseline-1, baseline-2 classification results are poor compared to Adaboost+RF, especially for baseline-1, its three key indicators are lower than other machine learning algorithms. The indicators of the Adaboost+RF classifier are better than those of other classifiers, with the highest Precision value of 40.1%, Recall value of 49%, F-Measure value of 41.1%, and ROC Area value of 71%. In the Adaboost+RF algorithm, different random factors (s) also have different effects on the experimental results. From the experimental results, it can be seen that Adaboost+RF algorithm performs best when the stochastic factor is 16, and the values of Precision, Recall, F1 and AUC value of the algorithm are the highest, which are 40.9%, 49.3%, 41.4% and 71.6%. For different random factors, the highest F1 value reached 41.4%, the lowest reached 39.8%, the highest ROC Area value reached 71.6%, the lowest reached 70.5%, with little change. Therefore, it is appropriate to apply Adaboost+RF algorithm to solve the problem. Therefore, it can be seen that the Adaboost+RF algorithm performs better, and the accuracy of CVD prediction can be further improved by using this algorithm. The reason for the analysis is as follows: In such a five-categories data set with a large number of missing values and unbalanced data, the Adaboost+RF algorithm is superior. It is a strong classifier obtained by combining Adaboost algorithm with a weak classifier random forest that is relatively simple and fast in training. It can iteratively fit even noisy data sets until the data points of each training set adapt to error-free, and it can also adapt to the respective training error rate of weak classifiers, effectively reducing the errors caused by unbalanced dataset.

5. CONCLUSION

In recent years, the mortality caused by CVD is still increasing, which has caused widespread concern in the health sector. The pathogenesis of CVD is complex and changeable, it is closely related to individual's bad behavior, such as smoking, unhealthy eating habits, long sedentary and other factors. Therefore, in the future work, we will first collect and analyze more different Occupation populations of data to extract more hidden features to more accurately assess and predict CVD risk. Second, mobile medical care is the future direction of medical development. We will develop mobile applications, provide better multilevel risk prediction services, and conduct daily health management of residents to reduce the incidence of CVD in patients.
6 COMPETING INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this paper.

ACKNOWLEDGMENT

This work is supported by the following fund projects: National Key Research and Development Program of China (No.2017YFB1401200); Integration of Cloud Computing and Big Data, Innovation of Science and Education (No.2017A11017); Purcell Network Next Generation Internet Technology Innovation Project (No. NGII20161202, No. NGII20170716); Key Science and Technology Projects of Henan (No.152102210249); Key Scientific Research Projects of Colleges and Universities in Henan (No.18A520049).

REFERENCES

[1] Prochaska, J. H., et al. Herz,"Prevention of cardiovascular diseases." 43.1:87-100(2018)
[2] Chen, W. W., et al. J Geriatr Cardiol, "China cardiovascular diseases report 2015: a summary;" 14.1:1-10(2017)
[3] Hood, L., R. Balling, and C. Auffray. Biotech J,"Revolutionizing medicine in the 21st century through systems approaches." 7.8:992-1001(2012)
[4] Hunter, Peter J., and Thomas K. Borg. NAT REV MOL CELL BIO,"Integration from proteins to organs: the Physiome Project." 4.3: 237 (2003)
[5] Weng, Stephen F., et al. PloS one , "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." 12.4: e0174944 (2017)
[6] Deo, Rahul C. Circulation, "Machine learning in medicine." 132.20 : 1920-1930 (2015)
[7] Shameer, Khader, et al. Heart, "Machine learning in cardiovascular medicine: are we there yet?." : heartjnl-2017 (2018)
[8] El-Saadawy, Hadeer, et al. "Electrocardiogram (ECG) heart disease diagnosis using PNN, SVM and Softmax regression classifiers." Intelligent Computing and Information Systems (ICICIS), 2017 Eighth International Conference on. IEEE, (2017)
[9] Ade, Ms RR, Dhanashree S. Medhekar, and Mayur P. Bote. INT J ENG SCI ,"Heart disease prediction system using svm and naive bayes." 2.5 (2013)
[10] Li, Runzhi, et al. IEEE International Conference on Bioinformatics and Biomedicine IEEE,"Multi-label classification for intelligent health risk prediction." 986-993,(2017)
[11] Li, Runzhi, et al. J HEALTHC ENG ,"An Ensemble Multilabel Classification for Disease Risk Prediction." , 2017,(2017-6-15), 2,1-10,2017
[12] Kilkenny, Monique F., et al. PloS one , "Knowledge of risk factors for diabetes or cardiovascular disease (CVD) is poor among individuals with risk factors for CVD." 12.2 : e0172941 (2017)
[13] Ristov, Saško, and Aleksandar Peckov. "Machine Learning Approach for Early Detection of Cardiovascular Deceases (CVD)." (2010)
[14] abbar, M. Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. "Heart disease prediction system using associativ classifcation and genetic algorithm." arXiv preprint arXiv:1303.5919 (2013)
[15] Dominic, Vinitha, D. Gupta, and S. Khare. Appl Med Inform, "An effective performance analysis of machine learning techniques for cardiovascular disease." (2015).
[16] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. EXPERT SYST APPL , "Effective diagnosis of heart disease through neural networks ensembles." 36.4: 7675-7680 (2009)
[17] Chen, Min, et al. IEEE Access,"Disease Prediction by Machine Learning over Big Data from Healthcare Communities." ,PP,99:1-1, (2017)
[18] LICHMANM, "UCI machine learning repository" in http://archive.ics.uci.edu/ml, Accessing 2015.
[19] UCI. Heart Disease Data Set[EB/OL]. http://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[20] Wager, Stefan. "Asymptotic theory for random forests." arXiv preprint arXiv:,1405.0352 (2014)
[21] Zhang, Yang, and Z. Zhao. International Congress on Image and Signal Processing, Biomedical Engineering and Informatics IEEE,"Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost." (2018)