A Classification Method of Arrhythmia Based on Adaboost Algorithm

Bing Zhang\textsuperscript{1,2,}\textsuperscript{*}, Jingye Wen\textsuperscript{1,2}, Huihui Ren\textsuperscript{1,2}

\textsuperscript{1} School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China;
\textsuperscript{2} The Key Laboratoy of Software Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China;

Email of all the authors: bingzhang@ysu.edu.cn; vanilla1023@sina.cn; 2456806242@qq.com;

* Corresponding Author: Bing Zhang; email: bingzhang@ysu.edu.cn; phone: 188-3046-4932.

Abstract: An arrhythmia classification model based on an adaptive boosting algorithm is proposed in this paper. According to the AAMI standard, 15 kinds of abnormal cardiac rhythms are grouped and the datasets are segmented by the non-crossover method. The electrocardiogram (ECG) signals are denoised by the filter method, and then divided into fixed-length ECG beats, and five features are extracted from time-domain and frequency-domain. Then, the base classifier of the algorithm and its optimal algorithm parameters is selected to realize the multi-classification of cardiac anomalies, aiming at mining hidden knowledge from human physiological data to detect human health status, making the diagnosis process more automatic, efficient, and intelligent.

1. Introduction

At present, the medical mechanism of arrhythmia is not very clear. Arrhythmia is characterized by a slow heartbeat, too fast, or irregular. There are numerous specific types of arrhythmia [1]. From the point of view of whether it is a threat to human health, arrhythmia can be divided into two groups: a serious threat and a potential threat. Patients with acute threatening arrhythmias need to be diagnosed and treated in time, and potentially threatening arrhythmias require patients to pay attention to health care and regular examination, so as not to worsen the disease and bring serious health problems. There are numerous ways to check abnormal heart rhythms. The most commonly used is the electrocardiogram (ECG) [2]. Doctors carefully observe the ECG signal with the naked eye, and then diagnose the patient whether there is an abnormal heart rhythm and what kind of abnormality based on personal clinical experience and professional knowledge. The traditional arrhythmia diagnosis process is time-consuming and labor-intensive, resulting in a waste of medical resources, and it is difficult to satisfy the needs of real-time arrhythmia detection.

In the past 20 years, a lot of research on the classification of arrhythmia based on machine learning methods has emerged in the medical field. Chazal et al.[3] extracted the morphological characteristics of the ECG beat, including four RR intervals, a QRS complex duration, P-wave duration, T-wave duration, and four sets of under-sampling of different bands (P, QRS, T-wave). Then the linear discriminator (LD) was utilized to solve the problem of arrhythmia classification, improved the
Likelihood Function of the linear discriminant model, and improved the classification effect of the model on the imbalanced datasets through the weighted likelihood function. Mar et al. [4] mainly extracted the morphological features of the ECG, and applied the sequential forward floating search (SFFS) algorithm for feature selection. Then the multilayer perceptron (MLP) model was used to classify the heart rhythm anomalies, and the classification effect of the LD model was compared. In the study of Li et al. [5], the combination of wavelet packet entropy feature and random forest model was used for the first time, and the dataset was established according to the arrhythmia grouping standard recommended by AAMI [6] and the non-intersection mode proposed in the paper. Yan Haolin [7] proposed a feature extraction method of ECG beat based on a convolutional neural network (CNN). The fully connected input of the final layer of CNN was taken as the feature of ECG beat, and all the features were finally input into an MLP classifier. Lu Tancheng et al. [8] extracted the RR interval characteristics, QRS waveform features, and discrete cosine transform features in ECG beat. Considering the real-time requirements in practical application, the minimum Euclidean distance classifier with minimal complexity was used to judge the abnormal ECG signal. Zhao Yong et al. [9] performed four-scale wavelet transform on ECG signals, calculated the maximum, minimum, average, and standard deviation of four scale detail wavelet coefficients (D1-D4) and the fourth scale subband wavelet coefficients (A4) as the characteristics of ECG beat, and selected the optimal feature subset based on the principle of maximum correlation and minimum redundancy, and finally established the classification model of arrhythmia based on support vector machine. Zhang Ruyi et al. [10] performed five-layer wavelet decomposition of the ECG signal, selected the data points from 80 milliseconds before R wave (R-80) to 12 milliseconds after R wave (R+120) in the third and fourth layers of wavelet coefficients, and took the variance and expectation as well as the expectation and variance of the autocorrelation function of the data as the characteristics of ECG beat. The method of clustering, linear classifier, and doctor-assisted was used to classify the arrhythmia accurately. Andreao et al. [11] performed a multi-scale wavelet transformation on the ECG beat, selected three-scale under-sampling coefficients as the characteristics of the ECG signal, and established a neural network classification model of arrhythmia. Lannoy et al. [12] extracted five groups of morphological features with different properties and established a support vector machine model to classify arrhythmia, including six kinds of RR intervals, and calculated 24 features from different bands (including the presence or absence of P, QRS, T waves, duration, maximum, minimum, waveform area, standard deviation, kurtosis, and skewness) and the corresponding 24 features of the normalized version (each feature is standardized using the feature mean of each patient). The other two sets of features are Higher-Order Statistical (HOS) parameter defined by Osowski et al [13] and the Hermite Basis Functions (HBF) parameter defined by Park et al. [14].

In this paper, a multi-classification method of arrhythmia (including normal heart rhythms) based on the Adaboost algorithm is proposed. The abnormal heart rhythm study used the MIT-BIH public dataset. First of all, according to the recommendation of AAMI, 15 types of heart rhythm are divided into 5 categories, and the training and test sets are divided according to inter method to ensure that the research conclusions are of practical significance. Secondly, denoising, segmentation of ECG beat, and feature extraction are performed on the dataset in sequence. In order to obtain more categorized ECG characterization information, five features in time-domain and frequency-domain are extracted from each ECG beat. Finally, a classification tree-based Adaboost arrhythmia classification model was designed, and the classification effect of the model with different number and depth of classification trees was analyzed, and 100 classification trees with a depth of 4 were selected as the base classifier of the Adaboost model. In order to verify the reliability of the method, the previous experimental results were compared, and 4 classification evaluation criteria were used to verify the accuracy of arrhythmia classification.
2. Evaluation index

The evaluation index of arrhythmia classification refers to the recommendations of AAMI and the practices of similar studies. We use the confusion matrix shown in Table 1 to represent the classification effect of the model, and count the accuracy, precision, and recall rates.

Table 1. Confusion matrix.

|                    | Model to predict heart rhythm | Sum  |
|--------------------|-------------------------------|------|
| True heart rhythm type | N     | S     | V     | F     | ΣN   |
| N                  | Nn    | Ns    | Nv    | Nf    | ΣN   |
| S                  | Sn    | Ss    | Sv    | Sf    | ΣS   |
| V                  | Vn    | Vs    | Vv    | Vf    | ΣV   |
| F                  | Fn    | Fs    | Fv    | Ff    | ΣF   |
| Sum                | Σ     | Σ     | Σ     | Σ     | Σ    |

The direction of the matrix row represents the true type of ECG beat and the direction of the column represents the type of ECG beat predicted by the model. For each of the heart rhythm types, the number of correct predicted values of the statistical model is TP, and the number of incorrect predictions is FP; the actual total amount of the other three types of heart rhythm is recorded as TN, and the total number of heart rhythm type incorrectly classified as the other three types by the model is recorded as FN. The calculation formulas of precision rate (precision), recall rate, and false positive rate (FPR) are as Equation (1)–(3).

\[
\text{precision} = \frac{TP}{TP + FP} \times 100\%
\] (1)

\[
\text{recall} = \frac{TP}{TP + FN} \times 100\%
\] (2)

\[
FPR = \frac{FP}{TN + FP} \times 100\%
\] (3)

For N-type ECG beats, the corresponding TP, FP, TN, FN are recorded as TPn, FPn, TNN, FNN, respectively, and their specific calculations are TPn = Nn; FPn = Sn + Vn + Fn; TNN = Ss + Sv + Vn + Vf + Fs + Fv + Ff; FNN = Ns + Nv + Nf.

3. Adaptive boosting (Adaboost) algorithm

3.1 Analysis of arrhythmia classification process

In this paper, the use of machine learning to solve the problem of abnormal heart rhythm classification mainly includes the following steps: preprocessing, datasets partitioning, model training, model prediction, and model evaluation. The whole process can be shown in Figure 1.
Figure 1. Arrhythmia classification process.

(1) First, preprocess the abnormal heart rhythm dataset to remove baseline drift and EMG interference noise in the ECG signal, and divide the continuous ECG signal into ECG beats of uniform length.

(2) Then perform feature extraction on the ECG beat. In this paper, five kinds of features in time-domain and frequency-domain are extracted from ECG beat and spliced into feature vectors with a length of 186 dimensions.

(3) Then the dataset is relabeled according to the type of arrhythmia recommended by AAMI, and the training set and test set are divided according to the non-crossover mode.

(4) The data processed through the above steps is used as the next model input. In this paper, the Adaboost algorithm in ensemble learning is used to deal with the multi-classification of arrhythmia, and the training set is used for model parameter optimization and model training. Finally, 100 classification trees with a depth of 4 are selected as the base classifier.

(5) In this paper, four classification evaluation indexes are utilized to detect the classification effect of the trained model on the test set, and we compare the previous research results.

3.2 Adaboost ensemble learning algorithm

Ensemble learning is a very popular machine learning method, which can be used to solve classification problems, regression problems, feature selection, anomaly detection, and other problems. The biggest difference between ensemble learning and other typical machine learning algorithms is that instead of using a single machine learning algorithm, it implements the learning process through the cooperation of multiple machine learning algorithms. Each machine learning algorithm in integrated learning is called a base learner. Multiple base learners in the integrated learning algorithm can be the same algorithm (homogeneous learner) or different algorithms (heterogeneous learner). Homogeneous learners are relatively more extensive. According to different cooperation models between base learners, integrated learning algorithms can be divided into two categories. One kind of base learner has no strong dependency relationship and can be trained in parallel. While the other kind of base learner has a strong dependency relationship, and the base learner needs serial training. The representatives of these two categories are bagging series algorithms and boosting series algorithms. Adaboost algorithm is a representative boosting algorithm.

In the learning process of the Adaboost lifting method, by changing the weight distribution of the training data, multiple rounds of training are carried out for different training data, and finally, multiple base learners are obtained, which are combined with the classification results of multiple base classifiers as the classification results of Adaboost algorithm. During the learning process, Adaboost increases the weight of the samples misclassified by the previous round of the base classifier in the
training set, while reducing the weight of the correctly classified samples in the training set. In this way, the training dataset required for the next round of training depends on the classification result of the previous classifier, and there is a "serial cooperation" relationship among multiple base learners. Adaboost uses a weighted voting method to count the prediction results of all base classifiers to obtain the final classification result of the model. For a base classifier with a larger classification error, a smaller weight is assigned to weaken its role in voting; for the base classifier with a better classification effect, it is given higher weight to strengthen its role in voting. The arrhythmia dataset processed in this paper has a significant difference in the number of arrhythmia types. In the process of multi-round training using the Adaboost algorithm, the weight of the samples in the training set can be changed, and the weight of a small number of abnormal arrhythmia types can be increased to solve the problem of dataset imbalance. Besides, this article uses a more widely used homogeneous learner, because many previous studies are based on classification trees and random forests to classify abnormal heart rhythms, and good classification results have been achieved. Therefore, the classification tree model with depth 4 is finally selected as the base classifier, and the number of base classifiers is set to 100. The specific modeling process is as follows.

1) Divide dataset reasonably

To ensure the application value of the research conclusions to the actual arrhythmia classification, this paper re-marks the tags in the arrhythmia database according to the type groups recommended by AAMI and divides the training set $DS_1$ and test set $DS_2$, according to the non-crossover mode to ensure that the ECG data trained and predicted by the model come from different people. After denoising, segmentation, and feature extraction, the ECG datasets $DS_1$ and $DS_2$ can be expressed as \[ \{(\mathbf{X}_1, y_1), \ (\mathbf{X}_2, y_2), \ldots, \ (\mathbf{X}_N, y_N)\}. \] $\mathbf{X}_i$ represents the feature vector, which is a feature vector of 186 dimensions that are joined by five features of $RRs$、$S_K$、$R_T$、$WVL$ and $WPD$; $Y_i$ is the classification tag, including N、S、V、F four types of arrhythmia tags.

2) Select the base classifier parameters of the model

In this paper, the classification tree is used as the base classifier to analyze the change of model classification error when the number and depth of classification trees are used as the base classifier of the AdaBoost algorithm. Abnormal heart rhythm classification capabilities of different parameter configuration models are shown in Figure 2 below. In the Figure 2, d1-d9 represents classification trees with a depth of 1-9, and the horizontal axis represents the number of classification trees. We can see that the error rate of the model decrease with the depth of the classification tree. When the depth of the classification tree increases in the range of 1-4, the error rate of the model decreases obviously; when the depth of the classification tree continues to increase in the range of 4-9, the error rate of the model remains unchanged. When the number of classification trees increases in the range of 0-100, the error rate of the model decreases greatly; when the number of classification trees continues to increase in the range of 100-400, the error rate of the model remains stable. Therefore, 100 classification trees with a depth of 4 are selected as the base classifiers of the AdaBoost algorithm.

3) Adaboost model training and prediction

The dataset $DS_1$ is used to train the Adaboost abnormal heart rhythm classification model containing 100 classification trees with a depth of 4, and the classification effect of the model is verified on the dataset $DS_2$ through four classification evaluation criteria. The model training process of Adaboost is shown in Algorithm 3.1.
Figure 2. The error rate of the model varies with the number and depth of classification trees.

**Algorithm 3.1** Adaboost arrhythmia classification model establishment algorithm  
Input: arrhythmia training set $DSI$.  
Output: arrhythmia classification model.  
1) Initialize 100 classification trees with a depth of 4 and the weight of each sample in the training set $DSI$ as $W=[1/(\text{length}(DSI)),\: 1/(\text{length}(DSI)),\: ...]$ Where $W$, is a vector of length $n$, $n=1/(\text{length}(DSI))$, $i$ corresponds to the serial number of the base classifier classification tree.  
2) for m in 100:  
3) Using the training dataset with the weight distribution of $W_m$ to learn the m-th basic classifier $G_m(x)$.  
4) Calculating the classification error of classifier $G_m(x)$.  
5) Calculate the weight coefficient $\alpha$ of the m-th base classification tree to the final classification result of the Adaboost model.  
6) Obtaining the weight vector set $W_{m+1}$ corresponding to the data for training the next base classifier  
7) End  
8) Combine the classification results of 100 base classifiers and the corresponding weight coefficient $\alpha$ to compute the final classification result of the Adaboost model.

4. Experimental results of arrhythmia  
4.1 Heart rhythm dataset analysis  
At present, many studies have defined the classification label of arrhythmia classification, which makes it difficult to objectively compare the results of different research methods. In addition, different types of arrhythmias have unusual potential effects on human health, and the treatment methods that need to be used are also different [15]. In this paper, according to the heart rhythm types recommended by AAMI [6], the 15 types of abnormal heart rhythm are classified into five categories, which are specifically classified as showed in Table 2.

In Table 2, each row represents a major category label of cardiac rhythm tags (its proportion in the dataset is marked after each category label), and each category contains one or more heart rhythm subtypes. On the one hand, the number of Q-type ECG beats is very small and can almost be ignored; on the other hand, Q-type arrhythmia rarely appears in the actual diagnosis situation. Therefore, when classifying the arrhythmia, this paper refers to the practice in [16], and Q category is removed from the dataset, and finally, the arrhythmia problems are classified into four categories that is, the dataset only contains N, S, V, F four kinds of rhythm tags.
Table 2. AAMI arrhythmia classification labeling standard.

| Heart rhythm label (proportion) | Content description | Include subcategories | Subclass label description |
|--------------------------------|---------------------|-----------------------|----------------------------|
| N (90%) Normal heart rhythm   |                     | L                     | Left bundle branch block   |
|                                |                     | N                     | Normal ECG beat            |
|                                |                     | R                     | Right bundle branch block  |
|                                |                     | e                     | Atrial escape              |
|                                |                     | j                     | Borderline escape          |
| S (3%) Supraventricular        |                     | A                     | Atrial premature beats     |
|                                | abnormalities       | J                     | Borderline premature beats |
|                                |                     | S                     | Supraventricular premature beats |
|                                |                     | a                     | Abnormal atrial premature beats |
| V (6%) Ventricular escape      |                     | E                     | Ventricular escape beat    |
| F (1%) Fusion Heartbeat       |                     | V                     | Ventricular premature beats |
| Q Unknown ECG beat            |                     | F                     | Ventricular fusion heartbeat |
|                                |                     | P                     | Pacing heartbeat           |
|                                |                     | U                     | Uncategorized ECG beat      |
|                                |                     | f                     | Pacing and normal fusion heartbeat |

In this paper, we define training set DS1 and test set DS2 according to the "between patients" model proposed by Chazal et al. [3] (the details are shown in Table 3) while avoiding the "internal patient" mode. The main difference between the two modes is that the former uses the ECG dataset of two groups of disjoint patients as the training set and test set of the model respectively, which is consistent with the actual medical application scenario. On the other hand, the training set and test set of the latter are randomly selected from the whole dataset, and the data of the same patient will be used to train and test the model, which can easily lead to the overestimation of the classification ability of the model.

Table 3. Confusion matrix of prediction results of Adaboost model.

| Model predicts heart rhythm | N  | S  | V  | F  | Sum  |
|-----------------------------|----|----|----|----|------|
| Real heart rhythm type      |    |    |    |    |      |
| N                           | 44072 | 47 | 51 | 69 | 44239 |
| S                           | 1797  | 35 | 5  | 0  | 1837 |
| V                           | 544   | 5  | 2670 | 1 | 3220 |
| F                           | 317   | 0  | 71 | 0  | 388  |
| Sum                        | 46730 | 87 | 2797 | 70 | 49684 |

4.2 Analysis of experimental results
The specific results of arrhythmia classification based on the Adaboost algorithm are shown in Table 3, and the precision, recall, and accuracy of the model for all kinds of ECG beat classification are shown in Table 4. Table 4 compares the classification effect of this paper with previous studies on four types of heart rate abnormalities.

Table 4. Compare the experimental results.

| Method | Accuracy | N  | S  | V  | F  | R/P/FPR | R/P/FPR | R/P/FPR | R/P/FPR |
|--------|----------|----|----|----|----|--------|---------|---------|---------|
| Chazal(2004) | 86.34 | 87.06/99.17/6.05 | 75.98/38.54/5.09 | 80.31/81.67/1.31 | 89.43/8.57/7.73 |
Zhang(2014) 86.60 89.20/98.07/14.26 56.94/30.59/5.34 78.36/41.82/7.60 0.00/0.00/0.0
Proposed 94.15 99.62/94.31/5.9 1.91/40.22/0.11 82.92/95.46/0.28 0.00/0.00/0.14

In this paper, the classification accuracy of the arrhythmia classification method is 94.15%, which is better than the previous experimental results. The classification effect of a relatively large number of two types of heart rhythm N and V is better than that of the study [3] and [10], while the classification of S and F is slightly worse.

5. Conclusions
Firstly, the MIT-BIH shared dataset, the classification, and combination of arrhythmias recommended by AAMI, and the training test set of inter modes are introduced. Then the denoising method and ECG beat division method used in this paper are introduced. Two median filters are used to remove baseline drift noise and a low-pass filter is used to remove high-frequency noise and EMG interference noise. In this paper, an ECG beat range is defined as 300 milliseconds forward of R wave peak and 500 milliseconds backward, and two types of features in time-domain and frequency-domain are extracted from each ECG beat, including five features: RR, S_K, R_T_wave, wvlt, and WPD. Finally, the Adaboost model is used to select the optimal number and depth of base classifiers (classification trees), which are 100 and 4, respectively, to establish a heart rhythm abnormality classification model.

Acknowledgements
This work was supported by the Preferential Funding for postdoctoral research project under Grant No. B2019003020. The authors are grateful for the valuable comments and suggestions of the reviewers.

References
[1] Huang B, Zhou J. (2018) Classification of common arrhythmias and corresponding treatment principles in communities. J. Chinese General Practice, 21(9): 1009-1019, 1029.
[2] Jiang H, Hong K, Li J. (2014) Value of the ECG tele-monitoring apparatus in the diagnosis of arrhythmias. J. Journal of Clinical Cardiology, 30(5):453-455.
[3] De Chazal P, O'Dwyer M, Reilly R B. (2004) Automatic classification of heartbeats using ECG morphology and heartbeat interval features. J. IEEE transactions on biomedical engineering, 51(7): 1196-1206.
[4] Mar T, Zaunseder S, Martinez J P, et al. (2011) Optimization of ECG classification by means of feature selection. J. IEEE Transactions on Biomedical Engineering, 58(8): 2168-2177.
[5] Li T, Zhou M. (2016) ECG classification using wavelet packet entropy and random forests. J. Entropy, 18(8): 285.
[6] ANSI/AAMI EC57 1998(R) 2003: Testing and reporting performance results of cardiac rhythm and ST-segment measurement algorithm. Association for the Advancement of Medical Instrumentation, Arlington, VA, USA.
[7] Yan H L, Yong A N, Wang H F. (2017) Feature extraction of ECG heartbeats based on convolutional neural networks. J. Computer Engineering and Design, 38(4): 1024-1028.
[8] Lu T C, Lv Y Y, Deng Y L, et al. (2016) Classification algorithm of ECG based on DCT. J. Beijing Biomedical Engineering, 35(3):259-266.
[9] Zhao Y, Hong W X, Sun S B. (2011) Cardiac arrhythmia classification based on multi-features and support vector machines. J. Journal of Biomedical Engineering, 28(2): 292-295.
[10] Zhang R Y, Liao J S, et al. (2013) Research on automatic ECG heart beat classification. J. Journal of Integration Technology, 2(2):46-51.
[11] Andreao R V, Dorizzi B, Cortez P C, et al. (2002) Efficient ECG multi-level wavelet classification through neural network dimensionality reduction. In: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, Martigny, Suisse, pp. 95-404.
[12] de Lannoy G, François D, Delbeke J, et al. (2010) Feature relevance assessment in automatic inter-patient heart beat classification. In: Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing, Valencia, Spain, pp. 13-20.

[13] Osowski S, Linh T H. (2001) ECG beat recognition using fuzzy hybrid neural network. J. IEEE Transactions on Biomedical Engineering, 48(11): 1265-1271.

[14] Park K S, Cho B H, Lee D H, et al. (2008) Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function. In: Computers in Cardiology, IEEE, Bologna, Italy, pp. 229-232.

[15] Martis R J, Acharya U R, Adeli H. (2014) Current methods in electrocardiogram characterization. J. Computers in biology and medicine, 48: 133-149.

[16] Zhang Z, Dong J, Luo X, et al. (2014) Heartbeat classification using disease-specific feature selection. J. Computers in biology and medicine, 46: 79-89.