ECG Identification Based on PCA and Adaboost Algorithm

Qi Liu[1], Yujuan Si[1,2(✉)], Liangliang Li[1], and Di Wang[1]

[1] College of Communication Engineering,
Jilin University, Changchun 130012, China
1508009282@qq.com, siyj@jlu.edu.cn
[2] Zhuhai College of Jilin University, Zhuhai 519041, China

Abstract. Electrocardiogram (ECG) is a weak electrical signal that reflects the process of heart activity, and has multiple excellent features such as uniqueness, stability, versatility, non-repeatability, easy collection and so on. As a new type of biometric authentication technology, the feature extraction and classification of ECG have become a hot research topic. However, there still exists some problems such as poor timeliness and low recognition accuracy. In order to solve these problems, in this paper, we propose an identification method based on Principal Component Analysis (PCA) and Adaboost algorithm. In this method, firstly, we remove the noise from the ECG signal and segment the ECG signal into multiple single heart beats based on detected R points. Then, PCA is used to process heart beat data to reduce feature dimension. Finally, the Adaboost algorithm is used to ensemble weak classifiers to construct a stronger classifier with higher accuracy. In order to validate the effectiveness of the proposed method, we tested our algorithm on 89 healthy subjects of the ECG-ID database. Experimental results show that the proposed method can achieve accuracy rate of 98.88% within 7 s, which demonstrates that the proposed method can provide an effective and practical way for ECG identification.

Keywords: ECG · Identification · PCA · Feature extraction · Adaboost

1 Introduction

Electrocardiogram (ECG) is a kind of weak electrical signal that reflects the beating law of the heart. As a common physiological signal in human body, ECG signals contain measurable characteristic discrepancies among different individuals. Generally, ECG signals are periodic and composed of the similar P-QRS-T waves. However, for different individuals, the position, period and amplitude of each characteristic point are different, which is the basis for ECG signals to be used in personal identification [1]. ECG signals include the following advantages: universality, uniqueness, stability and measurability, which are the necessary conditions of biometrics. In addition, ECG, as a biological feature of human body, is not easy to be stolen, and its safety is relatively higher. Meanwhile, since the ECG signal is one-dimensional, there is low computational complexity in preprocessing and feature processing.

V. G. Duffy (Ed.): HCII 2019, LNCS 11582, pp. 50–62, 2019.
https://doi.org/10.1007/978-3-030-22219-2_4

At present, the research on ECG identification algorithm can be divided into two categories [2]: feature extraction algorithm based on reference point detection and feature extraction algorithm based on non-reference point detection. The feature extraction algorithm based on reference point detection mainly extracts the amplitude, interval, slope, area, angle and other geometric features of ECG signals for identification. Chen et al. [3] extracted five features of Q wave position, S wave position, QRS interval, RQ amplitude difference and RS amplitude difference, this method used Support Vector Machine (SVM) to classify and recognize; Palaniappan et al. [4] intercepted the QRS segments of ECG signals, and extracted five feature points and one morphological coefficient of the segment, this method used the Back-Propagation neural network (BP) to classify and recognize. The feature extraction algorithm based on non-reference point detection is mainly based on transform features, such as time-frequency transform, wavelet transform and sparse coding. Zhao et al. [5] used Fast Fourier's matching tracking method and sparse decomposition of characteristic coefficients to classify and recognize by SVM; Chen et al. [6] used the singular value and dissimilarity distance of the wavelet transform matrix of ECG as the characteristic parameters, SVM was used to classify and 40 samples were identified, the recognition accuracy was 97.82%.

In the above literature, the feature extraction use the reference points overly relies on the positioning accuracy of the reference points, it only focuses on the local information and ignore the overall characteristics of the signal. The feature extraction algorithm based on non-reference point use all the information of the signal, so that this algorithm contains a large amount of redundant information, and the computation complexity is increased by feature transformation. For the common classification models, k-Nearest-Neighbor (KNN) is not regularized for identification, and class deviation is easy to occur in case of sample imbalance. Although SVM performs well, it is sensitive to missing data and the kernel functions should be selected carefully. The learning speed of neural network is slow, and it is easy to fall into local minimum.

Based on the above problems, the feature point location, feature redundancy and classification model selection are analyzed. An ECG identification algorithm based on PCA and Adaboost classifier is proposed in this paper. We extracted complete heart beats through R points, then the PCA was used to process the multidimensional features, removed inter correlation and redundant information, the PCA method can reduce the high-dimensional features to low dimensional features. Finally, Adaboost algorithm was used to construct strong classifier for classification and match. Adaboost algorithm has been successfully applied to face recognition [7], license plate recognition [8], disease diagnosis [9] and other fields because of its simple flow and ideal classification effect. The strategy of "reassigning weights" is adopted to combine weak classifiers weighted into strong classifiers with higher accuracy, which can improve the accuracy of ECG identification. In this paper, the simulation experiment based on ECG-ID database shows that the recognition accuracy and timeliness of the proposed algorithm are improved, which proves that the algorithm has a better performance.

2 ECG Signal Preprocessing and R Point Location

2.1 Denoising

ECG signal is a kind of weak electrical signal. It is easy to be disturbed by noise when collecting, which will affect the accuracy of identification. So the ECG signal needs to be denoised. The noises in the ECG signal mainly include baseline drift (<1 Hz), power frequency interference (50 Hz or 60 Hz), electromyographical interference (30–300 Hz). According to the frequency distribution of noise and ECG signal, we adopt the adaptive wavelet soft threshold method [10] for denoising. The sampling frequency of the signal in the ECG-ID database is 500 Hz. According to the Nyquist sampling theorem, we can know that the frequency of ECG signal is 0–250 Hz, the db4 wavelet is used to decompose the signal with 9 layers. The frequency distribution is shown in Table 1. The frequency of ECG signal mainly distributes in 0.05–100 Hz, of which QRS frequency is 3–40 Hz and S-T frequency is 0.7–10 Hz. So we use 9-layer db4 wavelet to decompose and reconstruct. According to the frequency distribution in the Table 1, the wavelet coefficients of the high frequency component in first layer can be directly set to zero, and the wavelet coefficients of the low frequency component in ninth layer can be set to zero. Then the soft threshold method is used to remove the noise which frequency is mixed with ECG signal.

Table 1. The 9 layer decomposition of ECG signal

Decomposition level	Low-frequency component (Hz)	High-frequency component (Hz)
1	0–125	125–250
2	0–62.5	62.5–125
3	0–31.2	31.2–62.5
4	0–15.6	15.6–31.2
5	0–7.8	7.8–15.6
6	0–3.9	39–7.8
7	0–2.0	2.0–3.9
8	0–1.0	1–2.0
9	0–0.5	0.5–1

First, the threshold of the high-frequency coefficients of each layer is determined, the formula is defined as follows:

$$thr = \alpha_k \times \sqrt{2\log(n)} \times \sigma \tag{1}$$

Where n is the signal length of the threshold processing. a_k is the weighted threshold coefficient of the decomposition of each layer, and σ is defined as follows:

$$\sigma = \frac{median(|d(k)|)}{0.6745} \tag{2}$$

Where $d(k)$ is the wavelet coefficients of each scale. k is the number of layers being processed.

In order to adapt to the wavelet decomposition thresholds of different layers, the weighted threshold method is adopted and the weight is designed, the formula is defined as follows:

$$\alpha_k = \begin{cases} 0.25 & f \leq 30\,\text{Hz} \\ 0.5 & 30\,\text{Hz} < f < 125\,\text{Hz} \\ 0 & f \geq 125\,\text{Hz} \end{cases} \tag{3}$$

Finally, the processed coefficients of the filtered coefficients are reconstructed, and the denoising results of individuals 3 and 9 are shown in Fig. 1. From the graph, we can see that the noise is basically removed, which can meet the needs of identification.

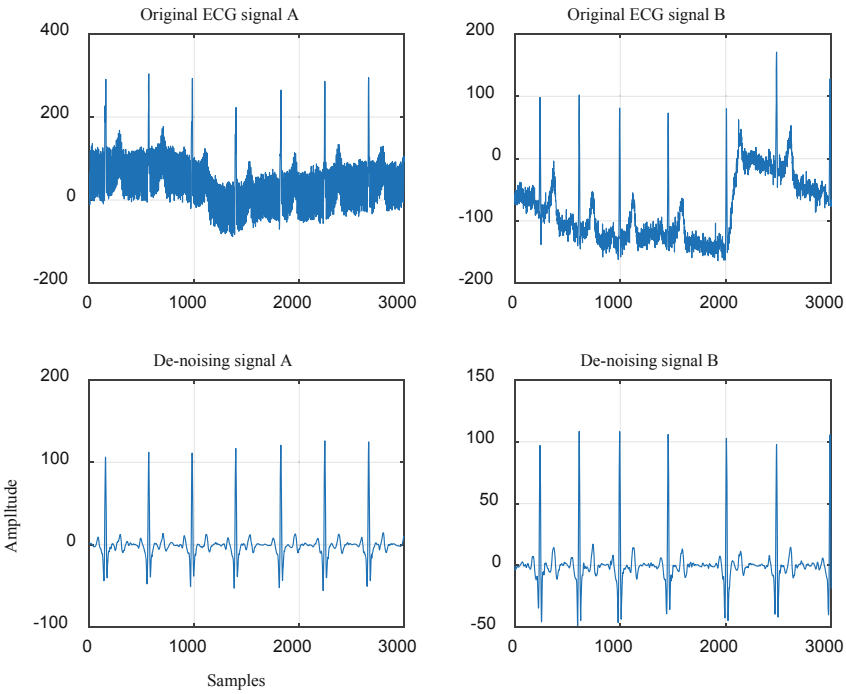

Fig. 1. The diagram of de-noising effect

2.2 R Point Location

The peak value of R wave is the largest in ECG signal, and its location is the simplest. In this paper, the second order difference threshold method [11] is used to locate the peak value of R wave. Figure 2 is the R point detection chart.

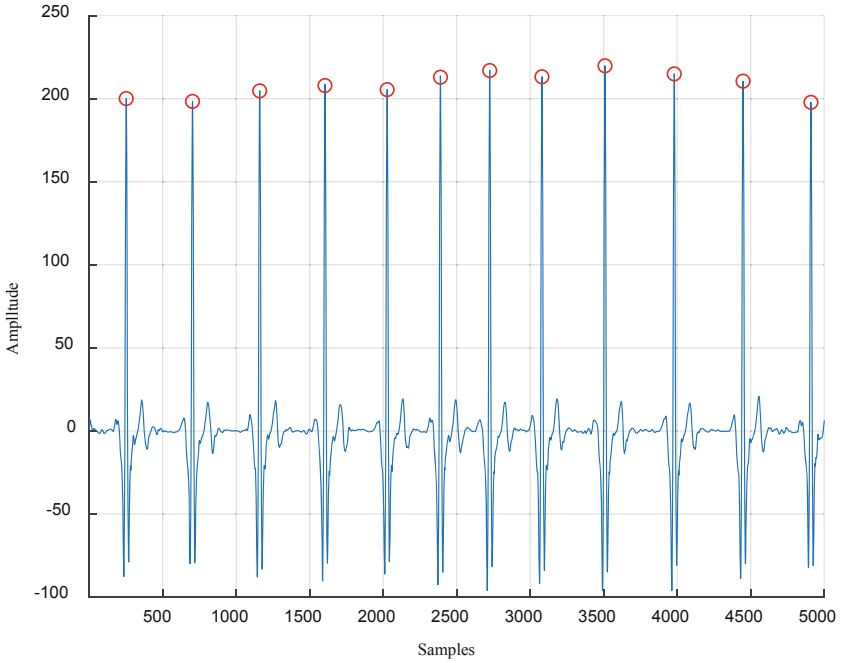

Fig. 2. The R point detection diagram

3 Our Algorithm

3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) [12] is a multivariate statistical algorithm for optimal orthogonal transformation in pattern recognition. It is mainly based on orthogonal projection to remove the correlation between data and maximize the variance of projection data. Using a few main features to replace the original ECG data can remove the correlation of ECG waveform characteristics. This algorithm reduces the data dimension and highlights the main characteristics of the data while retaining the main information of the ECG signal.

The process of principal component analysis is as follows:

(1) Suppose the sample set $X = (x^1, x^2, \ldots, x^m)$ is composed of m heart beat samples, the dimension of each sample is n. Take the sample set to remove the mean:

$$x_{ij}^{*} = x_{ij} - \bar{x}_i \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m) \tag{4}$$

(2) Calculate the covariance matrix:

$$\Sigma = \frac{1}{m} X X^T \tag{5}$$

(3) Calculate the eigenvalue $(\lambda^1 \geq \lambda^2 \geq, \ldots, \geq \lambda^n)$ of the Σ and its corresponding eigenvector $(\omega^1 \geq \omega^2 \geq, \ldots, \geq \omega^n)$.

(4) Determine the required low dimension d' according to contribution rate

$$\varphi = \frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{6}$$

Where d' is determined according to the demand of principal component contribution rate φ.

(5) Transform n-dimensional sample set X into d' dimension space:

$$Z = W^T X \tag{7}$$

$$W = (\omega_1, \omega_2, \ldots, \omega_{d'}) \tag{8}$$

3.2 Adaboost Algorithm

Adaboost algorithm was proposed and developed by Freund [13] in 1996. The Adaboost algorithm adopts the strategy of "reassigning weights" to train the weak classifiers for several rounds, and automatically improve the weights of the wrong samples in the previous training. The weights of weak classifiers with low misclassification rate are added, and the weight of every weak classifier is combined to the weight of the final strong classifier. Adaboost algorithm is often used in the study of binary classification problems, and ECG identification is a multi-classification problem, so the traditional Adaboost algorithm needs to be improved. Zhu et al. [14] proposed an improved algorithm called Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME). It extends the Adaboost algorithm directly to multiple types of problems. The SAMME algorithm reduces the requirement for the correct class rate of the weak classifier from 1/2 to 1/k, which means that in the multi-classification problem, the performance of the weak classifier can be accepted as long as it is better than random guess.

The steps of this algorithm are as follows:

The sample set is given: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ which x_i represent the ECG signal eigenvector, y_i represent the category label of x_i, and $y_i \in Y = (1, 2, \ldots, K)$.

Step 1: Initialize the weight of training samples:

$$w_1 = \frac{1}{m}(i = 1, 2, \ldots, m) \tag{9}$$

Step 2: For $t = 1, 2, \ldots, T$ (T is the number of iterations):

(1) Train the weak classifier h_t according to the sample distribution ω_t.

(2) Calculate the Prediction Error of Weak Classifier h_t:

$$e_t = \sum_{i=1}^{N} \omega_1(i) \bullet 1[y_i \neq h_t(x_i)] \tag{10}$$

Where $1[*]$ means that when $[*]$ is established, it equals 1, otherwise it equals 0.

(3) Calculate the weight of weak classifier α_t based on prediction error e_t:

$$a_t = \frac{1}{2}\ln(\frac{1 - e_t}{e_t}) + \log(K - 1) \tag{11}$$

Where K is the number of categories.

(4) Reassign the next training sample according to the weight a_t:

$$\omega_{t+1}(i) = \frac{\omega_t(i)}{B_T} * \begin{cases} e^{-a_t}, & if \quad y_i = h_t(x_i) \\ e^{a_t}, & if \quad y_i = h_t(x_i) \end{cases} \tag{12}$$

Where B_T is a normalization factor that normalizes ω_t^{t+1}.

Step 3: Built final strong classifier:

$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} a_t \bullet 1[h_t(x_i) = y_i] \tag{13}$$

3.3 The Flow of PCA_Adaboost Algorithm

Based on the simple and efficient Adaboost algorithm, a strong classifier is formed by combining PCA and Adaboost. It improves the timeliness and accuracy of ECG recognition, and has a good feasibility and practical significance. The flow chart is shown in Fig. 3.

4 Experiments and Results

4.1 Experimental Environment and Database

The experimental environment in this paper is the personal PC of the Windows10 operating system, the processor is Intel (R) Core (TM) i5-7500, the memory is 4 GB, and the compilation environments are Matlab2017b and Python3.6.

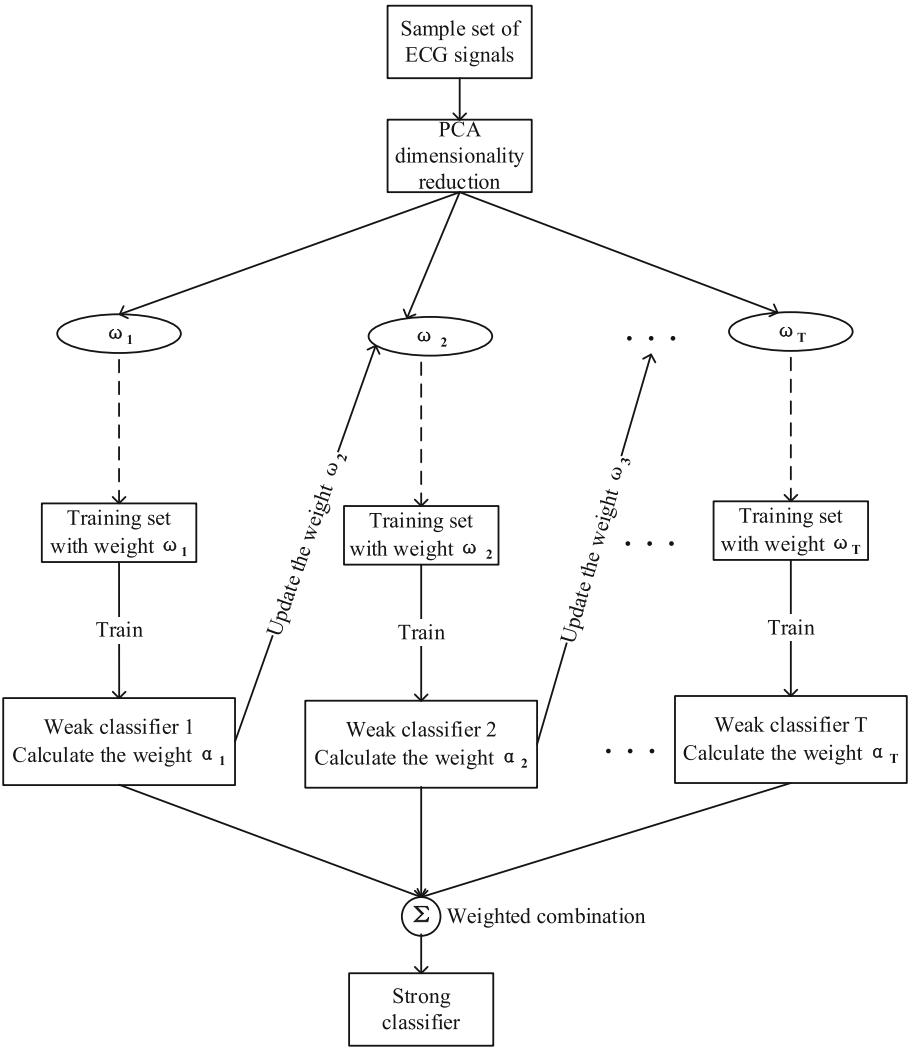

Fig. 3. The flow chart of PCA_Adaboost algorithm

This data is derived from the ECG-ID database [15] in Physionet, which contains ECG signals from 90 people. The signal acquisition frequency is 500 Hz and the resolution is 12bit, of which the seventy-fourth individual only collect one signal, and the remaining 89 people have at least two ECG signals collected at different times. Since only one signal was collected, the seventy-fourth individuals does not satisfy the experimental conditions of identification in this paper, we will eliminate the number seventy-fourth individual. This paper will use the remaining ECG signals from 89 people to carry out the identification experiment.

4.2 Experiment and Result Analysis

In this paper, two ECG signals of each people in the database are taken as training data and test data respectively. First, the pretreatment of denoising is used. Then, 150 points are intercepted forward based on R-point location, and 300 points are intercepted backward to extract a total of 450-dimensionals waveform features, including 2116 training dataset beats and 2110 test dataset beats. PCA is used to reduce the dimension of waveform features.

As shown in Fig. 4, the cumulative contribution rate of the first 10 principal components is 93.56%. In view of the special security requirements of identity recognition, combined with the actual effect of Adaboost classifier classification, we finally selected the 30 dimensions of principal components, the cumulative contribution rate is 99.72%, and the loss of information can be ignored.

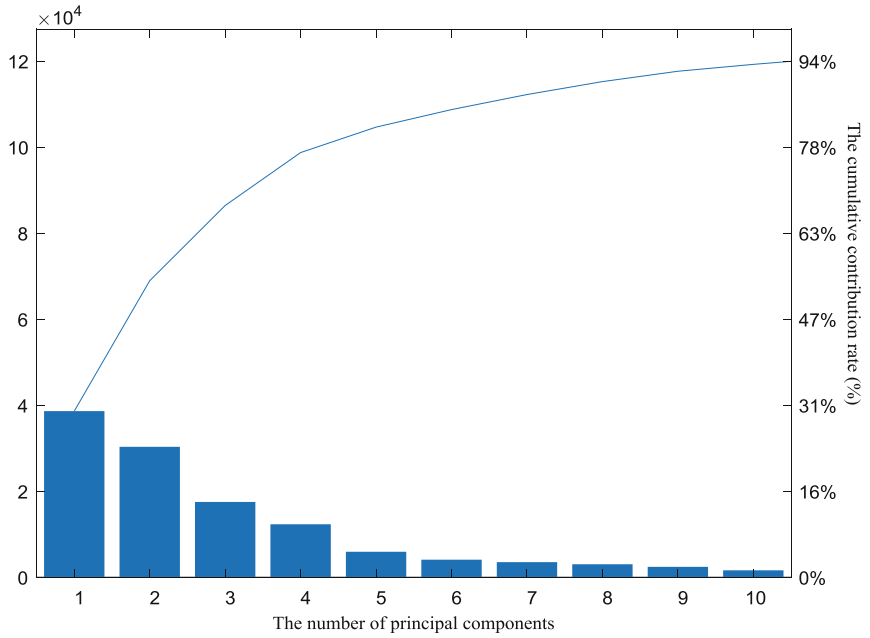

Fig. 4. The principal component contribution rate

Figure 5 is a comparison map of waveform characteristics and PCA dimension reduction characteristics. From the graph, we can see that the heart beat waveform of individual A is relatively close at different times, but it is quite different from individual B and individual C. For PCA dimensionality reduction features, it can be seen that the characteristics of the same individual are closer, and different individuals have more obvious differences. PCA dimensionality reduction highlights the main features of ECG data, reduces the dimensionality of ECG data, and reduces the correlation

between ECG signals and the redundant information that interferes with the recognition accuracy. This algorithm improves the recognition efficiency and recognition accuracy.

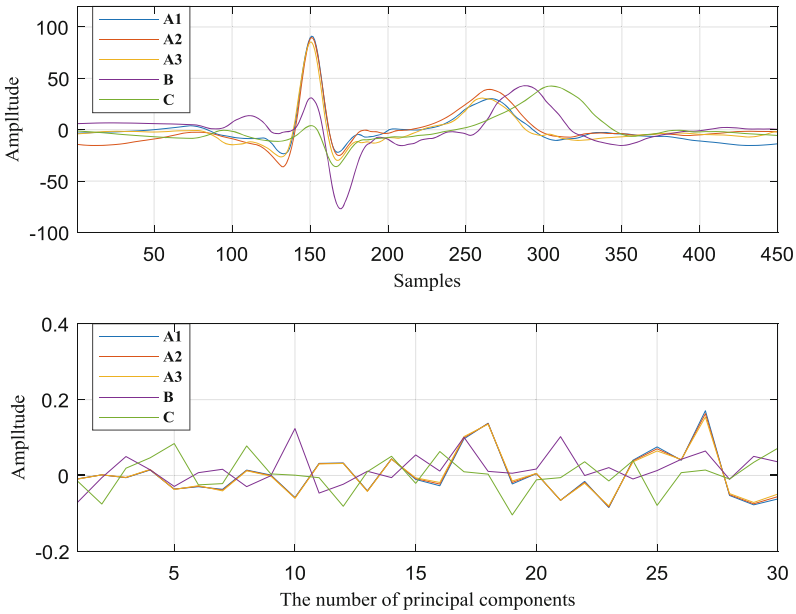

Fig. 5. The waveform features and PCA dimensionality reduction features

Under the same experimental standard, the decision tree is used as the base classifier. And we compared the recognition accuracy of Adaboost algorithm with PCA feature dimension reduction and without PCA feature dimension reduction at different iterations, as shown in Fig. 6.

It can be seen from the Fig. 6 that the recognition accuracy is positively correlated with the different number of iterations. The recognition accuracy is basically stable when the number of iterations is 40 without PCA feature reduction. The recognition accuracy is basically stable when the number of iterations is 30 with PCA feature reduction, and the recognition time is increased with the number of iterations. In this paper, 40 iterations were used to construct the recognition model.

In the same data set the algorithm proposed in this paper was compared with KNN, SVM, BP. And the accuracy rate and average time consuming is shown in Table 2. Among them, KNN classifier parameter K = 5; SVM classifier uses linear kernel function; BP neural network adopts 30-20-89 network [16], the transfer function is "tansig", the minimum gradient is 1.0−e7, and the upper limit of iteration is 500.

In this section, we defined the PCA and KNN recognition method is PCA_KNN, PCA and SVM recognition method is PCA_SVM, PCA and BP recognition method is PCA_BP. This algorithms were compared with KNN, SVM, BP, and Adaboost algorithms. And as shown in the Table 2, the methods combined with the PCA feature dimensionality reduction can improve the recognition accuracy and recognition

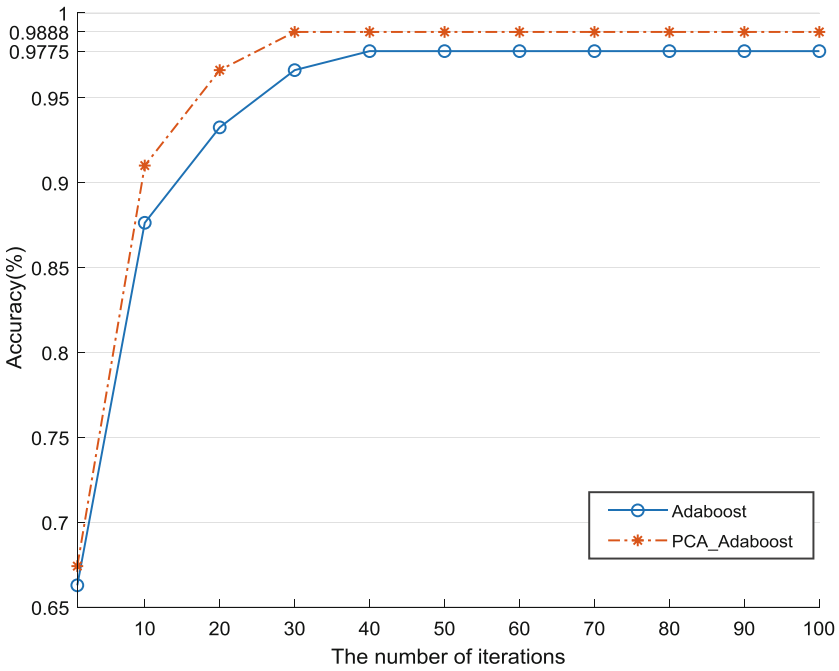

Fig. 6. The relationship between the accuracy and the number of iterations

Table 2. Comparison of different classifier recognition results

Algorithm	Accuracy rate (%)	Average time consuming (s)
KNN	89.89	24.00
BP	95.51	240.00
SVM	95.51	320.00
Adaboost	97.75	110.00
PCA_KNN	92.13	5.00
PCA_BP	96.63	12.00
PCA_SVM	96.63	20.00
PCA_Adaboost	98.88	7.00

timeliness. Among them, the accuracy of PCA_KNN algorithm is 92.13%, the accuracy of PCA_BP algorithm is 96.63%, the accuracy of PCA_SVM algorithm is 96.63%, and the accuracy of PCA_Adaboost algorithm can reach 98.88%. In view of the time-consuming of recognition, the PCA feature dimension reduction greatly reduces the time recognition required, while the average time consuming is basically reduced by one order of magnitude, and the timeliness of recognition is improved. Compared with the PCA_KNN algorithm, PCA _SVM algorithm PCA_BP method, the recognition time of the proposed algorithm is less than PCA_BP and PCA_SVM. Although it takes much more time than the PCA_KNN algorithm, the recognition

accuracy is much higher than PCA_KNN algorithm. Experiments show that PCA algorithm extracts the principal components of ECG signals instead of the original ECG waveform features, and removes the correlation between ECG signals and redundant information interfering the recognition accuracy. The PCA algorithm improves the recognition accuracy, reduces the data dimension of ECG signals, reduces the computational complexity of the algorithm, and improves the recognition efficiency. The Adaboost algorithm adopts the strategy of "weight assignment", to trains the training samples of ECG signals through weak classifiers, it also improves the weight of the wrong samples of the previous ECG signals, reduces the weights of the correct samples of the ECG signals, and increases the weights of the classifiers with small error rates, and reduces the weight of the classifier with large error rates. Several rounds of training and automatic adjustment are combined into a strong classifier with high recognition accuracy.

5 Conclusions

In order to improve the timeliness and accuracy of ECG identification, an ECG identity recognition method based on PCA and Adaboost algorithm was proposed in this paper. This paper mainly studies the dimensionality reduction and classification using waveform characteristics. Firstly, we extracted the single beat based on R point positioning, and PCA was used to reduce the dimension of this feature. Since PCA can remove the correlation and redundant information in original waveform features, the computation complexity of ECG identification using PCA features will be reduced. As a result, not only the running time of the algorithm can be decreased significantly, but also the need for timeliness of algorithm can be well satisfied. Then the Adaboost algorithm was applied to PCA features for model training. Specially, the "reassigning weights" strategy was adopted to combine multiple weak classifiers into one strong classifier for better classification effect. In our experiments, the algorithm was evaluated on ECG-ID database and accuracy of 98.88% could be achieved within 7 s. The experimental results show that our method has good performance on both identification accuracy and timeliness. The next step is to verify the reliability of the algorithm for the ECG signal data collected by ourselves.

Acknowledgments. This work was supported by the Science and Technology Development Plan Project of Jilin Province under Grant Nos. 20170414017GH and 20190302035GX; the Natural Science Foundation of Guangdong Province under Grant No. 2016A030313658; the Innovation and Strengthening School Project (provincial key platform and major scientific research project) supported by Guangdong Government under Grant No. 2015KTSCX175; the Premier-Discipline Enhancement Scheme Supported by Zhuhai Government under Grant No. 2015YXXK02-2; the Premier Key-Discipline Enhancement Scheme Supported by Guangdong Government Funds under Grant No. 2016GDYSZDXK036.

References

1. Biel, L., Pettersson, O., Philipson, L., et al.: ECG analysis: a new approach in human identification. IEEE Trans. Instrum. Meas. **50**(3), 808–812 (2002)
2. Babak, M.A., Sharafat, A.R., Setarehdan, S.K.: An adaptive backpropagation neural network for arrhythmia classification using R-R interval signal. Neural Netw. World **22**(6), 535–548 (2012)
3. Chen, X., Chen, G., Shen, H.: ECG sensor signal identification method based on SVM. Transducer Microsyst. Technol. **33**(10), 40–42 (2014)
4. Palaniappan, R., Krishnan, S.M.: Identifying individuals using ECG beats. In: International Conference on Signal Processing & Communications. IEEE (2005)
5. Zhao, Z., Yang, L., Chen, D.: Research of ECG identification based on FFT-matching pursuit algorithm. Chin. J. Sens. Actuators **26**(3), 307–314 (2013)
6. Chen, D., Zhao, Z.: Identification method of ECG signal based on SVD and dissimilarity analysis. Comput. Simul. **33**(2), 427–432 (2016)
7. Jammoussi, A.Y., Ghribi, S.F., Masmoudi, D.S.: Adaboost face detector based on joint integral histogram and genetic algorithms for feature extraction process. SpringerPlus **3**(1), 355 (2014)
8. Song, M.K., Sarker, M.M.K.: Modeling and implementing two-stage AdaBoost for real-time vehicle license plate detection. J. Appl. Math. **2014**, 1–8 (2014)
9. Ahmed, W., Khalid, S.: ECG signal processing for recognition of cardiovascular diseases: a survey. In: Sixth International Conference on Innovative Computing Technology. IEEE (2017)
10. Bin, H., Bai, Y., Zhang, Y.: Wavelet soft threshold ECG denoising based on different empirical mode decomposition. Math. Pract. Theory **46**(6), 136–144 (2016)
11. Yang, Z.: Real-time detection of ECG waveform based on differential algorithm. J. Mudanjiang Normal Univ. (Nat. Sci. Ed.) (4), 23–25 (2015)
12. Martinez, A.M., Kak, A.C.: PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2), 228–233 (2002)
13. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
14. Zhu, J., Zou, H., Rosset, S., et al.: Multi-class AdaBoost. Stat. Interface **2**(3), 349–360 (2009)
15. Lugovaya, T.S.: Biometric human identification based on electrocardiogram. [Master's thesis] Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation (2005)
16. Yu, J., Si, Y., Liu, X., Wen, D., Luo, T., Lang, L.: ECG identification based on PCA-RPROP. In: Duffy, V.G. (ed.) DHM 2017. LNCS, vol. 10287, pp. 419–432. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58466-9_37