Atrial Fibrillation Detection Based on EEMD and XGBoost

Zhang Yue\textsuperscript{1,*} and Zhu Jinjing\textsuperscript{2}

\textsuperscript{1,2} Division of Information Science and Technology, Tsinghua University, Shenzhen, China

* zhangyue@mail.tsinghua.edu.cn

Abstract. The electrocardiogram (ECG) is non-invasive, inexpensive and widely used in several applications, implemented to detect the physical condition and disease of the human body. Atrial fibrillation (AF) is the most common of many different forms of sustained arrhythmia. Therefore, early diagnosis of AF may help to improve doctor’s diagnostic efficiency and is essential to prevent further progression of Atrial fibrillation to other heart disease and stroke complications. With the popularity of the machine learning and deep learning, more and more researchers apply them in image recognition, speech recognition and so on. Naturally, there are also many studies which achieve the purpose of diagnosing diseases, such as detection of arrhythmia, biometric identification based on ECG signals and machine learning or deep learning. A novel approach to detect AF from ECG signals was developed on this study, we used great filter EEMD (Ensemble Empirical Mode Decomposition) and classifier XGBoost (eXtreme Gradient Boosting) to detect normal rhythm, AF and other rhythm. Finally, the great performance was achieved with an average F1 score of 0.84 and accuracy of 0.86.

1. Introduction

The electrocardiogram (ECG) has significant advantages such as non-invasive, simple, reliable and widely used in several applications \cite{1} and is recorded by multiple electrodes on the surface of the human body. For normal ECG signals, although the ECG waveforms collected by different leads are not the same, they all consist of P-wave, a QRS complex (include Q, R, and S waves) and a T wave (sometimes a U-wave with a small amplitude after the T-wave), as depicted in Figure 1. The ECG signal visually reflects the generation and conduction process of cardiac electrical excitation in the conduction system. Therefore, it can reflect the physiological condition of various parts of the heart objectively to a certain extent, which is one of the important bases for diagnosing heart disease and evaluating cardiac function. Achieving the diagnosis of heart disease, especially for the diagnosis and analysis of various arrhythmia is of great value.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{normal_ecg_trace.png}
\caption{A normal ECG trace}
\end{figure}

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
Cardiovascular diseases (CVDs) are the most main cause of death around world [2]. Atrial fibrillation (AF) is the most common of many different forms of sustained arrhythmia. Therefore, early diagnosis of AF may help to improve doctor’s diagnostic efficiency and is essential to prevent further progression of Atrial fibrillation to other heart disease and stroke complications. Early AF detection has some problems, AF is a supraventricular tachyarrhythmia characterized by rapid, disordered atrial electrical activity and thus it is difficult to detect AF by Sporadic monitoring of the cardiac activity in the hospital.

Many researchers propose multiple approaches for automated classification of ECG signals, such as AF detection, arrhythmia classification and biometric identification. Numerous methods of detecting AF are classified into three ways.

Firstly, according to doctor’s experience, diagnosing AF is based on two main ECG features: the irregularity of the RR intervals (RRI) and the absence of P-waves [3]. Censi et al [4] demonstrates the relationship between different P-wave morphologies and different patterns of interatrial conduction in patients with AF and Bettoni et al [5] demonstrates significant changes in HRV parameters before the onset of PAF (paroxysmal atrial fibrillation).

Secondly, a number of studies implement algorithms to distinguish normal and abnormal ECG signals by extracting time and frequency domain features, then using machine learning classifier such as decision tree, SVM, random forest. An approach is suggested to detect AF based on SVM classifier and wavelet decomposition [6]. Bin et al [7] divides the RR interval features into 30 small Features and finally classifies them with decision tree. Extracting 491 hand-crafted features and ranking these features, then researches select 150 features to detect AF signals and other rhythms by random tree [8].

Finally, more and more studies aim at deep learning methods which are popular due to automatically learning features of large dataset without extracting features by hand in recent years. Much research has been devoted to classify ECG signals by using deep learning methods. A method of identifying 12 heart arrhythmias, sinus rhythm and noise recordings is proposed based on a 34-layer convolutional neural network (CNN) [9]. Tomas et al [10] evaluates ECG signals as sequence and implements a recurrent neural network (RNN) to cluster the individual features. Due to success in the field of computer vision, Tae et al [11] tries to transform 1-D ECG rhythms into 2-D ECG images and enlarge data by augmenting the ECG images. Fernando et al [12] compares an ensemble of bagged trees based 169 hand-craft features with convolutional neural networks and explains the advantages and disadvantages of these two methods.

However, the algorithms of identifying ECG signals existing in the literatures now have three important points: pre-processing, classification and unbalance of dataset. In this study, we proposed a heart rhythm classifier based on XGBoost and EEMD that leverages features based on time, frequency and morphology characteristics of ECG signals. We classified ECG signals into three classes (normal sinus rhythm, AF, other rhythm) using the data provided by the AliveCor device, available from PhysioNet [13]. In this dataset, we used 8249 single lead ECG signals classified into three classes: normal rhythm, AF rhythm, other rhythm, lasting from 9s to 60s, recorded at a rate of 300Hz. The number of these ECG signals are listed in Table 1. Each signal is depicted in Figure 2.

The rest of this paper is shown in the following: The proposed method was described in detail in Section 2. Then in Section 3, we obtain the results. Conclusion is presented in Section 4.

| rhythm          | Normal sinus | AF   | Other rhythm |
|-----------------|--------------|------|--------------|
| numbers         | 5076         | 758  | 2415         |
2. Methodology
The novel method contains four main steps: pre-processing; R-peaks detection; feature extraction and AF classification. Figure 3 depicts four steps of the proposed method. We will describe these steps in detail as follows.

2.1. Pre-processing
As we know, the recorded ECG signals always contain three main types of noises. They are muscle noise, baseline wander, and power-line interferences, which will cause interference with identifying the ECG-types. Therefore, we chose EEMD as ECG pre-processing method to reduce ECG noise. According to Chang [14], compared with EMD (Empirical Mode Decomposition), EEMD has better filtering performance because of reducing mode mixing and also is more useful to remove composite
noise than traditional filters. According to Wu [15], the EEMD algorithm consists in four steps to filter ECG signals and the steps are as follow:

1) mix a white noise series in the ECG rhythms
2) decompose the mixed ECG rhythms into IMFs by EMD algorithm
3) repeat Steps (1) and Steps (2) in the loop, but with different white noise series of the same power at each time.
4) receive the means (ensemble) of the corresponding IMFs as output we want.

The original ECG signal, the typical EEMD decomposition and extracted IMF are depicted in Figure 4. Low frequency components compose the high-level IMF and high frequency components compose the low-level IMF. We chose 3IMF, 4IMF, 5IMF, 6IMF to be reconstructed, while we considered these 4 IMFs possess the important information. The reconstruction is the filtered ECG signal as expected.
2.2. R-peaks detection
In order to detect AF, valuable information is necessary to be extracted from the beat-to-beat intervals (RR). Pan-Tompkins algorithm [16] is a well-known algorithm to detect R-peaks and QRS complex of the ECG signals and extract RR intervals. This algorithm uses slope, amplitude, and width information of ECG signals to reliably detect QRS complexes and can automatically adjusts thresholds and parameters periodically to diverse signals characteristics, heart rate, and QRS morphologies of ECG signals. Hence, we used the adaptive algorithm providing for accurate use on ECG signals to obtain R-peaks and RR intervals and implement them for extracting features.

2.3. Feature extraction
There is large amount of methods about feature extraction for analysing and detecting ECG signals. In this work, we divided the features of the filtered ECG signals into three groups:
- Time domain features
  These features are mainly extracted by utilizing the temporal characteristics of the ECG signals. When the P-wave, QRS complex and T-wave are confirmed, we can obtain some characteristic values of the ECG data, such as average, maximum, range, variance, deviation, kurtosis and so on. In this work, some features are: P wave, QRS width, QT interval, Amplitude of QRS, RR interval, and so on.
- Frequency domain features
  The ECG signals can be transformed from time domain to frequency domain. FFT (Fast Fourier Transform) and DWT (Discrete Wavelet Transform) are two main methods to transform ECG signals. Then we can obtain Shannon entropy, SNR (Signal Noise Ratio), power, frequency band power.
- Non-linear features
  In addition to the features of time domain and frequency domain, some non-linear features can be extracted. Some matrix can be calculated by Poincare plots. We can also compute the heart rate variability, the coefficient of variation and density histograms, ECG signal-quality indices on frequency domain.

Figure 4. illustration of original ECG and IMF distribution.
2.4. AF classification

In this study, we used well-known classifier XGBoost to solve the imbalance of data and improve the accuracy of classification. XGBoost is initially started as a research by Tianqi Chen \[17\] and is a popular, powerful implementation of gradient boosting algorithm after its used in the winning solution of Higgs Machine Learning Challenge. XGBoost considers the case where the training data is sparse, and can specify the default direction of the branch for the missing value or the specified value. Due to the effective of the classification and the different length of ECG signals, we implanted XGBoost to evaluate the performance of our method. The difficulty in using XGBoost to detect AF is to improve the model, parameter tuning is much important. Some methods used are: choose a high learning rate, Tune tree-specific parameters and alpha parameters and then lower the learning rate. Finally, we decided the optimal parameters such as: max_depth=9, min_child_weight=8, colsample_bytree=0.8.

3. Results

In order to avoid the overfitting problem and obtain more accuracy evaluation, our proposed method used 5-fold cross validation. We chose 80% of randomly selected dataset in the entire dataset as training set and the remaining 20% were used in the test set. Evaluation of our result is done by F1 score combining both positive predictivity and sensitivity, F1 score is calculated as:

\[
F_1 = \frac{F_N + F_{AF} + F_O}{3}
\]  

(1)

Where \(F_N\), \(F_{AF}\), \(F_O\) is respectively F1 score of the three rhythms. And the F1 score of each rhythm can be calculated as:

\[
F_{\text{rhythm}} = \frac{2TP_{\text{rhythm}}}{TP_{\text{rhythm}} + FN_{\text{rhythm}} + FP_{\text{rhythm}}}
\]  

(2)

Our methodology achieved appreciable performance measures of 0.84 average F1 score and 0.86 average accuracy. We described the confusion matrix of the result in Figure 5.

![Confusion Matrix](image)

**Figure 5.** Confusion matrix of the result.
4. Conclusion
In this paper, we have presented EEMD and an ensemble classifier XGBoost to distinguish between three types of heart rhythms. The algorithm has shown significant performance, but there are some problems:
(1) the length of ECG signals is different and is distributed between 9s and 60s.
(2) the imbalance of the data.
(3) hand-crafted features are not sufficient to represent the characteristics of the ECG signals.
(4) the features should be selected according to the importance to classifying ECG signals.

Compared with machine learning, deep learning does not require hand-crafted features, has the ability to fully apply the characteristics of the ECG signals, and could use a larger amount of data to improve the diagnosis of atrial fibrillation. Future work will focus on using deep learning to detect AF, decreasing information loss, overcoming the class balance problem and improving the accuracy of detecting AF, normal rhythm and other rhythm. In addition, based on discussing with a cardiologist, future work will also apply the algorithm to the clinic.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No.61571628).

References
[1] P. Melillo, R. Castaldo, G. Sannino, A. Orrico, G. De Pietro, L. Pecchia, Wearable technology and ECG processing for fall risk assessment, prevention and detection, in: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2015, pp. 7740–7743.
[2] World health organization, fact sheet on cvds (fact sheet n0317). March 2015
[3] Dash S, Chon KH, Lu S, Raeder EA, Automatic Real Time Detection of Atrial fibrillation. Annals of Biomedical Engineering. NY, USA: Biomedical Engineering Society 2009; 37: 1701-1709
[4] Censi F, Corazza I, Reggiani E, Calcagnini G, Mattei E, Trivent1M, BorianiG. P-waveVariability and Atrial Fibrillation. Sci Rep 2016; 6:26799
[5] Bettoni M, Zimmermann M. Autonomic Tone Variations Before the Onset of Paroxysmal Atrial Fibrillation. Circulation 2002; 105:2753–2759
[6] S. Asgari, A. Mehrnia, and M. Moussavi, "Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine," Computers in Biology and Medicine, vol. 60, pp. 132-142, 2015.
[7] Bin, Guangyu Shao, Minggang Bin, Guanghong Huang, Jiao Zheng, Dingchang Wu, Shuicai, Detection of Atrial Fibrillation Using Tree Ensemble, Computing in Cardiology Conference. 2017
[8] Mortez Zabihi, Ali Bahrami Rad, Argelos K, Katsaggelos, Serkan Kiranyaz, Susanna Narkilahti, and Moncef Gabbouj. Detection of Atrial Fibrillation in ECG Hand-held Devices Using a Random Forest Classifier, Computing in Cardiology Conference. 2017
[9] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. arXiv170701836 2017
[10] Tomás Teijeiro*, Constantino A. Garcia, Daniel Castro and Paulo Félix, Arrhythmia Classification from the Abductive Interpretation of Short Single-Lead ECG Records, Computing in Cardiology Conference.2017
[11] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, Young-Hak Kim, ECG arrhythmia classification using a 2-D Convolutional neural network, arXiv:1804.06812v1 2017
[12] Fernando Andreotti*, Oliver Carr*, Marco A. F. Pimentel, Adam Mahdi, Maarten De Vos, Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG, Computing in Cardiology Conference. 2017
[13] Gari Clifford, Chengyu Liu, Benjamin Moody, Ikarso Silva, Qiao Li, Alistair Johnson, Roger Mark. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge 2017. Computing in Cardiology

[14] K.M. Chang, Arrhythmia ECG noise reduction by ensemble empirical mode decomposition. Sensors 10, 6063–6080 (2010)

[15] Wu, Z., and N. E Huang (2008), Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis. Vol.1, No.1. 1-41.

[16] J. Pan and W. J. Tompkins, “A Real-Time QRS Detection Algorithm,” IEEE Trans. Biomed. Eng., vol. BME-32, no. 3, pp. 230–236, 1985.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.