Computers in Biology and Medicine 61 (2015) 178–184

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A novel electrocardiogram parameterization algorithm and its application in myocardial infarction detection

Bin Liu [a,1], Jikui Liu [b,1], Guoqing Wang [c,**], Kun Huang [d], Fan Li [c], Yang Zheng [a], Youxi Luo [b,e], Fengfeng Zhou [b,*]

[a] Cardiovascular Disease Center, First Hospital of Jilin University, Changchun 130021, Jilin, PR China
[b] Shenzhen Institutes of Advanced Technology, and Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, PR China
[c] Key Laboratory of Zoonosis, Ministry of Education, Norman Bethune College of Medicine, Jilin University, Changchun 130021, Jilin, PR China
[d] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
[e] School of Science, Hubei University of Technology, Wuhan 430068, PR China

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 11 August 2014

Keywords:
Bioinformatics
Health informatics
Polynomial fitting function
Electrocardiogram
Myocardial infarction
Prediction

ABSTRACT

The electrocardiogram (ECG) is a biophysical electric signal generated by the heart muscle, and is one of the major measurements of how well a heart functions. Automatic ECG analysis algorithms usually extract the geometric or frequency-domain features of the ECG signals and have already significantly facilitated automatic ECG-based cardiac disease diagnosis. We propose a novel ECG feature by fitting a given ECG signal with a 20th order polynomial function, defined as *PolyECG-S*. The *PolyECG-S* feature is almost identical to the fitted ECG curve, measured by the Akaike information criterion (AIC), and achieved a 94.4% accuracy in detecting the Myocardial Infarction (MI) on the test dataset. Currently ST segment elongation is one of the major ways to detect MI (ST-elevation myocardial infarction, STEMI). However, many ECG signals have weak or even undetectable ST segments. Since *PolyECG-S* does not rely on the information of ST waves, it can be used as a complementary MI detection algorithm with the STEMI strategy. Overall, our results suggest that the *PolyECG-S* feature may satisfactorily reconstruct the fitted ECG curve, and is complementary to the existing ECG features for automatic cardiac function analysis.

1. Introduction

Myocardial infarction (MI) is the symptom of heart cell injury due to the lack of oxygen, and is usually caused by the buildup of white blood cells in the cardiac vessels [1]. MI is one of the top two cardiovascular diseases in the United States, and is notorious for its fatality rate and frequent recurrences [2]. MI may be diagnosed from electrocardiogram (ECG) signals, biochemical biomarkers and echocardiography imaging [3]. But its early and precise detection still remains challenging, due to the weak MI-specific association in the aforementioned data types, and the prolonged data generation steps [1,3]. Both data precision and clinician experience play essential roles in MI detection.

MI may be diagnosed by the detection of significant changes in the ST segment or Q wave in the subject's ECG signal [1]. The ECG represents one of the major features for MI determination since the data are easy to collect and require only a short sampling duration. But the diagnostic criteria are difficult to follow due to the inexplicitness and ECG individualized polymorphism. Even experienced cardiologists may only recognize 82% of the ST-segmental elevation in MI subjects [4]. To facilitate ECG-based MI detection for inexperienced clinicians or non-cardiologists, a number of ECG analysis algorithms were proposed for automatic MI detection. But the detection accuracy of these algorithms remains to be improved for practical clinical applications. After the wavelet transformation of the ECG signals, Al-Naima and his colleagues applied a Multi-Layer Perceptron Neural Network model to the detection of MI subjects, and achieved 90% accuracy [5]. Chang et al. combined the power of the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM), and achieved 82.5% accuracy [6].

This study proposes the development of a novel ECG feature by fitting the ECG signal with a polynomial function. The polynomial function fits the ECG curve with high accuracy, and the fitted

* Corresponding author. Tel.: +86 755 86392200; fax: +86 755 86392299.
** Corresponding author.
E-mail addresses: qing@jlu.edu.cn (G. Wang), FengfengZhou@gmail.com, ff.zhou@siat.ac.cn (F. Zhou).
URL: http://www.healthinformaticslab.org/ffzhou/ (F. Zhou).
[1] These authors have contributed equally to this work.

http://dx.doi.org/10.1016/j.compbiomed.2014.08.010
0010-4825/

coefficients are defined to be the ECG representing features. The optimal parameters of this algorithm were determined using clinically collected ECG data from the PTB database. This new ECG feature was applied and tested on the MI detection problem, and achieved 94.4% in overall accuracy.

2. Material and methods

2.1. Data sources

The ECG data were collected from the Physikalisch-Technische Bundesanstalt database (PTBdb) [7]. 148 myocardial infarction (MI) patients and 52 healthy subjects were digitalized to 1000 sampling frequency, and the signals from conventional 12 leads (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) were retrieved for this study.

2.2. Signal pre-processing

All the ECG signals were processed in the following steps. First, a discrete wavelet transformation (DWT) was conducted to remove high-frequency noise and baseline shifting. DWT works well on removing the inter-signal dependence among multiple signals, and can achieve non-redundant signal decomposition [8]. An ECG signal has the frequency range of 0.5–100 Hz [9]. The noise in the ECG signals mainly occurs at frequencies higher than 100 Hz, and the baseline shifting is usually distorted by the signals lower than 0.5 Hz in frequency [10]. In this study, the coif4 mother wavelet was used, because its shape is very similar to the ECG signals, and it is thus is a good representative for the signal [11].

Next, all the R peaks in the ECG signals were detected using the wavelet transformation. QRS waves are the major representative points in ECG signals, and most of the automatic ECG analysis algorithms are based on the correct detection of these waves. A peak usually occurs in the wavelet-transformed R waves, and this change may be captured by levels 3 and 4 of the quadratic spline mother wavelet, which ranges between 15 and 25 Hz [12,13]. Level 4 is excluded since it may be affected by the P waves [14].

All the ECG signals are split into ECG cycles by the detected R peaks. Although an ECG cycle is usually defined as a continuous segment of PQRST waves, no commonly accepted definition is available for the exaction of boundaries for each ECG cycle. The R peak is the most conspicuous wave pattern, and is the easiest wave to detect among the PQRST waves. Since this study requires an objective splitting of ECG signals into cycles, R peaks are chosen as the ECG cycle boundaries.

In order to make different ECG signals comparable to each other, the third step normalized a given ECG cycle into $[0, 1]\times[0, 1]$ on both the time and voltage axis, as described previously [15]. The ECG baseline drifting, respiration-induced QRS complex change, and other fluctuations may affect some normalization features. But this study proposes an objective ECG normalization algorithm, and its classification performance will be evaluated on a real dataset in this study. Such a scaled ECG curve is defined as a Unified ECG Cycle (UEC), and the scaling factors for time and voltage are defined as time (TF) and voltage factors (VF), respectively. The two factors TF and VF reflect the differences in the ranges of durations and voltages of different ECG cycles.

2.3. ECG polynomial fitting algorithm (PolyFit)

This study proposes a polynomial function to fit the ECG signals, and represents each ECG cycle as a vector of the coefficients of this polynomial function. With the rapid increase of the sampling rate, each ECG cycle has ~1000 data points under the current technology (1 kHz sampling rate). So a polynomial function may significantly reduce the data dimensionality, if its order is smaller in magnitude than 1000.

For a given short period of ECG signal $\{(x_i, y_i)\}$ $(i=1, 2,\ldots,m)$, a kth order polynomial fitting function $PolyFit(x)$ is defined as:

$$PolyFit(x) = a_0 + a_1 x + \cdots + a_k x^k \quad (1)$$

Then the least square rule is facilitated to minimize the sum of squared deviations between the fitting curve and the ECG cycle. The fitting optimization function is defined as:

$$\min R^2 = \min \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + \cdots a_k x_i^k)]^2 \quad (2)$$

The vector of coefficients $\langle a_0, a_1,\ldots, a_k\rangle$ is calculated using the function *polyfit* and *polyval* in Matlab.

2.4. PolyFit-based ECG parameterization algorithm (PolyECG)

A Unified ECG Cycle (UEC)$C=\{(x_i, y_i)\}$ $(i=1, 2,\ldots, m)$ is one ECG cycle between two R peaks, and its time and voltage factors are defined as $TF(C)$ and $VF(C)$. As defined in the above sections, $0 \le x_i, y_i \le 1$ $(i=1, 2,\ldots,m)$. If the polynomial fitting function $PolyFit(x)= a_0+a_1x+\cdots+a_kx^k$, the complete UEC curve C is transformed as a parameter vector $PolyECG\text{-}C(C, k)=\langle a_0, a_1,\ldots, a_k, TF(C), VF(C)\rangle$. A refined ECG parameterization algorithm is further defined as the *PolyECG*'s splitting version; $PolyECG\text{-}S(C, k)=PolyECG\text{-}C(C_1, k)\cup PolyECG\text{-}C(C_2, k)$, where $C=C_1\cup C_2$, and $x(C_1)\le 0.5$, $x(C_2)>0.5$. The hypothesis for this equal-sized splitting is that the PQR and ST segments in an ECG cycle have significantly different shape patterns, and may need different sets of fitting parameters. This hypothesis will be validated and supported in the following section. The parameters in the vector are regarded as the features of this UEC, and further screened for the best feature subset. There are $(k+1+2)$ and $(2k+2+2)$ parameters/features for $PolyECG\text{-}C(C, k)$ and $PolyECG\text{-}S(C, k)$, respectively.

2.5. Model evaluation with the Akaike information criterion (AIC)

The Akaike information criterion (AIC) is used to measure how relatively well a polynomial function $PolyECG\text{-}C(C, k)$ or $PolyECG\text{-}S(C, k)$ fits the ECG cycle curve C with different polynomial function order k. AIC evaluates a given fitting model from the perspectives of both the fitting goodness and complexity based on the information entropy [16], and is better with a smaller value [17]. The AIC is defined as $AIC=n\ln(RSS/n)+2(p+1)$, where p is the number of parameters, n is the number of data points in this ECG cycle C, $RSS=\sum_i(y_i-\hat{y}_i)^2$, and y_i and $\hat{y}_i$ are the real and predicted values for the ith data, respectively. Usually a more complex model tends to fit the data better, *i.e.* the first term $n\ln(RSS/n)$ decreases. But a larger p due to the complex model will increase the second term $2(p+1)$, and a more complex model may over-fit the dataset due to Runge's phenomenon [18]. This work calculates the AIC value for each trained model, and determines the best choice of the polynomial fitting function order k with the smallest AIC.

2.6. Feature selection and classification

The ECG signals of 148 myocardial infarction (MI) patients and 52 healthy subjects were processed and divided into training and testing datasets. The first and last 5 s of a given ECG signal were removed, and then the ECG signal was split into multiple Unified ECG Cycles (UECs). Each UEC curve C was transformed as a data entry of a feature vector by either $PolyECG\text{-}C(C, k)$ or $PolyECG\text{-}S(C, k)$, where k is the order of the polynomial fitting function. The data entries from the first halves of the ECG signals constituted the training dataset *TrainSet*, and the other data entries were in the testing dataset *TestSet*.