Ensemble of algorithms for coronary heart disease detection based on electrocardiogram

V.N. Guryanova[1]

[1] Lomonosov Moscow State University, Leninskie Gory, Moscow, Russia, 119991

Abstract. Coronary heart disease (CHD) is the leading cause of death in the world. This disease can be asymptomatic for a long time and over time can progress and result in death. Today electrocardiogram (ECG) can be done at home with the help of special equipment from CardioQvark. In this paper, the possibility of CHD detection based on such ECGs was explored. Different approaches to the classification of such electrocardiograms were surveyed. New algorithms and modifications to existing algorithms were proposed. A new method – ensemble of different algorithms – has shown the best quality.

Keywords: ECG Signal Processing, Ensemble Learning, Automatic CAD Detection, ECG classification.

1. Introduction

Coronary heart disease(CHD) [1] is a group of diseases, that is defined by lack of oxygen supply to the heart muscle through the coronary arteries. This disease is the leading cause of death in the world. It is very important to identify the CHD in time to slow the course of the disease and prevent the patient's death.

Currently, it is extremely relevant to create a device that will help determine the disease or its probability at home. Such devices will allow the person to be sent to a doctor in case of a high probability of having a CHD. CardioQvark [2] has created a device in a form of a smartphone case that allows you to make ECG measurements at home. In this work, the possibility of CHD detection based on such ECGs was explored.

There are different approaches to ECG classification problem. Some of them were surveyed in this work. New algorithms and modifications to existing algorithms were proposed. In order to improve the quality of the classification, an ensemble of 5 different methods was built. Each of these methods will be briefly described below.

2. Data Description

The sample that was used for this task consists of 1798 cardiograms. It contained 1055 cardiograms of healthy patients and 743 cardiograms of patients with CHD. The sampling frequency was 1000 Hz.

Signals were preprocessed before applying machine learning algorithms. For preprocessing, a low-pass and high-pass Butterworth filters [3] of the second order were used. For low-pass filter cutoff frequency was 0.3 Hz. For high-pass filter cutoff frequency was 15 Hz. The signal trend was extracted using a median filter [3]. Then the trend was subtracted from the signal.

3. Algorithms Description

The first algorithm uses images obtained via bispectrum decomposition of a signal as a feature space. The bispectrum is a function of two variables f_1 and f_2 that specify the frequencies, expressed by the following formula [4]:

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2), \tag{1}$$

where $X(f)$ is the Fourier transform of the signal, and $X^*(f)$ is the complex conjugate of it. Bispectrum of the signal can be represented as two-dimensional matrix whose elements are complex numbers. The elements of this matrix can be signed as a_{ij}, so some image can be assigned to each signal. This image is calculated in the following way. A new matrix $B = ||b_{i,j}||$ is computed, the elements of which are equal to:

$$b_{i,j} = \sqrt{\mathrm{Re}^2 a_{i,j} + \mathrm{Im}^2 a_{i,j}}, \tag{2}$$

where Re is the real part of the complex number, and Im is the imaginary part of the complex number. As the image, a contour plot was used.

Signal bispectrum was successfully used earlier to determine CHD, however, the classification took place without the use of machine learning algorithms. It was proposed to classify the obtained images with the help of neural networks, which now successfully solve many problems associated with the image classification [5]. In this paper, a neural network with the architecture described in Table 1 was used. Leaky ReLU [5] and sigmoid [5] functions were used as activation functions.

Table 1. Neural Network Structure

Input Layer	Shape = (3, 80, 80)
Convolutional Layer	Filter Shape = (32, 5, 5) Stride = (2, 2)
Dense Layer	Units Number = 30 Activation Function = leaky ReLU
Dense Layer	Units Number = 1 Activation Function = sigmoid

The basis of the second algorithm is the transformation of the signal into text, consisting of code words. To perform this transformation, the following actions are performed. The approximation coefficients from the signal wavelet transformation are used as a new signal representation. In this paper, Daubechies wavelets [6] were used as wavelet functions. After obtaining the coefficients of the wavelet transform, local segments are extracted from the signal. To do this, it is necessary to use a “window” of some length w, with step s along the signal. All elements that fall within the same window are recorded in a separate segment.

All local segments that enter the training sample are clustered into k clusters by the k-means method. After that, each segment is replaced by the cluster number to which it belongs. Thus, each signal is represented in the form of a text consisting of code words corresponding to the selected clusters. When implementing the given approach, the following parameter values were used: $w = 100$, $s = 30$, $k = 200$. In the test sample, each local segment is replaced by that cluster, closer to which it is located. This approach was used earlier to classify patients via their ECG [7].

In order to get the dependencies between local segments in the signal, it was proposed to use the features based on word2vec [8] technology. The word2vec model was built on a training sample with a vector length of 80. Average of all words vectors in the text corresponding to the signal was used as a feature. Logistic Regression was used as the classification model for this algorithm.

The third algorithm was proposed earlier to determine the necessity of patient's hospitalization based on his ECG [9]. The feature space for this approach is constructed in the following way. First, the neighborhoods of the signal R-peaks are allocated: 200 points before the R-peak and 500 points after. Then, the averaged neighborhood is used as a feature space. Neural network with the architecture described in Table 2 was used as the classification model for this algorithm.

Table 2. Neural Network Structure

Input Layer	Shape = (700)
Dense Layer	Units Number = 90 Activation Function = sigmoid
Dense Layer	Units Number = 1 Activation Function = sigmoid

The fourth algorithm uses a set of features derived from the HRV signal. The HRV signal is a signal that is obtained by calculating the distances (RR) between the R peaks in the original ECG signal and then transforming these distances as 60/RR. Various groups of features are constructed based on the HRV signal [10].

The first group of features includes various entropic features, which indicate a measure of unpredictability in the signal. The following types of entropies are used: approximate entropy, sample entropy, and Shannon entropy.

The next group of features is based on a recurrence plot. This plot shows the frequency and duration of recurrences in the signal. Based on this plot, the following features are calculated: density of points on the plot, percentage of points that form diagonal lines, the average length of diagonals, the entropy of diagonal lines length, the entropy of vertical lines length.

Also features based on the Poincare plot, detrended fluctuation analysis, and correlation dimension are used. Gradient boosting was used as the classification model for this algorithm.

The fifth algorithm consists of three different groups of features. Statistical features of the signal: mean, standard deviation, signal minimum, signal maximum, selective quantiles of order: 0.1, 0.25, 0.5, 0.75, 0.9, sums and sums of signal values squares that are above / below certain values of quantiles: 0.1, 0.25, 0.5, 0.75, 0.9, skew, kurtosis. Hjorth's parameters: activity, mobility, complexity [11]. Uspensky features [12] are used in this algorithm. These features are constructed in the following way. The signal is transformed into code sequence using increments of R-peak amplitudes, increments of distances between R-peaks, and also inverse tangents of their ratios. After signal code representation is received, three-gram selection is performed. The feature space is the number of occurrences of each of the possible three-gramms in a given code sequence obtained from the signal. Gradient boosting was used as the classification model for this algorithm.

The ensemble of algorithms was built with the help of the majority vote and EM algorithm [13]. This EM algorithm was proposed to aggregate the data received from different people about the same event in order to obtain the true result. Since the purpose of constructing an ensemble of algorithms is the aggregation of the answers of various algorithms to obtain the true result, this algorithm can be used to construct an ensemble. As a result of the work of this algorithm, the probabilities of data belonging to each class are obtained. The answer is the class, the probability of belonging to which is the greatest.

4. Evaluation of Algorithms

Cross-validation was used to evaluate the quality of algorithm. To avoid overfitting the ECGs of one patient did not fall simultaneously into the training and test set. The following quality criteria was introduced:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{n_i}\mathbf{I}_{t_{ij}=p_{ij}}}{n_i}, \tag{3}$$

$$\mathbf{I}_{t_{ij}=p_{ij}} = \begin{cases} 1 & \text{if } t_{ij} = p_{ij}, \\ 0 & \text{if } t_{ij} \neq p_{ij}, \end{cases} \tag{4}$$

where t_{ij} is the true value of the target variable for the cardiogram j of the patient i, p_{ij} – the predicted value of the target variable for the cardiogram j of patient i, n_i – the number of cardiograms of the patient i, N – the number of patients. This criteria is called patient quality. It allows us to evaluate how well the algorithm determines a person's disease by any of his cardiograms. In addition, this quality criterion does not depend on the number of cardiograms for each patient. ROC-AUC score [14] and F-score [15] were used for models evaluation.

5. Results

The results of evaluations are shown in Table 3, where first column shows algorithms number or ensemble type. The second algorithm is included in two variants, with word2vec and without.

Table 3. CHD detection results

Algorithm	Patient Quality	ROC-AUC	F-score
1	0.7207	0.7418	0.7244
2	0.741	0.8	0.6967
2+word2vec	0.7501	0.8	0.6990
3	0.7602	0.7988	0.703
4	0.763	0.744	0.662
5	0.7632	0.8042	0.70256
majority	0.806		0.77
EM	0.8108	0.8738	0.7784

6. Conclusion

In the course of this paper, the following results were obtained. When CardioQvark equipment is used, it is possible to determine CHD with an accuracy of more than 0.81 for patient quality, with an accuracy greater than 0.77 for F-score and with an accuracy greater than 0.87 for ROC-AUC score. Word2vec can increase the quality of the classification method based on the wavelet transform. Bispectrum can be used to classify CHD. The EM algorithm is applicable for ensemble and in this case, shows the best quality of classification for all selected quality criteria.

7. References

[1] Gorbachev, V.V. Cardiac ischemia. — Minsk: High school, 2008. — P. 479. (in Russian).

[2] CardioQvark Web Site [Electronic resource]. — Access mode: http://www.cardioqvark.ru (21.11.2017).

[3] Rangayyan, R.M. Biomedical signal analysis. — John Wiley & Sons, 2015. — Vol. 33.

[4] Al-Fahoum A., Al-Fraihat A., Al-Araida A. Detection of cardiac ischaemia using bispectral analysis approach // Journal of medical engineering & technology. — 2014. — Vol. 38(6). — P. 311-316.

[5] Goodfellow I., Bengio Y., Courville A. Deep learning. Book in preparation for MIT Press. — Access mode: http://www. deeplearningbook.org (21.11.2017).

[6] Liu, C.L. A tutorial of the wavelet transform //NTUEE, Taiwan. — 2010.

[7] Mikolov, T. Efficient estimation of word representations in vector space //arXiv preprint:1301.3781. — 2013.

[8] Wang, J. Bag-of-words representation for biomedical time series classification //Biomedical Signal Processing and Control. — 2013. — Vol. 8(6). — P. 634-644.

[9] Ripoll, V.J.R. ECG assessment based on neural networks with pretraining //Applied Soft Computing. —2016. — Vol. 49. — P. 399-406.

[10] Dua, S. Novel classification of coronary artery disease using heart rate variability analysis // Journal of Mechanics in Medicine and Biology. — 2012. — Vol. 12(4). — P. 1240017.

[11] Hjorth B. EEG analysis based on time domain properties //Electroencephalography and clinical neurophysiology. – 1970. – Vol. 29(3). — P. 306-310.

[12] Uspenskiy V.M. Information function of the heart. Theory and practice of diagnosis of diseases of the gastrointestinal tract by methods of methodical analysis of electrocardiograms. - M.: Economics and Informatics, 2008. — P. 116. (in Russian).

[13] Dawid A.P., Skene A.M. Maximum likelihood estimation of observer error-rates using the EM algorithm //Applied statistics. — 1979. — P. 20-28.

[14] Fawcett, T. An introduction to ROC analysis // Pattern recognition letters. — 2006. — Vol. 27(8). — P. 861-874.

[15] Sokolova M., Japkowicz N., Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation // Australian conference on artificial intelligence. — 2006. — Vol. 4304. — P. 1015-1021.