Feature Selection Using Genetic Algorithms for Premature Ventricular Contraction Classification

Yasin Kaya[1], Hüseyin Pehlivan[2]

Department of Computer Engineering, Karadeniz Technical University, Trabzon, Turkey
[1]yasin@ktu.edu.tr, [2]pehlivan@ktu.edu.tr

Abstract

Cardiac arrhythmia is one of the most important indicators of heart disease. Premature ventricular contractions (PVCs) are a common form of cardiac arrhythmia caused by ectopic heartbeats. The detection of PVCs by means of ECG (electrocardiogram) signals is important for the prediction of possible heart failure. This study focuses on the classification of PVC heartbeats from ECG signals and, in particular, on the performance evaluation of selected features using genetic algorithms (GA) to the classification of PVC arrhythmia. The objective of this study is to apply GA as a feature selection method to select the best feature subset from 200 time series features and to integrate these best features to recognize PVC forms. Neural networks, support vector machines and k-nearest neighbour classification algorithms were used. Findings were expressed in terms of accuracy, sensitivity, and specificity for the MIT-BIH Arrhythmia Database. The results showed that the proposed model achieved higher accuracy rates than those of other works on this topic.

1. Introduction

Cardiovascular disease (CVD) is listed as a major underlying cause of death, accounting for 47.73% of all deaths in Turkey [1]. In order to reduce the mortality rate caused by CVD, watching heart cycles for the recognition of early complications is a vital concern for cardiologists and related medical staffs. Electrocardiogram (ECG) is a signal that conveys important information for indicating the abnormal status of cardiovascular system. Detection and classification of different types of heart beats in the ECG is great importance for diagnosis of cardiac dysfunction. It is necessary to record long-term ECG signals to be aware of cardiac problems. Cardiologists must be deep examination in these records. It takes more time to examine each beat of these signals. Developing algorithms to be used ECG analysis will facilitate to examine of long ECG signals.

Among the various abnormalities, premature ventricular contraction (PVC) is one of the most important arrhythmias. PVC results from the early depolarization of the myocardium originating in the ventricular area and is a common form of arrhythmia in adults. It is usually associated with structural heart disease and increases the risk of sudden death. Additionally, its assessment and treatment are complex [2], [3]. This paper focuses on the classification of PVC arrhythmias using selected features with genetic algorithms.

In recent years, numerous studies have been conducted on detection and classification of arrhythmia problems. Researchers attempting to classify PVC arrhythmias have used different feature extraction and classification methods. Inan et al. attempted to classify PVC beat. They combined wavelet-transformed ECG waves with timing information as their feature set for classification. They used neural network (NN) classifier to classify PVC beat and their classification accuracy was 95.16% over 40 ECG signals in the MIT-BIH Arrhythmia Databases [4]. In [5], authors focused morphological transformation and cross-correlation technique for detection PVC beat. They used a modified morphological filtering (MMF) technique for signal pre-processing and Multiscale Morphological Derivative was performed on the MMF conditioned signal to detect PVC beat present in the signal. MIT-BIH Arrhythmia Database was used in the experimental result and they achieved sensitivity and specificity rates of 99.67% and 95.2%, respectively. In [3], RR-interval, QRS-width and QRS-pattern was used in their algorithm. They used simple decision rule to detect PVC beats. The performance of proposed method was 91.05% of sensitivity and 99.55% of specificity. In [6], authors used NN to classify PVC beats and they obtained 99.7% of normal beat correct detection rate and 98.5% of PVC correct detection rate.

In this paper, an effective approach was developed for the classification of PVC arrhythmias. The main objective was to improve classification accuracy of the system and examine the performance of selected features from genetic algorithms. Selected features and time series of the signal was used to evaluate performance metrics for classification. Neural networks (NN), k-nearest neighboor (k-NN) and support vector machines (SVM) classifiers were applied using different schemes to obtain the experimental results. The test data used in the analysis were selected MIT-BIH Arrhythmia Database[7]. The results showed that the proposed approach gathered the considerably high accuracy rates of 99.69% and provided better detection performance than other works studied previously.

2. Material and Methods

We used all ECG signals containing normal and PVC beats from MIT-BIH Arrhythmia Database in this work. Preprocessing step was used for noise reduction. Beat parsing was performed on noise free signal and 200 samples were selected as one ECG beat. After beat parsing step genetic algorithm feature selection process was implemented to reduce the size of feature vector. Both time series and selected features used for classification stage and their results compared. Fig. 1 shows the block diagram of the proposed approach.

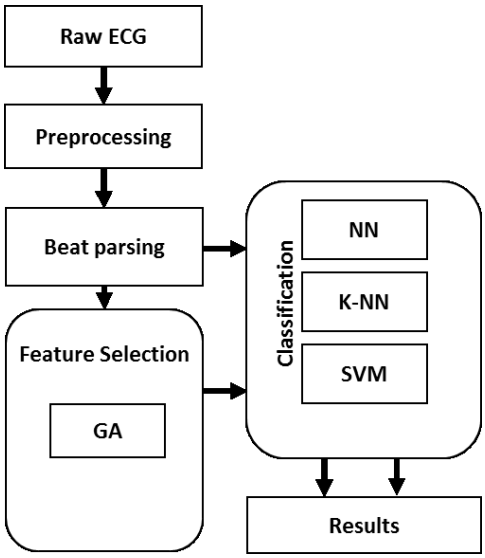

Fig. 1. Block diagram of the proposed approach.

2.1. ECG Database

In this study the MIT-BIH Arrhythmia Database [7], [8] was used as the data source. The database contains 48 signals of 30 min duration each, and two leads – Lead II and one of the modified leads (V1, V2, V4, or V5). The signals of the database were sampled at 360 Hz. Twenty-three files were randomly selected to serve as a representative sample of routine clinical recordings and 25 files were selected to comprise rare complex ventricular, junctional, and supraventricular arrhythmias. The database was annotated both in timing information and beat label. The annotation labels were used to locate the beats in the signal files in this work. We didn't used QRS detection algorithm. Approximately 100 normal beats were selected for the test from each file. The data used consisted of 3500 (from 35 files) normal (N) beats and 3500 (from 33 files) PVC beats. The PVC beats were intermittently selected from the files because these beats were unevenly distributed in the files. All the details of the dataset we used can be accessed our previous study [10] on this topic.

2.2. Preprocessing

Noise in the ECG signals is a significant problem. There are numerous noise factors in the ECG: EMG noise, power line noise, baseline wander, and composite noise [9]. Instabilities in the amplitude of ECG signals have a negative effect on the calculated feature vectors. The differences in ECG signals are minimized by performing normalization and pre-processing operations.

In this study signal mean was set to zero. Thereafter a median filter was used to reduce noise. A cascade low-pass filter was applied in the final signal to remove frequency components below 0.5 and 2 Hz [10]. Fig. 2 shows the block diagram of the cascade filter.

2.3. Beat Parsing

200 points was established of each beat's window length from the filtered ECG signal according to the location of the R point in the QRS. The related location of the R points collected the annotation files of the MIT-BIH Database. The selected beats constituted a 7000 × 200 data matrix [10].

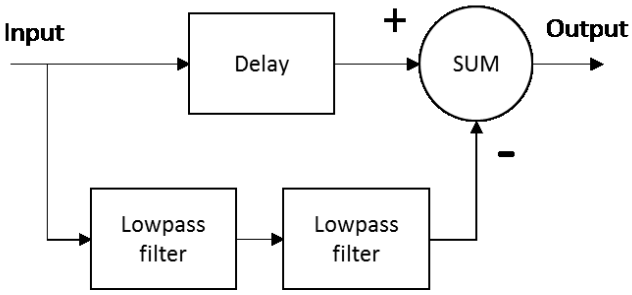

Fig. 2. Block diagram of cascade filter.

2.4. Feature Selection with Genetic Algorithm

Pattern classification and knowledge discovery problems require selection of a subset of features to characterize the patterns to be classified[11]. Representing the data with a subset of the dataset can be improved performance of the classifier. In this work GA used to reduce the size of feature vector.

A standard genetic algorithm [12] with tournament based selection strategy was used in this work. The achieved results are based on 10-fold cross-validation for each classification task with the following parameter settings:

- Population size: 5
- Maximum number of generations: 30
- Probability of crossover: 0.5
- Probability of mutation: -1.0
- Minimum number of attributes: 15
- Maximum number of attributes: 25 – 50 – 100

Use of large number of attributes is not feasible because of exhaustive search. Therefore, different selection operation was made limiting the maximum number of attributes 25, 50, and 100 and the results were evaluated. When the maximum number of attributes parameter limited to 100, 50, and 25, GA selected 95, 50, and 25 samples as a subset, respectively. Fig. 3 shows times series of one normal beat and GA selected 25 samples on it.

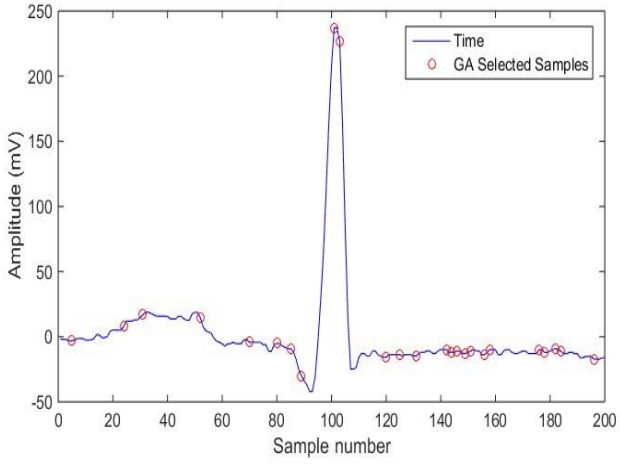

Fig. 3. Time series of a normal beat and its GA selected 25 points.

2.5. Classification

In this work, NN, k-NN, and SVM classification algorithms were used for classification. A three-layered feed forward NN was used for classification. The input layer was composed of 200 nodes corresponding to the 200 points of one beat. Furthermore, results of the GA feature selection were also tested using this method. In that case, the sizes of the input layer were 95, 50, and 25. The output layer consisted of two nodes corresponding to two class labels.

k-NN was used as a classifier in this work. k value of the k-NN classifier was selected as one after the parameter optimisation stage. Euclidean distance function was used as the measure function. Another popular classifier especially for binary classification in machine learning was applied for experimental results in this work. Following parameter set was used SVM classification stage: C=100, Gamma=4, and kernel-type=polynomial [10].

3. Experiments and Results

The proposed approach was tested on 200 time series samples of one beat and selected features with GA as an input feature vector. A parameter optimization step was implemented to obtain optimum parameter values. In the NN classifier, a hidden layer consisting of 10 neurons was used. The size of the hidden layer was selected by empirical observation. The NN was trained by a back propagation algorithm. At the training and testing stage, training cycle and learning rate parameters were set as 500 and 0.3, respectively. The error threshold parameter was set as 0.00001.

In the k-NN classification stage a grid search was applied to find best k value. The tests showed that the best k value of the k-NN algorithm was found at one; however, all k values in the test range achieved high results. k-NN classifier achieved highest classification accuracies in this work.

For the SVM classification experiments, parameters were determined using a grid search like that done with the k-NN experiments. After the optimization stage following parameters was found: the kernel function = polynomial, C=100, Gamma=4.

A 10-fold cross-validation method was used in this study for training and testing of the classification algorithms. The average accuracy of these 10 trials was calculated as a classification result. The cross-validation accuracy is the percentage of data which are properly classified.

The classification performance of the classifiers can be measured by calculating the accuracy, sensitivity, and specificity. These performance parameters are defined as shown in Equations (1)-(3).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

where TP and TN symbolize the total number of correctly classified PVC beat (true positive) samples and N beat (true negative) samples. The FP and FN symbolize the total number of misclassified PVC beat (false positive) samples and N beat (false negative) samples.

Table 1. Classification performances of classifiers for time series and GA selected features.

Classifier	Feature	Acc.	Sen.	Spe.
k-NN	Time	99.56%	99.29%	99.83%
	GA95	**99.69%**	**99.46%**	**99.91%**
	GA50	99.66%	99.43%	99.89%
	GA25	99.54%	99.23%	99.86%
NN	Time	98.10%	99.06%	97.14%
	GA95	98.53%	98.91%	98.14%
	GA50	97.86%	99.11%	96.60%
	GA25	98.01%	99.00%	97.03%
SVM	Time	98.09%	96.86%	99.31%
	GA95	98.10%	96.89%	99.31%
	GA50	98.11%	96.91%	99.31%
	GA25	97.70%	96.57%	98.83%

Table 1 shows the classification performance parameters (accuracy, sensitivity, and specificity) of classifiers using the time series of the one beat signal and GA selected features as an input vector. Classification results showed that k-NN classifier achieved the highest accuracy for the 95 selected features from GA. The other GA features (50 and 25 features) gained high classification accuracies for k-NN classifier.

4. Conclusions

In this paper, an approach was proposed to correctly classify PVC beats. The accuracy, sensitivity, and specificity were calculated in order to compare the training algorithms. When compared to our previous work [10] on this topic, this study used GA to reduce the size of the feature vector and achieved slightly better classification rates. GA selected features and time series of the signal was used to classify PVC beat. k-NN, NN and SVM classification methods was used with different schemes.

In terms of recognition accuracy, it can be seen that the k-NN classification algorithm using GA features accomplished the best performance according to the experiments. This study showed that high classification accuracy can be achieved without implementing any feature extraction method and by using time series of the signal for input. GA can be used to reduce the size of the input vectors representing the data.

5. References

[1] N. Tosun, Y. Erkoç, T. Buzgan, B. Keskinkılıç, D. Aras, N. Yardım, S. Gögen, G. Sarıoğlu, and M. Soylu, "Türkiye Kalp ve Damar Hastalıklarının Önleme ve Kontrol Programı (2010-2014)," Anıl Matbaası, Ankara, 2014.

[2] G. K. Lee, K. W. Klarich, M. Grogan, and Y.-M. Cha, "Premature ventricular contraction-induced cardiomyopathy: a treatable condition.," *Circ. Arrhythm. Electrophysiol.*, vol. 5, no. 1, pp. 229–36, Feb. 2012.

[3] S. Ittatirut, A. Lek-Uthai, and A. Teeramongkonrasmee, "Detection of Premature Ventricular Contraction for real-time applications," in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2013*, 2013.

[4] O. T. Inan, L. Giovangrandi, and G. T. A. Kovacs, "Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval

features.," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12 Pt 1, pp. 2507–15, Dec. 2006.

[5] S. Nahar and S. Bin Munir, "Automatic detection of premature ventricular contraction beat using morphological transformation and cross-correlation," in *3rd International Conference on Signal Processing and Communication Systems, ICSPCS'2009 - Proceedings*, 2009.

[6] I. Christov and G. Bortolan, "Ranking of pattern recognition parameters for premature ventricular contractions classification by neural networks," *Physiol. Meas.*, vol. 25, no. 5, pp. 1281–1290, Oct. 2004.

[7] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database.," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.

[8] G. Moody and R. Mark, "The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it," in *[1990] Proceedings Computers in Cardiology*, 1990, pp. 185–188.

[9] K.-M. Chang, "Arrhythmia ECG noise reduction by ensemble empirical mode decomposition.," *Sensors (Basel).*, vol. 10, no. 6, pp. 6063–80, Jan. 2010.

[10] Y. Kaya and H. Pehlivan, "Classification of Premature Ventricular Contraction in ECG," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, pp. 34–40, 2015.

[11] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 44–49, Mar. 1998.

[12] M. Mitchell, "Genetic algorithms: An overview," *Complexity*, vol. 1, no. 1, pp. 31–39, 1995.