ECG Beat Classification by Using Discrete Wavelet Transform and Random Forest Algorithm

Nahit Emanet[1]

[1]Computer Engineering Department, Fatih University, Istanbul, Turkey.

[1] emanetn@fatih.edu.tr

Abstract

Until now, there has been no study in the literature that uses Random Forest algorithm for the classification of ECG beats. In this study, the ECG signals obtained from the MIT/BIH database were used to classify the five heartbeat classes (N, L, R, V, P). Feature extraction from the ECG signals for classification of ECG beats was performed by using discrete wavelet transform (DWT). The Random Forest was then presented for the classification of the ECG signals. Five types of ECG beats were classified with a success of 99.8%. Since Random Forest algorithm works very fast, gives excellent performance and there is no cross validation, it can be useful for long-term ECG beat classification.

1. Introduction

The electrocardiogram (ECG) is a non-invasive diagnostic and monitoring tool that records the electrical activity of the heart at the body surface. It provides very accurate information about the performance of the heart and cardiovascular system. The heart generates an electrochemical impulse, initiated by a group of nerve cells called the sinoatrial node (SA) that results in a process called depolarization. Depolarization is propagated from cell to cell across the entire heart. This wave of depolarization causes the cells contract and relax in a timely order and makes the heart beat. Because this action represents a flow of electricity, it can be measured by skin electrodes, placed at designated locations on the surface of the body, in the form of ECG signal. The pattern of electrical propagation is not random, but spreads over the structure of the heart in a coordinated pattern. A typical ECG signal waveform of a normal heart beat is shown in Figure 1. The ECG signal is characterized by six upward and downward voltage reflections. The first upward deflection, P, is due to atrial complex. Other deflections, Q, R, S, T, are all due to the action of the ventricular complexes. Any deviation from the norm in a particular ECG measurement is an indication of possible heart disease or abnormality. Early detection of heart diseases enables patients to enhance the quality of their life through more effective treatments. Therefore, numerous researches have been conducted in an attempt to analyze and classify the ECG signal [1-6]. A heart disease can be identified by knowing the classification of heartbeats. However, this is very tedious task, because some heart diseases appear infrequently, and very long ECG measurements are needed to capture them. Analysis of such a large number of data is very time consuming, thus automated analysis and classification can be very helpful.

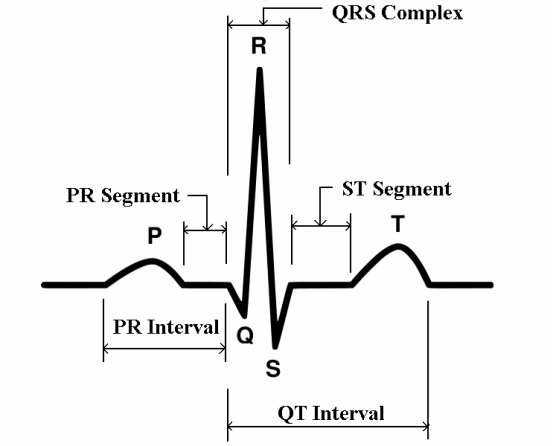

Figure 1*: A typical ECG signal waveform*

Figure 2 depicts the stages of an automated ECG analysis suitable for heartbeat classification. It consists of three stages: a preprocessing stage, a processing stage, and a classification stage. The preprocessing stage removes artifact signals from the ECG signal. These artifact signals include baseline wander, power line interference, and high-frequency noise. The preprocessing stage is of great importance since it contributes significantly to the overall classification result.The processing stage consists of heartbeat detection and feature extraction modules. The heartbeat detection module attempts to locate all heartbeats. The feature extraction module forms a feature vector from each heartbeat. The feature extraction modules are required, because greater classification performance is often achieved if a smaller number of discriminating features are first extracted from the ECG. The classification stage contains one or more classifier units which select one of the required classes in response to the input feature vector. The most difficult problem faced during automated ECG analysis is that there is a great variety of morphologies among the heartbeats belonging to one class, even for the same patient. Moreover, heartbeats belonging to different classes are morphologically similar to each other. A number of classification systems have been previously reported by other researchers. These methods include linear discriminant systems [7], back propagation neural networks [1], self organizing maps (SOM) [8], learning vector quantization (LVQ) [8], support vector machines (SVM) [9], fuzzy or neuro-fuzzy systems [4], and the combination of different neural-based solutions, so-called hybrid systems [3].

9781-4244-3428-2/09/$25.00 ©2009 IEEE

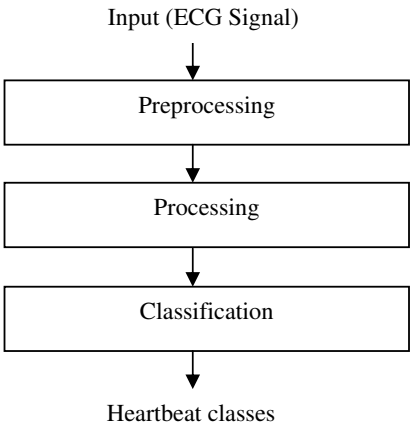

Figure 2*: Heartbeat classification stages*

Until now, there has been no study in the literature that uses random forests for the classification of ECG beats. In the present study, the ECG signals obtained from the MIT/BIH database [10] were used. Feature extraction from the ECG signals for classification of ECG beats was performed by using discrete wavelet transform (DWT) [11]. The Random Forests [12] were then presented for the classification of the ECG signals. The Random Forest classifies the five classes (N, L, R, V, P) of the ECG signals when discrete wavelet coefficients defining the behaviour of the ECG signals were used as inputs.

The organization of the paper is as follows. Section 2 gives a description of the dataset used for training and testing. Section 3 provides a detailed description of the Feature extraction stage. The Random Forest algorithm is presented in Section 4. The correct classification rates and convergence rates of the system are examined and then performance of the system is reported in Section 5. Finally, some conclusions are drawn concerning the classification of the ECG signals in Section 6.

2. Data Set

The 48 records from MIT/BIH ECG arrhythmia database were used for the development and evaluation of the Random Forest classifier. This database comprises 48 half-hour two-channel ECG recordings obtained from 47 subjects with several types of arrhythmias. Twenty-three records of this database were chosen randomly from a set of over 4,000 recordings collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital (BIH), and they are intended to serve as a representative sample of routine clinical recordings; the remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias like ventricular, junctional, and supraventricular arrhythmias. The subjects were 25 men aged 32–89 years, and 22 women aged 23–89 years (two records came from the same male subject). In most records, the upper signal is a modified limb lead II, obtained by placing the electrodes on the chest.

Each record in the MIT/BIH database has an annotation file in which each ECG beat has been identified by expert cardiologist. These labels are used in training the classifiers and also to evaluate the performance of the classifiers in testing phase. The availability of annotated MIT/BIH database has enabled the evaluation of performance of the proposed beat classification algorithm.

We tried to remove the baseline wandering by using median filtering as a preprocessing stage.

3. Feature Extraction Using Wavelet Transform

Original ECG signal vectors were formed by 256 discrete data in the intervals of R–R for all arrhythmias so that each feature vector contains a single ECG beat. When feature vectors are only formed by the magnitudes of the ECG signals, computational cost is low. However, feature vectors are affected by the determination of the R-peak position. If the R-peak position is not correctly determined, vectors will scatter in the feature space.

Mathematical transformations are applied to the signals to obtain further information from the signal that is not readily available in the raw signal. Since feature vectors are formed by using transforms, computational cost is high. Transformation methods prevent the scattering of vectors in the feature space.

Some of the most widely used transformations are linear transformations, such as principle component analysis (PCA) and linear discriminant analysis (LDA). Although PCA and LDA are very commonly used, they are not necessarily the best ones. In fact, depending on the application, Fourier-based transformation may be more appropriate. For nonstationary signals, a wavelet-based time-frequency representation may be the better feature extraction technique. The basic idea of the wavelet transform is to represent any arbitrary function f as a superposition of wavelets. Any such superposition decomposes f into different scale levels, where each level is then further decomposed with a resolution adapted to the level. In practice, it is easy to define f as a discrete superposition, hence a discrete wavelet transform (DWT).

In order to obtain our wavelet analysis, we used the Matlab program, which contains a very good "wavelet toolbox". In our analysis, Daubechies db2 wavelet [13] was used, because it gives better accuracy compared to other wavelets [14]: Symmlet sym6, Symmlet sym10, Coiflet coif2, Coiflet coif4, Daubechies db1, Daubechies db6. For each ECG feature vector, formed by 256 discrete data in the intervals of R–R for all arrhythmias, the detail wavelet coefficients (dk, k = 1, 2, 3, 4) at the first, second, third and fourth levels (129 + 66 + 34 + 18 coefficients) and the approximation wavelet coefficients ($a4$) at the fourth level (18 coefficients) were computed. Then 265 wavelet coefficients were obtained for each ECG segment. These coefficients were presented as an input feature vector to the Random Forest algorithm.

4. Random Forest

Random Forest is an ensemble of unpruned classification trees. It gives excellent performance on a number of practical problems, because it is not sensitivite to noise in the data set,

and it is not subject to overfitting. It works fast, and generally exhibits a substantial performance improvement over many tree-based algorithms.

The classification trees in the Random Forest are built recursively by using the Gini node impurity criterion which is utilized to determine splits in the predictor variable. A split of a tree node is made on variable in a manner that reduces the uncertainty present in the data and hence the probability of misclassification. Ideal split of a tree node occurs when Gini value is zero. The splitting process continues until a "forest", consisting of multiple trees, is created. Classification occurs when each tree in the forest casts a unit vote for the most popular class. The Random Forest then chooses the classification having the most votes over all the trees in the forest. Pruning is not needed as each classification is produced by a final forest that consists of independently generated trees created through a random subset of the data, avoiding over fitting. The generalization error rates depend on the strength of the individual trees in the forest and the correlation between them. This error rate converges to a limit as the number of trees in the forest becomes large.

Another advantage of RF is that there is no cross validation or a separate test set to get an unbiased estimate of the test set error. Test set accuracy is estimated internally in RF by running out-of-bag (OOB) samples. For every tree grown in RF, about one-third of the cases are out-of-bag (out of the bootstrap sample). The out-of- bag (OOB) samples can serve as a test set for the tree grown on the non-OOB data.

Random Forest (RF) algorithm can be summarized as follows:

1. A number n is specified much smaller than the total number of variables N (typically $n \approx \sqrt{N}$)
2. Each tree of maximum depth is grown without pruning on a bootstrap sample of the training set
3. At each node, n out of the N variables are selected at random
4. The best split on these n variables is determined by using Gini node impurity criterion.

Reducing n reduces the strength of the individual trees in the forest and the correlation between them. Increasing it increases both. Using the OOB error rate, an optimum value of n can be found. This is the only adjustable parameter to which random forests is sensitive.

The computational complexity for each tree in RF is $\sqrt{N}\, S\, log(S)$, where S is the number of the training cases. Therefore, it can handle very large number of variables with moderate number of observations.

5. Experimental Results

Classification results of the Random Forest are displayed by confusion matrix. A confusion matrix displays the number of correct and incorrect predictions made by the classifier compared with the actual classifications in the data set. The confusion matrix is n-by-n, where n is the number of classes. Each column of the matrix represents the predictions, while each row represents the actual classifications. By using confusion matrix, one can see if the system is mislabeling classes. The confusion matrices showing the classification results of the Random Forest algorithm for the ECG beats are given in Table 1 and Table 2 for training and test sets, respectively. The waveforms of five different ECG beats classified in the study are normal beat (N), left bundle branch block beat (L), right bundle branch block beat (R), premature ventricular contraction (V), paced beat (P). Training and test sets are formed by data obtained from records 100, 106, 107, 109, 111, 118, 124, 201, 202, 210, 212, 213, 214, 217, 219, 220, 221, 231 of MIT-BIH database. For five classes, training set is formed by choosing 600 vectors (120 vectors from each class), and test set is formed by 300 vectors (60 vectors from each class). The final error rates are 0.41% and 0.16% for the training and test set, respectively. The number of variables to split on at each node, n, was 12. This value gave the smallest OOB error rate.

Table 1*: Training set confusion matrix*

	P	L	R	V	N
P	240	0	0	0	0
L	0	239	0	1	0
R	0	0	239	1	0
V	0	1	0	237	0
N	0	0	1	1	240

Table 2*: Test set confusion matrix*

	P	L	R	V	N
P	120	0	0	0	0
L	0	119	0	0	0
R	0	0	120	0	0
V	0	0	0	120	0
N	0	1	0	0	120

6. Conclusions

In this study, Random Forest algorithm for the classification of ECG beats was presented for the first time in the literature. Five types (N, L, R, V, P) of ECG beats were classified with a success of 99.8%. Since Random Forest algorithm works very fast, gives excellent performance and there is no cross validation, it can be useful for long-term ECG beat classification.

7. References

[1] Yeap T.H, Johnson F., and Rachniowski M., "ECG Beat Classification by a Neural Network", Proceedings Annual International Conference of the IEEE EMBS Society, 167-173, 1990.

[2] Hu Y.H., Palreddy S., and Tompkins W.J., "A patient-Adaptable ECG Beat Classifier Using a Mixture of Experts Approach", *IEEE Trans. on Biomedical Engineering*, vol. 44, 891–900, 1997.

[3] Dokur Z., and Ölmez T., "ECG Beat Classification by a Novel Hybrid Neural Network", *Comput. Methods Programs Biomed.*, vol. 66, 167-181, 2001.

[4] Özbay Y., Ceylan R, and Karlik B., "A Fuzzy Clustering Neural Network Architecture for Classification of ECG Arrhythmias", *Comput. Biol. Med.*, vol. 36, 376–388, 2006.

[5] Engin M., "ECG Beat Classification using neuro-fuzzy network", *Pattern Recog. Lett.*, vol. 25, 1715-1722, 2004.
[6] Chazal P., and Reilly R.B., "A Patient-adapting Heartbeat Classifier using ECG Morphologhy and Heartbeat Interval Features", *IEEE Trans. Biomedical Engineering*, vol. 53, 2535-2543, 2006.
[7] Yeh Y.C., Wang W.J., and Chiou C. W., "Cardiac Arrhythmia Diagnosis Method Using Linear Discriminant Analysis on ECG Signals" *Measurement*, vol. 42, 778-789, 2009.
[8] Palreddy S., Tompkins W.J., and Hu Y.H., "Customization of ECG Beat Classifiers Developed using SOM and LVQ", *Engineering in Med. And Biol. Soc.*, vol. 1, 20-25, 1995.
[9] Ülbeyli E.D., "ECG Beats Classification Using Multiclass Support Vector Machines with Error Correcting Output Codes", *Digital Signal Processing*, vol. 17, 675- 684, 2007.
[10] Physiobank Archieve Index, MIT-BIH Arrhythmia Database,http://www.physionet.org/physiobank/database 2009.
[11] Mallat S., A Wavelet Tour of Signal Processing, Academic Press, 1997.
[12] Breiman, L., "Random Forests", *Machine Learning*, vol. 45, 5-32, 2001.
[13] Daubechies, I., "The Wavelet Transform, Time-Frequency Localization and Signal Analysis", *IEEE Trans.on Inform. Theory,* vol. 36, 961-1005, 1990.
[14] Güler I., and Ülbeyli E.D., "ECG Beat Classifier Designed by Combined Neural Network Model", *Pattern Recognition*, vol. 38, 199- 208, 2005.