

Investigating Cardiac Arrhythmia in ECG using Random Forest Classification

R. Ganesh Kumar
Research Scholar,
Dept.of CSE, Sathyabama University,
Chennai, INDIA.

Dr. Y S Kumaraswamy
Senior Professor & HOD, Dept. of MCA.,
Dayananda Sagar College of Engineering,
Bangalore, INDIA.

ABSTRACT

Electrocardiogram (ECG) is used to assess the heart arrhythmia. Accurate detection of beats helps determine different types of arrhythmia which are relevant to diagnose heart disease. Automatic assessment of arrhythmia for patients is widely studied. This paper presents an ECG classification method for arrhythmic beat classification using RR interval. The methodology is based on discrete cosine transform (DCT) conversion of RR interval. The RR interval of the beat is extracted from the ECG and used as feature. DCT conversion of RR interval is applied and the beats are classified using random tree. Experiments were conducted using MIT-BIH arrhythmia database.

KEYWORDS

ECG, ECG Arrhythmia classification, MIT-BIH ECG data, RR interval, DCT, Random forest.

1. INTRODUCTION

The electrocardiogram (ECG) is the bioelectrical activity signal of the heart which represents the cyclic rhythm of contraction and relaxation of the heart muscles. Arrhythmia is an irregular heartbeat; it causes irregular rhythms such as slow or fast heartbeat [1]. Arrhythmia forms the basis of heart disease diagnosis, which may indicate the serious problems in the heart. The problem faced using ECG for heart disease diagnosis is that the normal ECG reading differs from person to person and similarly the arrhythmia also differs, thus complicating the heart disease diagnosis. Diagnosis of certain arrhythmia by visual inspection takes time and is a tedious process. Thus, automatic classification of heartbeat is useful and helps medical professionals for fast and accurate diagnosis of heart beat. Utilization of pattern classifier techniques can improve ECG arrhythmia diagnoses.

The ECG signal is acquired through number of electrodes. The ECG recordings contain noise and the amplitude of the beats varies from person to person. Thus preprocessing of the signal is required for beat detection and feature extraction. The P, Q, R, S and T waves of the ECG signal contains all important features [2]. A sample ECG waveform is shown in figure 1. The features of these waves are used to classify different types of arrhythmia. The ECG signals can be divided into vital and plain parts. The most important information in the ECG signal is concentrated in the P wave, QRS complex and T wave [3].

The diagnoses of disease are generally based on the position and magnitudes of PR interval and segment, ST interval and segment, QRS interval and QT interval.

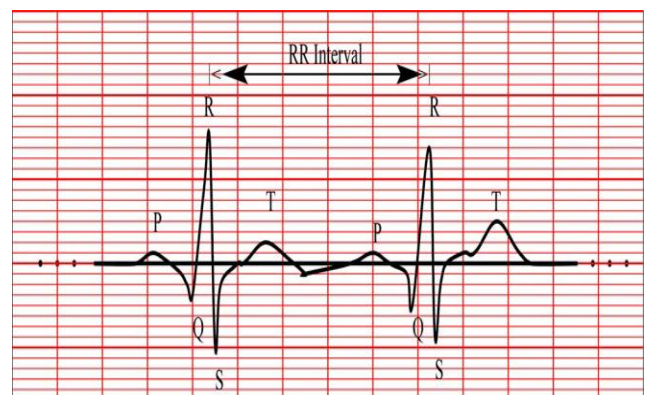


Fig 1: A sample ECG signal with P, Q, R, S and T marked.

The heartbeats are classified into six types as

- Normal beats (N)
- Premature Ventricular Contractions (PVC)
- Paced beats (PB)
- Atrial Premature Beat (APB)
- Left Bundle Branch Block beats (LBBB)
- Right Bundle Branch Block beats (RBBB)

Arrhythmia is mainly determined by the QRS complex wave. ECG signal analysis includes QRS waveform detection, feature selection and heartbeat classification. There are various studies analyzing the ECG signal for cardiac arrhythmia. Studies based on QRS pattern [4, 5], automatic detection of discerning arrhythmia using Hidden Markov models [6], wavelet transform [7], Hermite function [8] and Neural networks [9,10]. Feature selection is crucial for efficient automatic classification of ECG. The features are either selected from the time domain or frequency domain. The wavelet transform methods are widely used for denoising of the ECG signal [11, 12]. The wavelet transforms decompose the ECG signal into various components of different scales. The phase information is maintained due to the linear operation of the transforms. Wavelet functions support symmetry and compactness and have shown to achieve high accuracy for ECG signals.

In this work, it is proposed to investigate the classification accuracy of random forest classifier on ECG data. The time series data obtained from MIT-BIH is converted to frequency domain using Discrete Cosine Transform (DCT) for feature extraction. Two types of arrhythmia namely LBB and RBB are considered in this work.

2. LITERATURE REVIEW

Khoureich Ka, [13] proposed a method of ECG beat classification based on waveform similarity and RR interval. The proposed classifier is patient-specific classifier. A beats database of the patients ECG is created for the classifier. In the proposed method the wavelet transform based techniques were used to denoise the ECG signal. And also a java implementation to reduce noise due to the baseline wander cancellation and high oscillation noise is used. RR intervals are extracted and used as features. Classification is done using neural network or fuzzy logic. Experimental results achieved high accuracy for classification of beats.

Tsipouras, et al., [14] proposed an efficient method based on use of knowledge-based systems for arrhythmia beat classification and arrhythmic episode detection and classification. The proposed method was based on RR interval extracted from the ECG signal. A set of rules and a deterministic automaton based on RR interval behavior during arrhythmic episodes is formed, which is used to classify beats and arrhythmic episode classification. The proposed method was evaluated using MIT-BIH arrhythmia database. Experimental results showed that the proposed method achieved high accuracy in both beat and episode classification.

In this paper, it is proposed to classify arrhythmia of ECG signals. The ECG data for the arrhythmia classification is obtained from MIT-BIH PhysioBank arrhythmia database. The data is preprocessed to find RR interval, DCT is applied to convert the RR intervals and Random tree is used for classification. The paper is organized in the following sections, section 1 contains introduction, section 2 discusses the materials and methods used in this paper, section 3 describes the experimental setup and section 4 presents the results and conclusions.

3. METHODOLOGY

The QRS complex along with the RR interval plays an important role in arrhythmia identification [15]. The peaks are detected as follows:

- The moving average of the signal is calculated using n number of records.
- New signal is derived by subtracting the moving average from the original signal.
- Peak of the signal 'R' is found.
- Peaks of P, Q, S and T are found by relative position.

The peaks amplitude is measured from k line. The k line is given by:

$$k = \text{Max}(\theta_i, i = 1, 2, \dots, 11) + c$$

where θ is the greatest amplitude, i type of heartbeat, and c is a constant.

The ECG datasets signals for training and testing are taken from MIT-BIH arrhythmia database (mitdb) [16]. The selected Arrhythmias are LBBB (Left Bundle Branch Block), RBBB (Right Bundle Branch Block) and normal. Each ECG beat used is a matrix (275x1) for one ECG lead. Every ECG signal has five distinct points (P, Q, R, S and T) used for the interpretation of the ECG as shown in figure 2.

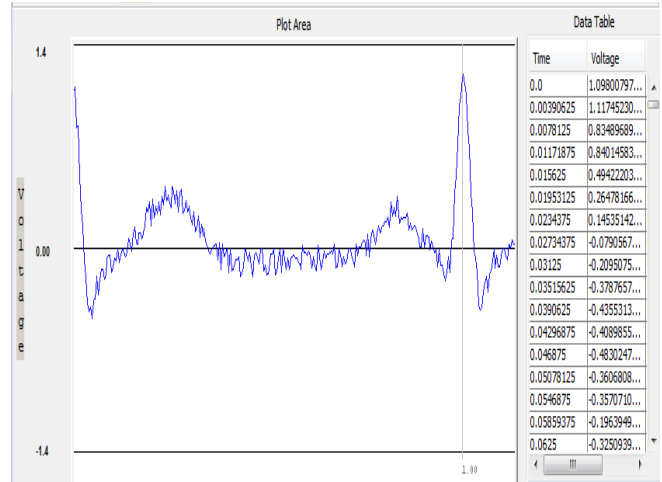


Fig 2: Sample of ECG signal with P, Q, R, S and T peaks in the dataset used

DCT translates time series signal into basic frequency components. The ability of DCT to compress signals is advantageous in ECG signal classification. The process of decomposing an image into a set of cosine basis functions is called forward discrete cosine transform (FDCT) and process of reconstructing is called inverse discrete cosine transform (IDCT).

The FDCT [17] of a list of n real numbers $s(x)$, $x = 0, \dots, n-1$, is the list of length n given by:

$$S(u) = \sqrt{2/n} C(u) \sum_{x=0}^{n-1} s(x) \cos \frac{(2x+1)u\pi}{2n}$$

where $C(u) = 2^{-1/2}$ for $u=0$ or otherwise $C(u) = 1$

The constant factors are chosen so that the basis vectors are orthogonal and normalized.

The inverse cosine transform (IDCT):

$$S(x) = \sqrt{2/n} \sum_{u=0}^{n-1} C(u) s(u) \cos \frac{(2x+1)u\pi}{2n}$$

where $C(u) = 2^{-1/2}$ for $u=0$ or otherwise $C(u) = 1$

Growing an ensemble of random trees for classification using a probabilistic scheme is called random forest of trees. Classification accuracy is high as the trees vote for the most popular class. Trees drawn at random from a set of possible trees is called random tree. Random tree is a decision tree that considers K randomly chosen attributes at each node. The class

probabilities on each node are based on backfitting with no pruning [18]. The steps involved in growing a random tree are:

- The training set for growing the tree is obtained by selecting N cases at random but with replacement from the original dataset.
- A random number of attributes m are chosen for each tree. These attributes form the nodes and leafs using standard tree building algorithms. The best split on m is used to split the nodes and m is held constant.
- Each tree is grown to the fullest extent possible without pruning.

A new object is classified using its input vector down each of the trees in the forest. The forest chooses the class with most votes, and new object input vector is classified. The advantages of random forests are that it is easy to calculate a measure of “variable importance” and out of bag data can be used to estimate the classification error. The error rate of the random forest depends upon the correlation between the trees in the forest and the strength of the trees. Higher the correlation between two trees, higher is the error rate. A tree is a strong classifier if it has lower error rate, increasing the strength of each tree decreases the error rate of the random forest.

4. RESULT AND DISCUSSION

The extracted beats contain 68 instances of left bunch bundle block, 30 instances of right bunch bundle block and 56 normal instances. Using 10 fold cross validation 1,5,10 and 20 trees were constructed. The classification accuracy is shown in figure 3.

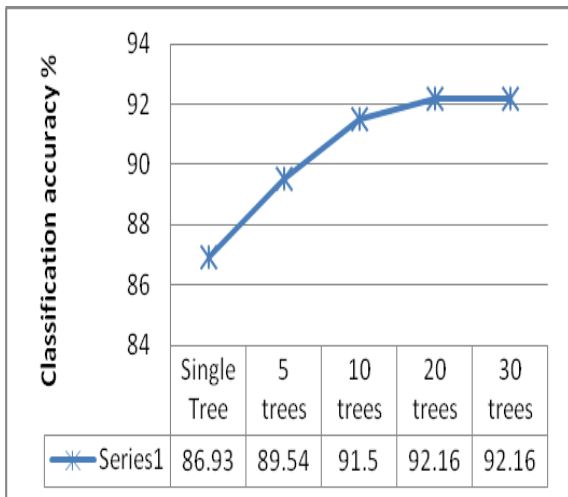


Fig 3: Classification accuracy of random forest with different number of trees.

The root mean squared error is shown in figure 4.

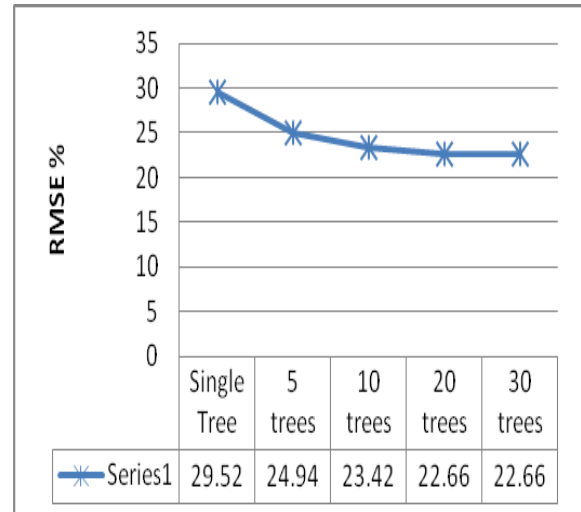


Fig 4: Root mean squared error.

From the figure 3, it is seen that as the number of trees used in the computation crosses 20, the classification accuracy remains constant with no further increase. The optimal number of trees to get the best classification accuracy can be in the range of 15 to 20 trees based on the time within which the data needs to be computed.

5. CONCLUSION

In this paper, it was proposed to implement a frequency domain method using Discrete Cosine Transform (DCT) for feature extraction. Two different arrhythmia conditions were investigated along with normal ECG. ECG data for this investigation was obtained from phyiobank website. The average classification accuracy obtained by the proposed method is over 90%. Future direction should focus on data reduction to improve the classification time.

6. REFERENCES

- [1] Sandoe E, Sigurd B. Arrhythmia – A Guide to Clinical Electrocardiology. Bingen: Publishing Partners Verlags GmbH, 1991.
- [2] J. Pan and W. J. Tompkins, “A Real-time QRS detection algorithm,” IEEE Trans. Biomed. Eng., vol. BME-32, no. 3, pp. 230–236, Mar. 1985.
- [3] R.M. Rangayyan, Biomedical Signal Analysis: A Case-Study Approach, Wiley–Interscience, New York, 2001.
- [4] I. Christov, G. Gómez-Herrero, V. Krasteva, I. Jekova, A. Gotchev, K. Egiazarian, Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification, Med. Eng. Phys. 28 (2006) 876–887.
- [5] P. Chazal, M. O’Dwyer, R.B. Reilly, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Trans. Biomed. Eng. 51 (2004) 1196–1206.
- [6] A. Koski, Modelling ECG signals with Hidden Markov Models, Artif. Intell. Med. 8 (1996) 453–471.

- [7] C.W. Li, C.X. Zheng, C.F. Tai, Detection of ECG characteristic points using wavelet transform, *IEEE Trans. Biomed. Eng.* 42 (1995) 21–28.
- [8] P. Laguna, R. Jane, S. Olmos, N.V. Thakor, H. Rix, P. Caminal, Adaptive estimation of QRS complex wave features of ECG signal by the Hermite model, *Med. Biol. Eng. Comput.* 34 (1996) 58–68.
- [9] Z. Dokur, T. Olmez, ECG beat classification by a novel hybrid neural network, *Comput. Methods Program Biomed.* 66 (2001) 167–181.
- [10] Y.P. Meau, F. Ibrahim, S.A.L. Narainasamy, R. Omar, Intelligent classification of ECG signal using extended EKF based neural fuzzy system, *Comput. Methods Program Biomed.* 82 (2006) 157–168.
- [11] J. P. Martinez, S. Olmos, and P. Laguna, “Evaluation of a wavelet-based ECG waveform detector on the QT database,” *Comput. Cardiol.*, vol. 27, pp. 81–84, Sep. 2000.
- [12] P. M. Agante and J. P. Marques de Sa, “ECG noise filtering using wavelets with soft-thresholding methods,” *Comput. Cardiol.*, vol. 26, pp. 535–538, Sep. 1999.
- [13] Ahmad Khoureich Ka, ECG beats classification using waveform similarity and RR interval, Arxiv preprint arXiv:1101.1836, 2011.
- [14] Tsipouras MG, Fotiadis DI, Sideris D. An arrhythmia classification system based on the RR interval signal. *Artif Intell Med* 2005;33:237—50
- [15] M. Engin, “ECG beat classification using neuro-fuzzy network,” *Pattern Recognition Letters* 25 (2004) 1715-1722.
- [16] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.* *Circulation* 101(23):e215-e220.
- [17] N. Ahmed, T. Natarajan, *Discrete-Time Signals and Systems*, Reston Publishing Company, 1983.
- [18] Frederick Livingston: Implementation of Breiman’s Random Forest Machine Learning Algorithm, in *ECE591Q Machine Learning conference*, Fall 2005.