Design of Classifier for Electrocardiography Classification

Duong Van Binh, Nguyen Thanh Nghia, Nguyen Thanh Hai and Nguyen Manh Hung

FEEE, HCMC University of Technology and Education, Vietnam
11141013@student.hcmute.edu.vn, 1627003@student.hcmute.edu.vn,
nthai@hcmute.edu.vn, hungnm@hcmute.edu.vn

Abstract. The Electrocardiography classifier is an essential tool for helping doctors in diagnosing early heart problems. This paper proposes with an electrocardiography classifier for analyzing accuracy in case of non-long-tail effect. Data are obtained from MIT-BIH arrhythmia database. Therefore, a discrete wavelet transform decomposition algorithm is employed for feature extraction and a principal component analysis is used for dimension reduction of data. In addition, the heart beat can be classified using a neural network method. In order to evaluate the classifier accuracy, the confusion matrix and Receiver Operating Characteristic curve are applied.

Keywords: Principal component analysis, Discrete wavelet transform decomposition, Neural networks, confusion matrix and Receiver Operating Characteristic curve.

1 Introduction

In recent years, dead people due to heart disease have fast increased. Early diagnosis of arrhythmia is very necessary for doctors. Electrocardiography (ECG) shows the electrical activity of heart changes over time through displaying on a screen or paper pages for presenting ECG data. Therefore, doctor can show clinical diagnosis of heart disease based on the ECG graphical presentation with waveform characteristics P, Q-R-S, T [1]. In this waveform, some of characteristics such as PR interval, PR segment, QRS complex interval, ST segment, ST interval, QT interval, and RR interval contains features of ECG data that doctor can use for diagnosis.

The noisy component in ECG data needs to be considered. In particular, noisy source in ECG data is from all of leads and the variety of frequency bands of the system. Thus ECG data need to be re-moved the noisy component by using filters. Some algorithms for removing ECG noise is often used such as baseline wander, low pass and high pass filters, wavelet filter, and discrete wavelet transform [2,3].

ECG data after removing the noisy component is converted into one beat, in which each heart beat is considered as one feature. Because dimensions of feature are large, some techniques are applied to reduce the feature dimensions. In particular, the techniques applied for dimensional reduction often are Principal Component Analysis

(PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA) [4,5]. The new ECG data with the low dimension will allows the action of classifier faster.

To determine the heart beat type, a classifier is applied for recognition. The classifier for recognition of ECG data can be one of the following kinds such as Support Vector Machine (SVM), Neural Networks (NNs), fuzzy logic, Hidden Markov Model (HMM) [6,7]. The ECG feature after dimensional reduction will be employed for training the classifier. In particular, a part of ECG feature will be used for testing in the classifier. Moreover, some algorithms are utilized for testing the classifier such as confusion matrix, accuracy index, and Receiver Operating Characteristic (ROC).

2 Materials and Methods

2.1 A. Heart beat database

ECG data for research is obtained from MIT-BIH data-base [8]. The ECG data have 48 ECG signals corresponding to 48 patients and are converted into each ECG heart-beat containing QRS complex. In addition, each ECG heartbeat has 200 samples, in which 100 samples in the left side and 99 samples in the right side of R peak. Basically, ECG signal has five heart disease types and every heart disease is considered as one class.

Table 1. ECG data for experiment

	Class 1	Class 2	Class 3	Class 4	Class 5	Total
Original ECG data	90249	2776	7226	802	6906	107959
ECG data with random extraction in each class	802	802	802	802	802	4010
ECG data with small class duplicated	90249	90249	90249	90249	90249	451245

ECG data has 2 channels and it is measured with difference leads (LMII, V1, V2, and V5). In this paper, only data on lead LMII is considered. In particular, the normal beat (class 1) is too large but another class is small, so the long-tail effect is always existed in classification [9]. ECG data is considered in two situations: Firstly, large ECG data will reduce randomly to equal to another class; Secondly, the class of small ECG data will be duplicated to equal to large class as shown in Table 1.

2.2 Discrete Wavelet Transform Decomposition algorithm

For extraction of feature of heartbeat disease, a Discrete Wavelet Transform Decomposition (DWTD) algorithm is applied. In particular, the DWTD coefficients of the ECG heartbeat are obtained, including detail coefficient and approximate coefficient and these coefficients are de-scribed as follows:

$$a_j(k) = \sum_{n=0}^{N-1} x[n]h(2k-n)$$
$$d_j(k) = \sum_{n=0}^{N-1} x[n]g(2k-n) \quad (1)$$

in which a_1 and d_1 are obtained when x[n] passes through the low pass filter and high pass filter. After that, a_2 and d_2 are obtained when a1 passes through the low pass filter and high pass filter again and it is similar for others.

in which, a_j, d_j are the approximate and detail coefficient at the j^{th} level.
$x[n]$ is the ECG heartbeat.
$h(2k\text{-}n)$, $g(2k\text{-}n)$ is the low pass, high filter.
N is the length of x[n].

In this paper, the Mayer wavelet function (dmey) is applied at the 4th level and the approximate coefficient a4 and detail coefficient d4 are the features of ECG heartbeat.

2.3 Feature Dimensional Reduction and Classifier

The ECG heartbeat feature is reduced the dimension using the PCA algorithm for high accuracy. After reduction of the dimension, six approximate coefficients and six detail coefficients are considered as twelve ECG features used for training and testing. In this paper, the PCA algorithm is applied for reducing dimensions of ECG data after the DWTD. Therefore, In the PCA, the covariance matrix C is calculated as follow:

$$C = (x - \bar{x})(x - \bar{x})^T \quad (2)$$

in which, $\bar{x}$ is a mean vector of x.

The matrix of eigenvectors U and diagonal matrix of eigenvalue D is described as follows:

$$U^{-1}CU = D \quad (3)$$

ECG data after reducing the dimension is obtained as the following equation:

$$y = [U^T(x - \bar{x})^T]^T \quad (4)$$

where y is one ECG data after using the PCA.

To identify the heartbeat type, a neural network algorithm is employed. The neural network model has three layers, including one input layer, one hidden layer and one output layer. In particular, the input layer consists of twelve neurons corresponding to twelve ECG features, the output layer has six neurons corresponding to six classes of the heartbeat type and the hidden layer uses ten nodes. In this neural network model, an error back-propagation method is utilized to update neural network weights. Thus, the maximum value of a Mean Square Error (MSE) method will be set 0.0001 and it is the error between the desired response and the actual response.

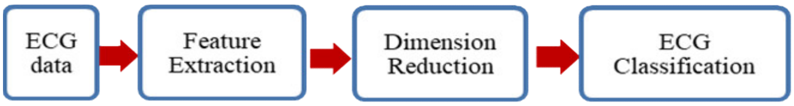

Fig. 1. The block diagram of the propose method

In order to classify ECG data, steps are shown as in Fig. 1. In particular, in the first stage, ECG data is split into heartbeats, the second stage is that ECG data feature extraction is executed using the DWTD algorithm, the next stage means that the ECG data dimensions is reduced using the PCA algorithm, and at the final stage, the neural network is employed to classify ECG heartbeats.

3 Result and Discussion

Features of ECG data are extracted using the DWTD with Mayer wavelet function, in which Approximate coefficients and detail coefficients at fourth level is obtained. The original ECG heartbeat, approximate coefficient (a_4) and detail coefficient (d_4) are described in Fig. 2.

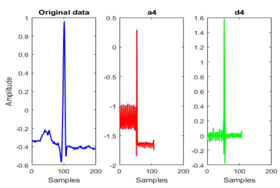

Fig. 2. Original data, approximate (*a*4) and detail (*d*4)

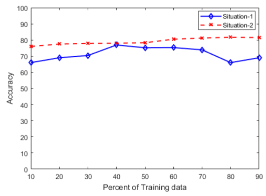

Fig. 3. The accuracy of the classifier

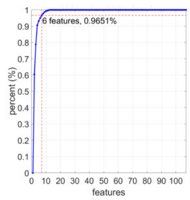

Fig. 4. The cumulative sum distributions of the relative weights

The ECG features of a_4 and d_4 consist of 214 samples corresponding to 214 dimensions of ECG data. To reducing the ECG data dimensions, a PCA algorithm is applied. Fig. 4 is shown the cumulative sum distribution of the information of data in approximate a_4. The information of data in first six features of a_4 is 96,5% as shown in Fig. 4. After reducing dimensions, coefficients of a_4 and d_4 are twelve dimensions and the ECG features is included in the classifier. Therefore, the neural network with twelve input neurons is applied for calculating the accuracy (ACC) of classifier as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

in which parameters are defined as the True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

The ECG heartbeats are classified for evaluating the performant of classifier and a confusion matrix is described as in Table 2.

Table 2. Structure of the confusion Matrix

	True condition		
Predicted		***Positives***	***Negatives***
	Positives	True positives (*TP*)	False positives (*FP*)
	Negatives	False negatives (*FN*)	True negatives (*TN*)

Table 3. The accuracy of classifier in case of ECG data with large class is extracted randomly

Training (%)	10	20	30	40	50	60	70	80	90
Testing (%)	90	80	70	60	50	40	30	20	10
ACC (%)	66.0	69.0	70.4	76.9	75.2	75.3	73.9	66.0	69.0

Table 4. The accuracy of classifier in case of ECG data with small class was duplicated

Training (%)	10	20	30	40	50	60	70	80	90
Testing (%)	90	80	70	60	50	40	30	20	10
ACC (%)	80.7	82.8	83.1	82.7	84.2	83.9	84.1	85.1	84.5

In this paper, the training and testing data are designed as in Table 3 and Table 4 for evaluation of the accuracy. In particular, In Table 3, the ECG data in large class (situation-1) is extracted randomly to equal another class. Because of the number of heartbeats in each class is small, the accuracy of classifier is low. While the ECG data in small class (situation-2) is duplicated for equal to large class, so the accuracy of this classifier is higher as described in Table 4. In addition, both cases above are recognized as the long-tail effect which is cancelled.

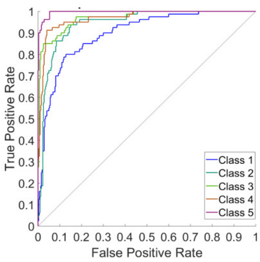

Fig. 5. ROC curve of classifier in situation-1

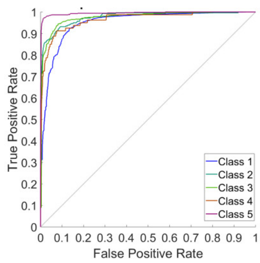

Fig. 6. ROC curve of classifier in situation-2

From Table 3. the accuracy of two classifiers are described in Fig. 3, the blue line with the accuracy in case of ECG data is extracted randomly in each class and the red line with the accuracy of ECG data is duplicated. As a result of the classifier, the accuracy of the ECG classifier of situation-2 is higher than that of situation-1.

In addition, the ROC curves of classifiers shown in Fig. 5 and Fig. 6 of Situation-1 and Situation-2 are to present the true positive rates and compare them together for evaluation. While authors of research in [5] just shows the high accuracy of classifier, but the number of heartbeat in each class is not mentioned. The original data in Table

1 shows that the number of normal heartbeats (in class 1) is 83.5% of the total heartbeat. Therefore, the long-tail effect still exists on the classifier. In this paper, ECG data on each class is equal, so the distribution of ECG data on each class is similar to the classifiers as shown in Fig. 5 and Fig. 6.

4 Conclusions

The DWT algorithm was applied for feature extraction of ECG data from MIT-BIH database in this study. Therefore, the PCA algorithm was applied to reduce the number of dimensions of feature vectors for accurate evaluation. The ECG heartbeats were classified using the neural network. The experimental results showed the better performance of the classifiers of two situations and the distribution of ECG data between classes. In addition, the accuracy of the classifiers with the non-long-tail effect by duplicating ECG data is a little lower long-tail effect, but the distribution of ECG data between classes is better.

Acknowledgement

The authors would like to acknowledge the support of Ministry of Education and Training, Vietnam with Grand No. B2017.SPK.03 and HCMC University of Technology and Education, Vietnam.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

1. Gajendra, E., Kumar, M.J.: A Novel Approach of ECG Classification for Diagnosis of Heart Diseases: Review. International Journal of Advanced Research in Computer Engineering & Technology **4**(11), 4096-4010 (2015).
2. Hesar, H.D., Mohebbi, M.: ECG Denoising Using Marginalized Particle Extended Kalman Filter with an Automatic Particle Weighting Strategy. IEEE Journal of Biomedical and Health Informatics **21**(3), 635-644 (2017).
3. J, V.M., Sadasiv, S.: A Comparative Study On Removal Of Noise In ECG Signal Using Different Filters. International journal of innovative research & development **2**(4), 915-927 (2013).
4. Bakir, C.: Classification of ECG Signals with the Dimension Reduction Methods. Journal of Mathematics and Statistical Science **Vol. 2017**(No. 12), 353-363 (2017).
5. Martis, R.J., Acharya, U.R., Min, L.C.: ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. Biomedical Signal Processing and Control **8**(5), 437-448 (2013).
6. Behadada, O., Chikh, M.: An interpretable classifier for detection of cardiac arrhythmias by using the fuzzy decision tree. Artificial Intelligence Research **2**(3), 45-58 (2013).
7. Liu, S.-H., Cheng, D.-C., Lin, C.-M.: Arrhythmia Identification with Two-Lead Electrocardiograms Using Artificial Neural Networks and Support Vector Machines for a Portable ECG Monitor System. Sensors (Basel, Switzerland) **13**(1), 813-828 (2013).
8. Physionet: MIT-BIH Arhythmia Database. http://physionet.org/physiobank/database/mitdb/ (2014, November 10).
9. Thai, N.H., Nghia, N.T., Binh, D.V., Hai, N.T., Hung, N.M.: Long-tail Effect on ECG Classification. Proceedings of the International Conference on System Science and Engineering 2017 (ICSSE 2017), 34-38 (2017).

BÀI BÁO KHOA HỌC

THỰC HIỆN CÔNG BỐ THEO QUY CHẾ ĐÀO TẠO THẠC SỸ

Bài báo khoa học của học viên

có xác nhận và đề xuất cho đăng của Giảng viên hướng dẫn