Classification of Paper-based Electrocardiogram

Garn Wungkobkiat , Dusit Thanapatay
Department of Electrical Engineering
Kasetsart University
Bangkok, Thailand

Chusak Thanawattano
National Electronics and Computer Technology Center
National Science and Technology Development Agency
Pathumthani, Thailand

Akinori Nishihara
The Center for Research and Development of Educational Technology
Tokyo Institute of Technology
Yokohama, Japan

***Abstract*—A method for the automatic classification of paper-based electrocardiogram (ECG) is presented. An automated classification system of digital ECG has been developing for a few years. However, in reality, ECG signal usually recorded on the paper which cannot be directly analyzed by the computer. To extract the feature of signal, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Hybrid Discriminant Analysis (HDA) have been used to perform in this issue. ECG shape form scanned image was detected by many image processing techniques. Example data was obtained from MIT-BIH database. This investigation uses Support Vector Machine (SVM) to create the classifier model. This experiment resulted in an accuracy of 98.73%**

Keywords-Electrocardiogram (ECG); feature extraction; paper-based ECG; ECG classification

I. INTRODUCTION

Electrocardiogram (ECG signal) is a bio-electrical signal that reflects the performance and the properties of human heart. The ECG test uses the skin electrodes to measure the voltage fluctuations from patient heart and record on a chart recorder paper.

The pattern of ECG signal composed with three components as shown in figure 1. The left component is P wave which generated by the atrial depolarization, the center hump is QRS complex which represent the ventricles depolarizing and the right component is T wave which occurs with the re-polarization of ventricles.

Three major clinical contexts are diagnosis, therapy and monitoring. Automatic classification of Electrocardiogram has been an important role in the field of clinical diagnosis. Over the past few years, automated ECG signal classification is one of the content for a clinical monitoring [1]. The automated ECG classification is a monitoring system for patients who suffer from a life-threatening condition. The system should be able to detect the signs of cardiac disorder before it occurs and provide a warning system to save the life of the patient.

There are many researches that develop the classification system of ECG in digital format [2-5]. But in reality, especially in the rural area, the ECG usually stored in the paper-based which cannot be directly analyzed by the computer. Previous works for ECG printout digitization is referred in [6-9] but they did not working along with the classification system. So the investigation on ECG classification method for ECG printout is aimed to relieve this problem.

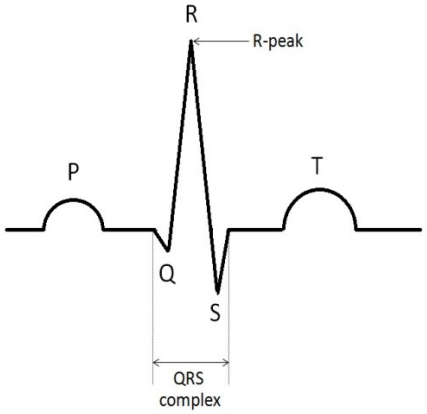

Figure 1. A normal pattern of ECG signal

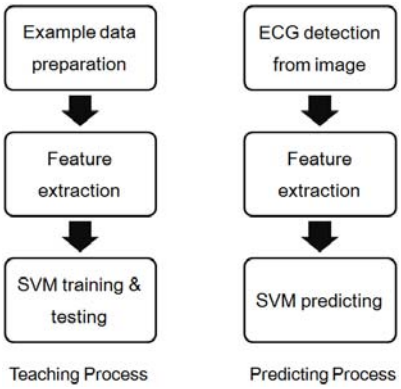

Figure 2. Flow chart of overall system

Proc. of the International Conference on Advances in Electronics and Electrical Engineering – AEEE 2012
Edited by Dr. R. K. Singh.

ISBN: 978-981-07-2683-6

II. METHOD AND INVESTIGATION

The classification system of ECG paper-base is separated into two parts. The first part is called "Teaching Process". This process is the method for teaching the system to learn how to cluster the shape of ECG signal. The second part is composed with many image processing techniques for detecting ECG signal from the image and predicts the class of ECG by the shape of signal which has been extracted from image. This part is named as "Predicting Process". Figure 2. is showing every steps of each processes.

III. TEACHING PROCESS

A. Example data preparation

The example of ECG signal is necessary for studying about development of the ECG classifying system. However, collecting an example from paper-based ECG is a complex and time consuming work. A digital form of ECG signal can be used in the classifying system because the important features of signal will be extracted or decoded by its shape. In this investigation, the example data were downloaded from MIT-BIH arrhythmia database [10]. MIT-BIH is a collaboration project between MIT and Beth Israel Deaconess Medical Centre to collect ECG recordings for approximately 110,000 cycles of heartbeat. They were all digitized at 360Hz. Each of signal cycle has its own annotation at the R-peak to describe the class of signal. Then the series of data was cropped by 315ms before and after R-peak into one beat (one beat has 231 samples). There are 20 different annotations (19 abnormal classes and one normal class) that appear in annotation files. 11 categories of annotations that appear over than 100 beats were selected to use in the investigation. The eleven types which we choose are normal beat (N), left bundle branch block beat (L), right bundle branch block beat (R), atrial premature beat (A), aberrated atrial premature beat (a), premature ventricular contraction (V), fusion of ventricular and normal beat (F), ventricular flutter wave (!), ventricular escape beat (E), Paced beat (/) and fusion of paced and normal beat (f). The other classes are to less to be classified.

The example data and the data which will be predicted should have the same frequency. So, all of example beat have been re-sampling to 200Hz (128 Samples). Then Fast Fourier Transform (FFT) is used for noise reduction. Most of ECG signal has a frequency between 0.5-40 Hz. So the information in a frequency domain below or over 0.5-40Hz will be removed out. Finally, normalize and zero-mean each beats to ensure that the data at each scale are comparable to each other.

B. Feature extraction

In pattern recognition, the input data usually have much insignificant data and too large to be analyzed. Feature extraction has the main purpose is transforming input data into a reduced representation with sufficient information to be processed.

1) PCA: Principal Component Analysis is the most popular and easiest way to performs feature extraction. The main idea of PCA is projecting the data onto the plain that input data distribute to.

Setup n number of input data $X = \{x_1, x_2, \ldots, x_n\}$, x is positioned in column vector, and use PCA to find the direction which information is most distributed to. W was defined to be the vector of that direction. To find W, let W be a unit vector ($\|W\| = 1 = W^TW$) then project X on W like dot product. The outcome are x_1^TW, x_2^TW,… x_n^TW. The bigger variance the higher data distribution. So X was assumed that its average zero for finding the variance.

$$Var(W,X) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T W)^2 \quad (1)$$

$$Var(W,X) = \frac{1}{n}\sum_{i=1}^{n}(W^T x_i)(x_i^T W)$$

$$Var(W,X) = W^T\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) W$$

$$Var(W,X) = W^T C W \quad ; C = \text{covariance matrix}$$

Equation(1) is showing a solution to find Var(W,X). The maximum value of W^TCW can be obtained when W is a unit vectors. The solution is CV = VD. V is matrix whose columns are the corresponding Eigenvectors and D is a diagonal matrix of Eigen-values. Eigen-values are variance of W direction. So principal components can be found by the Eigen-vector with high Eigen-value. If we want to reduce the dimension of data to m, the m Eigen-vector with the highest Eigen-value will be selected to be a transformation matrix. This matrix is used to multiply with input data and will be used with the input data in Predicting part later. In this investigation, size of reduced data was chosen to be 10 as same as LDA for the reason mentioned in the following section.

2) LDA: Linear Discriminant Analysis can be used for both feature extraction and pattern recognition. Compared to PCA, LDA gives the most discriminant features while PCA creates the most descriptive features. LDA main concept is finding subspace which projected the data with the same class closer than the other class. Instead of finding the variance of X, we need to find the between-class covariance and within-class covariance.

$$S_b = \sum_k n_k (\mu_k - \mu)(\mu_k - \mu)^T \quad (2)$$

$$S_w = \sum_k \sum_{i \in k}^{n_k} (x_i - \mu_k)(x_i - \mu_k)^T \quad (3)$$

S_b is between-class covariance (2) and S_w is within-class covariance (3). n_k is amount of k class data, μ_k is average vector of k class and μ is average vector of whole data To separate the data into group $|W^TS_bW|$,between-class variance, must be maximized and $|W^TS_wW|$,within class variance, must be minimized. In conclusion, we need to find the Eigen-vectors and Eigen-values of $S_w^{-1}S_b$. Maximum number of dimensions that performed by LDA is equal to number of class minus by one (equal to 10 in this investigation) because the other dimensions has very small Eigen-value and inconsiderable.

Financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology and Kasetsart University (KU).

3) HDA: Hybrid Discriminant Analysis [11] is a combining technique between PCA and LDA. The maximum variance can be found by

$$W_{opt} = \arg\max_W \frac{|W^T[(1-\lambda)S_b+\lambda C]W|}{|W^T[(1-\eta)S_w+\eta I]W|} \quad (4)$$

(λ, η) are two parameters in the range of (0, 0) to (1, 1).

- (λ=0, η=0) the transformation reduces to full LDA.
- (λ=1, η=1) the transformation recovers to full PCA.
- (λ=0, η=1) gives a subspace that is mainly defined by maximizing the scatters among all the classes with minimal effort on clustering each class.
- (λ=1, η=0) gives a subspace that mainly preserves the most energy while minimizing the scatter matrices of within-classes.
- (λ=0.5, η=0.5) gives a subspace that is a trade-off between LDA and PCA.

The best dimensional reduction can be found from this method by the summation of selected Eigen-values divided by summation of all Eigen-values. In this investigation (λ=0, η=1) yields the most complete dimensional reduction when the output dimension is 10.

C. SVM training & testing

Support Vector Machine (SVM) is supervised machine learning for classification. Example signal was divided into 2 equal halves. The first half was used for training SVM to create the boundaries to separate training data with different class. The second half was used to test the performance of the classifying system. LIBSVM [12] is an open sources library for support vector classification. It performs an automate training and testing. Classifier model from LIBSVM yields the best accuracy as possible.

In TABLE I., although LDA is an algorithm for classification, it becomes the worst method in this investigation. When the training data set is small, PCA can outperform LDA. PCA is less sensitive to different training data sets [13]. In this investigation, some class has a small of number if compared with N class, L class or / class. So, it is normally if PCA is better than LDA in this situation. HDA contains compressed important information more than PCA. However, it still cannot overcome PCA. The completeness of data compression cannot indicate the performance of classification. From teaching process, PCA was selected to be a method for feature extraction in overall classification system.

TABLE I. RESULT FROM VARIOUS METHOD

ECG class	Number of testing data	Number of correctly classified		
		PCA	*LDA*	*HDA*
N (normal beat)	26036	25928	25833	25934
L	3242	3235	3153	3233
R	2794	2769	2722	2764
A	301	203	145	195
a	58	40	34	40
V	1518	1440	1337	1418
F	371	281	225	276
!	50	30	29	33
E	50	48	48	49
/	3470	3454	3454	3458
f	460	435	380	426
All of disorder class	12314	12032	11677	12000
Accuracy (%)	-	98.7301	97.4185	98.6336

IV. PREDICTING PROCESS

A. Beat detection from image

1) Image binarization: The image is scanned with 300 dpi which is equivalent to 295 Hz. This process converted an image into black and white colors or binary image by selecting a suitable threshold limit. From the gray-scale scanning image, it should have color levels varied from 0 to 255 in any pixels (absolutely black is 0 and absolutely white is 255). After selecting the threshold limit (130 in this experiment) the pixels with the color level over than threshold limit is replaced with white pixels and black pixels is placed on the others. This approach aims to remove unnecessary object like background, stain or red line grid which usually appear in the original chart paper as shown in Figure 3.

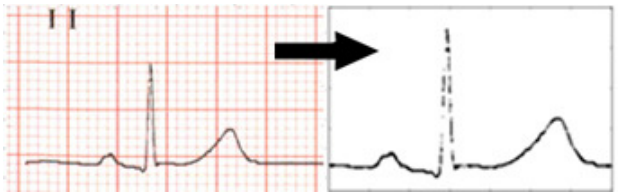

Figure 3. Image binarization

2) Removing black grid: Some of image are scanned from a copy of ECG. All of these image have a black grid which cannot be removed totally after the binarization. Median filter has been used to removing thin continuous lines in this situation. The main concept of median filter is sliding the pattern of neighbors, called "window", pixel by pixel and replacing each pixel with the median value of neighboring pixels over the entire image.

3) Thinning image: The purpose of thinning process is creating the 1-pixel thickness or also called "skeleton image"

of the ECG signal for eliminating unnecessary the repetition of data. The moving average algorithm is chosen to operate in this process. Similarly to median filter, moving average algorithm has a window to slide pixel by pixel. Instead of finding the median value, the average of neighbor pixel in the same column is what this algorithm focuses on (size of window was chosen to be 5 pixels in this experiment). The only one pixel in the column that still be black pixel is the pixel with the largest average value and other pixel will be replaced with white pixel. The result was shown in Figure 4.

4) Time series data extraction: Time series data extraction idea is searching black pixel of each column. A row index of black pixel(start form bottom) will be recorded as the 1-dimension data. The process was repeated on every column from left to right as shown in Figure 5.

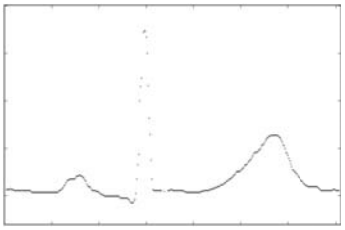

Figure 4. Skeleton image

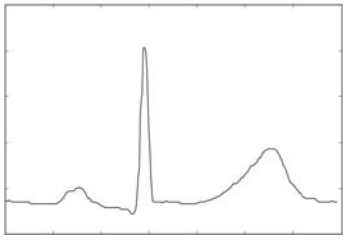

Figure 5. Extracted data from scanned image

B. Feature extraction

Extracted data is digitized at 295Hz. (189 samples per beat) but example data from teaching process was resample to 200Hz. So the first step is re-sampling to 200Hz. (128 samples per beat). Then following the teaching process, the extracted data must be passed to FFT for eliminating noise. After that, normalize and zero-mean the data. Finally, feature extraction must perform in this step before predicting class by multiply the extracted data with transformation matrix which obtained from PCA process in teaching process.

C. SVM predicting

Finally, the extracted data from image is ready to be predicted the class by LIBSVM. This method also supports digital form classification because the feature of ECG signal is based on shape of signal.

V. CONCLUSION

This paper has presented a method for developing a classification of ECG printout system. The system has been separated into two parts. In teaching process, the example data were downloaded from MIT-BIH and have been extracted the feature by PCA. Classifier model was created from example data by SVM. In predicting process, ECG signal was retrieved from scanned image. The feature was extracted by using transformation matrix in teaching process. In the end, SVM was used to predict the class of ECG data.

The system in this research must be control manually. The automated system will be developing in future work.

ACKNOWLEDGMENT

This research is financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology and Kasetsart University (KU).

REFERENCES

[1] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," Biomedical Engineering, IEEE Transactions, pp. 1196 - 1206, July 2004.

[2] H. Zhang, and L. Q. Zhang, "ECG analysis based on PCA and Support Vector Machines," Neural Networks and Brain, 2005. ICNN&B '05. International Conference, pp. 743 - 747, 13-15 Oct. 2005.

[3] M. S. Khadtare, and J.S. Sahambi, "ECG Arrhythmia Analysis by Multicategory Support Vector Machine," Computer Science, Volume 3285/2004, pp. 100-107, 2004.

[4] R. Besrour, Z. Lachiri, and N. Ellouze, "ECG Beat Classifier Using Support Vector Machine," Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference, pp. 1-5, 7-11 April 2008.

[5] S. Osowski, L. T. Hoai, and T. Markiewicz, "Support Vector Machine-Based Expert System for Reliable Heartbeat Recognition," Biomedical Engineering, IEEE Transactions on, pp. 582-589, July 2004.

[6] S. Mitra, and M. Mitra, "An automated data extraction system from 12 lead ECG images," Comput Methods Programs Biomedicine, pp. 33 – 38, May 2003.

[7] S. Mitra, M. Mitrab, and B.B. Chaudhuri, "Generation of digital time database from paper ECG records and Fourier transform-based analysis for disease identication," Computers in Biology and Medicine, pp. 551 - 560, October 2004.

[8] F. Badilini, T. Erdem, W. Zareba, and A. J. Moss, "ECGScan: a method for conversion of paper electrocardiographic printouts to digital electrocardiographic files," Journal of Electrocardiology, pp. 310 - 318, 8 October 2005.

[9] D. Thanapatay, C. Suwansaroj, and C. Thanawattano, "ECG beat classification method for ECG printout with Principle Components Analysis and Support Vector Machines," Electronics and Information Engineering (ICEIE), 2010 International Conference, pp. V1-72 - V1-75, 1-3 August 2010.

[10] R. Mark and G. Moody, "MIT-BIH arrhythmia database directory," Massachusetts Inst. Technol. (MIT).

[11] Q. Tian, J. Yu, and T. S. Huang, "Boosting Multiple Classifiers Constructed by Hybrid Discriminant Analysis Lecture Notes in Computer Science, Volume 3541/2005, pp. 653 - 657, 2005.

[12] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines", 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] A. M. Marto Anez, and A. C. Kak, "PCA versus LDA," Pattern Analysis and Machine Intelligence, IEEE Transactions, pp. 228 - 233, February 2001.