ECG Feature Extraction and Classification Using Wavelet Transform and Support Vector Machines

Qibin Zhao
Department of Computer Science
and Engineering
Shanghai Jiaotong University
Shanghai, 200030, China
E-mail: qbzhao@sjtu.edu.cn

Liqing Zhang
Department of Computer Science
and Engineering
Shanghai Jiaotong University
Shanghai, 200030, China
E-mail: zhang-lq@cs.sjtu.edu.cn

Abstract— **This paper presents a new approach to the feature extraction for reliable heart rhythm recognition. This system of classification is comprised of three components including data preprocessing, feature extraction and classification of ECG signals. Two different feature extraction methods are applied together to obtain the feature vector of ECG data. The wavelet transform is used to extract the coefficients of the transform as the features of each ECG segment. Simultaneously, autoregressive modelling(AR) is also applied to obtain the temporal structures of ECG waveforms. Then the support vector machine(SVM) with Gaussian kernel is used to classify different ECG heart rhythm. Computer simulations are provided to verify the performance of the proposed method. From computer simulations, the overall accuracy of classification for recognition of 6 heart rhythm types reaches 99.68%.**

I. Introduction

The electrocardiogram (ECG) is routinely used in clinical practice, which describes the electrical activity of the heart. In physical checkups at hospitals, physicians record the ECG after the patient has exercised to check his/her cardiac condition. The Holter ECG device is used most frequently for recording the ECG. Physicians apply the device to a patient when they need to monitor his/her ECG to find the few abnormal cycles in the ECG throughout the day. Physicians then interpret the shapes of those waves and complexes. They calculate parameters to determine whether the ECG shows signs of cardiac disease or not. The parameters are the height and the interval of each wave, such as RR interval, PP interval, QT interval, and ST segment. Recognition of the fiducial points and calculations of the parameters is a tedious routine for the physician. Therefore, there is an urgent need for an automatic ECG recognition system to reduce the burden of interpreting the ECG.

Various studies have been done for classification of various cardiac arrhythmias [1][2][3][4]. In this paper, we propose the combination of wavelet transform and AR model as the feature extraction method, then use the SVM to classify the ECG heartbeat. The proposed approach is validated in the MIT-BIH Arrhythmia Database[5] and get high accuracy of classification.

II. ECG data and preprocess

All ECG data were obtained from MIT-BIH arrhythmia database that contains records of many patients with heart troubles or abnormalities. The frequency of the ECG data was 360HZ. Each record has its respective annotation file that indicate the class of the heartbeat. A single channel ECG is collected and used to algorithm evaluation. Since there are few categories of abnormal QRS complexes in one record, we select different abnormal QRS complexes from several records. Six types of QRS complexes appeared frequently in the database. Therefore, we mainly deal with six types heartbeats which include normal beat(NORMAL), left bundle branch block beat(LBBB), right bundle branch block beat(RBBB), paced beat(PACE), premature ventricular contraction(PVC) and atrial premature contraction(APC).

In the data preprocessing process, continuous ECG signals must be separated into many segments which contain one heartbeat. The extracted data of ECG complexes is centered around R peak. Considered that some PVC duration is great and sometimes R peak detection may be not the center of the complex, we have selected segment of 250ms before the fiducial point and 400ms after that with the R peak point is the 90th point. The R peak is detected using the Pan and Tompkins algorithm[6]. Thus, each segment must contain one ECG heartbeat. Fig.1 shows typical waveforms of six types of ECG segments.

III. Feature extraction

The recognition of heart rhythms requires generation of the feature vector which represents the original ECG segment. A good recognition system should depend on the features representing the ECG signals in such a way, that the differences among the ECG waveforms are suppressed for the waveforms of the same type but are emphasized for the waveforms of belonging to different types of heartbeats. We perform the recognition process of heart rhythms on the single heartbeat of the ECG, proposing the description or representation by wavelet transform and AR model.

0-7803-9422-4/05/$20.00 ©2005 IEEE

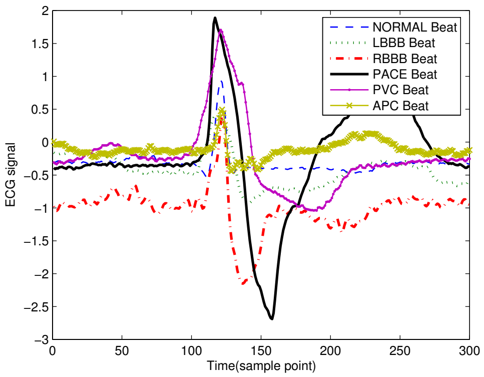


Fig. 1. Typical ECG heartbeat waveforms of six types

A. Wavelet Transform

The WT of a signal $f(x)$ is defined as

$$W_s f(x) = f(x) * \Psi_s(x) = \frac{1}{s}\int_{-\infty}^{+\infty} f(t)\Psi(\frac{x-t}{s})dt \quad (1)$$

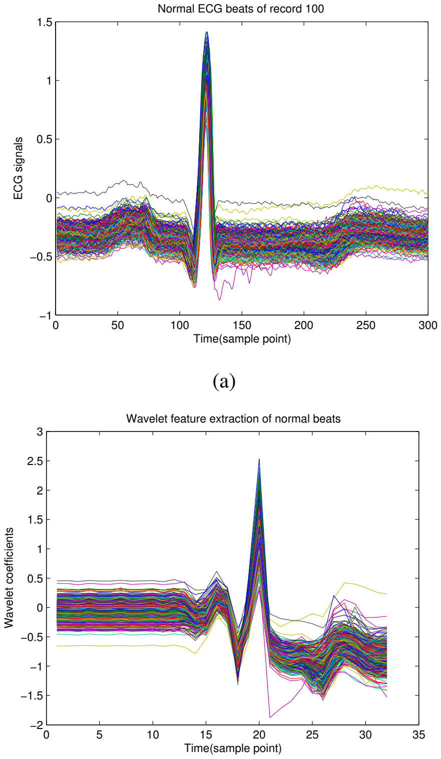
where s is scale factor. $\Psi_s(x) = \frac{1}{s}\Psi(\frac{x}{s})$ is the dilation of a basic wavelet $\Psi(x)$ by the scale factor s. Let $s = 2^j (j \in Z, Z$ is the integral set), then the WT is called dyadic WT[7]. The dyadic WT of a digital signal $f(n)$ can be calculated with Mallat algorithm as follows:

$$S_{2^j} f(n) = \sum_{k \in Z} h_k S_{2^{j-1}} f(n - 2^{j-1}k) \quad (2)$$

$$W_{2^j} f(n) = \sum_{k \in Z} g_k S_{2^{j-1}} f(n - 2^{j-1}k) \quad (3)$$

where S_{2^j} is a smoothing operator. $S_{2^j} f(n) = a_j, a_j$ is low frequency coefficients that is the approximation of original signals while $w_{2^j} f(n) = d_j, d_j$ is high frequency coefficients that is the detail of original signals. It is known that the WT is better suited to analyzing nonstationary signals, the discrete wavelet transform performs an adaptive time-frequency decomposition of a presented pattern. By the multiresolution representation it is possible to describe the signal structure by only a few coefficients in the wavelet domain.

The selection of appropriate wavelet and the number of decomposition level is very important in analysis of signals using the WT. The number of decomposition levels is chosen based on the dominant frequency components of the signal. The level are chosen such that those parts of the signal that correlate well with the frequencies required for classification of the signal are retained in the wavelet coefficients. The wavelet used in this work is one member of the Daubechies families. The number of decomposition levels was chosen to be 4. Thus, the ECG signals were decomposed into the details $d_1 - d_4$ and one approximation a_4. Usually, tests are performed with different types of wavelets and the one which gives maximum efficiency is selected for the particular application. Therefore, the Daubechies wavelet of order 8 was chosen through some tests.



(a)

(b)

Fig. 2. (a) Original normal heartbeats (b) Wavelet coefficients of normal heartbeats

Feature selection is also very important component, it has two meanings: which components of a pattern or which set of inputs best represent a given pattern. The computed discrete wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. For each heartbeat, the detail wavelet coefficient d_1 are usually noise signals that must be eliminated and d_2, d_3, d_4 represent high frequency parts of the ECG signal. Since the approximation wavelet coefficients a_4 represent the main feature of each heartbeat, then we chosen a_4 to be feature of each heartbeat. For one heartbeat, the original signal contains 300 points, and a_4 contains 32 points. Fig.2(a) shows some normal heartbeats in record 100, and Fig.2(b) shows the wavelet feature extraction of these normal heartbeats.

B. Autoregressive Modelling

In an AR process of order p, the signal $x[n]$ at time instant n may be represented as a linear combination of p previous values of the same signal. Specifically, the process is modelled as

$$x[n] = \sum_{i=1}^{p} a[i]x[n-i] + e[n] \quad (4)$$

where $a[i]$ is the ith coefficient of the AR model, $e[n]$ is a white noise with mean zero, and p is the AR order. Various methods are currently used to estimate the coefficients of an autoregressive process. The criterion used to evaluate the

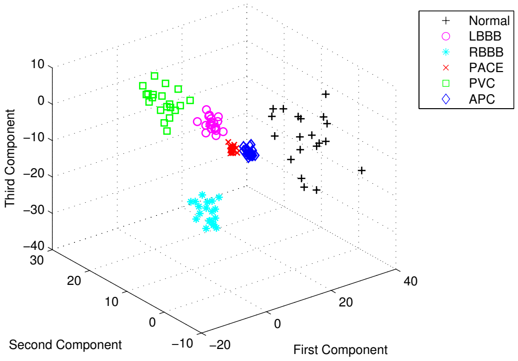


Fig. 3. Data Distribution in Feature Space

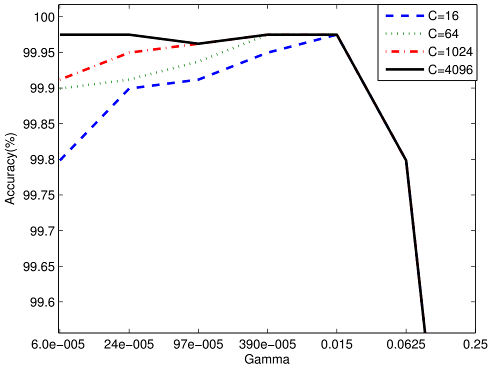


Fig. 4. Crossing-validation result by different parameters

model order selection in this work was ARfit method[8]. Eventually, we select 4-order AR model to represent each heartbeat of ECG signals.

C. Generation of Feature Vectors

The feature vector was composed of wavelet coefficients and AR coefficients. For each ECG heartbeat, the original signal is a vector of 300 dimensions $\{y_1, y_2, ..., y_{300}\}$. The wavelet coefficient for it is a vector of 32 dimensions $\{w_1, w_2, ..., w_{32}\}$, and the AR coefficient for it is a vector of 4 dimensions $\{a_1, a_2, a_3, a_4\}$. Then wavelet coefficient and AR coefficient can be concatenated together to form the feature vector $\{w_1, w_2, ..., w_{32}, a_1, a_2, a_3, a_4\}$ for classification.

Every feature vector represents one point in the feature space. Those points of same class should be closer and points of different classes should be far from each other. In order to view the data distribution of all types points of the ECG heartbeat in figure, PCA was used to reduce the dimension of each heartbeat's feature vector to three dimensions. Fig.3 shows the data distribution of six classes ECG heartbeats in three dimensions feature space. From the figure, we can see most points of different types heartbeats have be separated, which proved that the process of feature extraction got the good effect.

IV. ECG Classification

Support Vector Machine(SVM) is one of the pattern recognition methods, and is proposed by V.Vapnik and his coworkers[9][10]. SVM separates an input example $X = (x_1, ...x_d)$ of dimension d into two classes. A decision function of SVM separates two classes by $f(X) > 0$ *or* $f(X) < 0$. The size of training set N is $(y_i, X_i), i = 1, ..., N$. Where $X_i \in R^n$ is the input pattern for the ith example, and $y_i \in -1, 1$ is the class label. Support Vector classifiers implicitly map X_i form input space to a higher dimensional feature space which depend on a nonlinear function $\phi(X)$. A separating hyperplane is optimized by maximization of the margin. Then SVM is solved as the following quadratic programming problem,

$$Maximize : \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j), \quad (5)$$

$$Subjectto : 0 \leq \alpha_i \leq C(i = 1, ...n), \sum_{1}^{n} \alpha_i y_i = 0 \quad (6)$$

Where $\alpha \geq 0$ are Lagrange multipliers. When the optimization problem has solved , many α_i will be equal to 0, and the others will be Support Vectors. C is positive constant which chosen empirically by the user. This parameter expresses degree of loosing constraint. A large C can classify training examples more correctly. $K(X, X^{'})$ is the kernel function which is inner-product defined by $K(X, X^{'}) = \phi(X) \cdot \phi(X^{'})$. Then the SVM decision function is

$$f(X) = \sum_{X_i \in SV} \alpha_i y_i K(X_i, X) + b. \quad (7)$$

A common kernel is the Gaussian radial basis function (RBF),

$$K(x, x^{'}) = e^{-\|X - X^{'}\|^2/2\sigma^2} \quad (8)$$

In this work, the Gaussian kernel is used as the kernel function.

Classification consists of two steps: learning and testing. Our classifier is a learning machine of the supervised type and Multicategory SVM(MC-SVM). Firstly, all ECG segments that contain a special type heartbeat are mapped into feature space using wavelet and AR model which have been explained above. In the learning phase, SVM receives some patterns as input. These patterns are heartbeats represented by m feature parameters that can be seen as points in m-dimensional space. Then the machine becomes able to find the labels of new vectors by comparing them with those used in the learning phase.

V. Results & Discussion

In the numerical experiments, we have used the ECG data from the MIT-BIH Arrhythmia Database corresponding to the normal heartbeat and 5 types of arrhythmias. Each type heartbeat was extracted from the record which contained most beats of this type. In this work, the NORMAL beats, LBBB, RBBB, PACE, PVC and APC were extracted respectively from the record 100, 109, 118, 107, 208 and 232. Then each type heartbeat has a data set that contains many heartbeats. Due to the scarcity of data corresponding to some beat types the

TABLE I
THE RESULT OF CLASSIFICATION

Beat type	N	L	R	P	V	A	TOTAL
Training number	1566	1743	1516	1454	694	967	7940
Testing number	671	747	649	623	298	415	3403
Correctly beats	671	737	649	623	297	415	3392
Misclassified to this beat	0	0	7	1	0	3	11
Accuracy%	100	98.66	100	100	99.66	100	99.68

number of data belonging to each heartbeat type was variable. We chosen 70% of data set to be training data and 30% of data set to be testing data. The total number of data used in training was equal 7940. Another 3403 data points have been left for testing.

In order to get a best optimal SVM classification, we should adjust two parameters $(C, \gamma), \gamma = \frac{1}{2\sigma^2}$ which is very important for classification effect. Therefore, we must first find the best parameter C and γ using crossing-validation. Then use the best parameter C and γ for training and testing. Form Fig.4 we can see the relation of crossing-validation accuracy and two parameters. Then the best parameters can be founded according to the Accuracy. In this work, we set $C = 65536$, and $\gamma = 2.44e - 004$.

Table I shows the result of SVM classification for all classes of heartbeats. Row 1 lists the six typical heartbeats which was considered in our experiment. N represents the normal beat, L represents the LBBB, R represents the RBBB, P represents the PACE beat, V represents the PVC beat and A represents the APC beat. Training number represents the number of learning beats for all classes of heartbeats, testing number represents the number of test beats for all classes of heartbeats and misclassified to this beat represent the number of classified to the beat which in fact is not belong to this beat. It is evident that the best results have been obtained for SVM, and the NORMAL heartbeats can be entirely separated form other abnormal heartbeats with 100% accuracy.

The objective of this study is to model single-lead ECG signals for extracting classifiable features in order to improve the classification results using wavelet and AR modelling. The wavelet transform analysis provides robust features in presence of background continuous noise. Dominant and important features in the ECG data are extracted to provide robust information retrieval for classification. AR coefficients are also be used to extract the feature of ECG data. Because of the reduced dimensions of feature vectors, the classification can be done quickly. The experiment results show that using wavelet and AR model together can result in high accuracy of classification.

VI. CONCLUSION

The experiments of recognition of 5 types of arrhythmias and normal beat were carried out on MIT-BIH Arrhythmia Database. The SVM used for classification of the ECG beat was trained, cross validated and tested with the extracted features from discrete wavelet transform and AR model of the ECG signals. Computer simulations showed that our approach gives the excellent performances of successful recognition, accuracy is found to be 99.68%. Because of single-lead ECG was used in this study, Other further work can try to increase the number of ECG leads for more accuracy.

ACKNOWLEDGMENT

The project was supported by the national natural science foundation of China under grant 60375015.

REFERENCES

[1] S chen , "A two stage discrimination of cardiac arrhythmias using a total least squares-based prony modeling algorithm," IEEE Trans on BME, vol.47,No.10,pp 1317-1327,October 2000

[2] İnan Güler, ElifDerya Übeyli, "ECG beat classifier designed by combined neural network model", Rattern Recognition 38(2005)199-298

[3] Yasushi Kikawa and Koji Oguri, "A Study for Excluding Incorrect Detections of Holter ECG Data Using SVM", N.R.Pal et al.(Eds.):ICONIP 2004, Lncs 3316,pp.1223-1228,2004.

[4] Stanislaw Osowski,Linh Tran Hoai,and Tomasz Markiewicz,"Support Vector Machine-Based Expert System for Reliable Heartbeat Recognition", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING,VOL.51,NO.4,APRIL 2004

[5] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220

[6] J.Pan and W.J.Tompkins, "A real-time QRS detection algorithm"IEEE Trans.Biomed.Eng., Vol.BME-32, PP.230-236,1985

[7] Cuiwei Li,Chongxun Zheng, and Changfeng Tai, "Detection of ECG Characteristic Points Using Wavelet Transforms", IEEE Transactions on biomedical engineering, VOL.42,NO.1,pp.21-28,January 1995

[8] A. Neumaier and T. Schneider, 2001: "Estimation of parameters and eigenmodes of multivariate autoregressive models". ACM Trans. Math. Softw., 27, 27C57.

[9] Vladimir N. Vapnik:"The Nature of Statistical Learning Theory 2nd edn" Springer Verlag,(1999)

[10] Nello Cristianini and John Shawe-Taylor:"An Introduction to Support Vector Machines" Cambridge University Press,(2000)