Study of Feature Extraction Based on Autoregressive Modeling in ECG Automatic Diagnosis

GE Ding-Fei[1] HOU Bei-Ping[1] XIANG Xin-Jian[1]

Abstract This article explores the ability of multivariate autoregressive model (MAR) and scalar AR model to extract the features from two-lead electrocardiogram signals in order to classify certain cardiac arrhythmias. The classification performance of four different ECG feature sets based on the model coefficients are shown. The data in the analysis including normal sinus rhythm, atria premature contraction, premature ventricular contraction, ventricular tachycardia, ventricular fibrillation and superventricular tachycardia is obtained from the MIT-BIH database. The classification is performed using a quadratic discriminant function. The results show the MAR coefficients produce the best results among the four ECG representations and the MAR modeling is a useful classification and diagnosis tool.

Key words Autoregressive model, ECG features, classification, automatic diagnosis.

1 Introduction

One of the most important tasks is the reliable detection and classification of the arrhythmias for automatic monitoring and diagnosis. Among those threatening arrhythmias, ventricular tachycardia (VT) and ventricular fibrillation (VF) are most dangerous because they produce the haemodynamic deterioration. Other arrhythmias like premature ventricular contraction (PVC) *etc.* are not so lethal, but are also important for diagnosing the heart diseases. Various studies have been proposed for classification of various cardiac arrhythmias, such as analysis of peaks in the short-term autocorrelation function[1], time-frequency analysis[2], nonlinear dynamical modeling method[3,4], total least squares based Prony modeling algorithm[5], correction waveform analysis[6], and artificial neural networks for decimated ECG analysis[7]. Generally, these techniques classify only two or three arrhythmias, therefore there is a need for extending the identification technique for a larger number of arrhythmias and easy real-time implementation.

Multivariate autoregressive (MAR) modeling provides an approach to analyse the bio-signals. For example, MAR modeling was widely used to model heart rate (HR), blood pressure (BP) and respiration (RESP) for assessment of interaction between them[8]. MAR modeling was used to extract the features from the human electroencephalogram with which mental tasks can be discriminated[9]. However, in the study of ECG arrhythmia recognition problems, researches have not done too much using MAR model and multiple lead ECGs. Scalar autoregressive (AR) modeling has been widely utilized to model bio-signals for the purpose of analysis, such as AR modeling of scalar time signals based on Kalman filter for calculating instantaneous measures of linear dependence[10], AR modeling used to model heart rate variability (HRV) and for power spectrum estimation of ECG and HRV signals[11], AR coefficients used as ECG features for classification of cardiac arrhythmias using fuzzy ARTMAP[12]. It is noted that normal ECG QRS complexes are usually prominent in ECG lead II and normal beats are frequently difficult to discern in ECG lead VI although ectopic beats will often be more prominent. Thus, two-lead ECG signals contain more information than one-lead ECG signals, and the classification results can be improved by using two-lead ECG signals significantly.

The purpose of the present work is to explore the feasibility of MAR and AR modeling to extract the classification features from two-lead ECG signals in order to classify more types of cardiac arrhythmias with higher accuracy. In this study, MAR and AR modeling were performed on the ECG data including normal sinus rhythm (NSR), atria premature contraction (APC), PVC, VT, VF and superventricular tachycardia (SVT). There were four ECG representations based on the model coefficients, and the classification was performed using quadratic discriminant function (QDF) based classifier. Three hundred sample patterns each from the six classes were selected for analysis. A training data set consisted of 150 sample patterns each from the six classes, and the remaining data was used for testing. The results showed that the MAR coefficients could classify better than other three representations. Thus, MAR modeling is a useful classification and diagnosis tool for the cardiac arrhythmias.

2 Methods

2.1 Preprocessing

The data in the analysis was obtained from the MIT-BIH database. The NSR, PVC and APC were sampled at 360Hz, the VT and VF were sampled at 250Hz, and the SVT was sampled at 128Hz. The data including NSR, PVC, APC and SVT was subsampled in order that all the two-lead ECG signals in the analysis had a frequency of 250Hz. All ECG data have been filtered to remove the noise including respiration, base line drift and wandering *etc.* The high-pass filter is of a linear phase characteristic based on the frequency of 250Hz. The cut off frequency of the high-pass filter is 2Hz. Thus, the drift caused by respiration at about 0.2Hz is sufficiently removed. The other noise caused by the motion from the electrode is also minimized.

The R peaks of the ECGs were detected using Tompkin′s algorithm[13]. A normal ECG refers to the usual case in the health adults where the heart rate is 60∼100 beats per minute. In the current study, the sample size of the various segments was 0.9 seconds. 0.3 seconds before R peak and 0.6 seconds after R peak were picked for modeling. It is adequate to capture most of the information from a particular cardiac cycle.

Received January 16, 2006; in revised form May 24, 2006
Supported by Natural Science Foundation of Zhejiang Province of P. R. China (Y104284)
1. School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310012, P. R. China
DOI: 10.1360/aas-007-0462

2.2 MAR and scalar AR modeling

A common form of a MAR model of order P is given by[8,9].

$$\boldsymbol{X}(k) = -\sum_{i=1}^{P} A(i)X(k-i) + \boldsymbol{e}(k) \tag{1}$$

where $\boldsymbol{X}(k)$ is a 2-dimensional column vector of observations at time k, $\boldsymbol{e}(k)$ is a 2-dimensional column vector of unknown, zero-mean, uncorrelated random variable, $A(i)$, for $i = 1, 2, \ldots, P$ is the 2×2 matrices of MAR model coefficients to be estimated.

It is important to determine the model order which best fits the data when constructing a MAR model. The model was estimated from 225 points of data (0.9seconds) from two ECG leads in this research. The model order selection was performed on the six types of two-lead ECG signals including in the analysis. Pre-selected model orders from one to eight were investigated for model order selection. Burg′s algorithm was used to estimate the MAR coefficients. The criterion used to evaluate the model order selection was the sum-squared error (SSE) in this work[10].

Scalar AR modeling was performed on each of the two ECG leads for the six types of ECG signals. The AR model order was estimated based on the SSE, and was calculated over all estimates in the 225-point window segmented from single lead.

2.3 ECG features

In this study, four different representations of ECG signals were used for classification: the MAR coefficients, the K-L MAR coefficients, the scalar AR coefficients based on two-lead ECG, and the scalar AR coefficients based on single-lead ECG.

2.3.1 ECG features based on MAR and K-L MAR coefficients

A MAR process of order P has been applied to the two-lead ECG signals from the six classes. The number of MAR coefficients representing a two-lead ECG segment was $4P$.

In order to reduce the redundancy of features, K-L MAR coefficients was computed and used as features. The K-L transform can reduce the dimension of feature space by projecting the original feature vectors onto a small number of eigenvectors. The K-L transform in this study was performed as follows[14]:

1) Calculate the within-class scatter matrix. 2) Calculate the eigenvalues and eigenvectors of the within-class scatter matrix. 3) The set of m eigenvectors which correspond to the m largest eigenvalues was chosen to transfer the original data, the corresponding eigenvectors in this study was determined by the index i for which $r_i/r_{max} \leq 0.001$, where $i = 1, 2, \ldots, 4P$, r_i′s are in the descending order. 4) Generate the K-L transform by projecting each $4P$-dimentional pattern onto these chosen eigenvectors. Thus the dimension of the features based on K-L MAR coefficients was m.

2.3.2 ECG features based on scalar AR coefficients

A scalar AR process of order P has been performed on each ECG lead from the six classes. The scalar AR coefficients were estimated from each lead and concatenated together to form the feature vectors for the classification. The number of the scalar AR coefficients representing a two-lead ECG segment was $2P$, the number of the scalar AR coefficients representing a single-lead ECG segment was P.

2.4 QDF-based classification

The ECG features described as above were utilized to classify the cardiac arrhythmias. The various cardiac arrhythmias have been classified by a stage-by-stage QDF-Based algorithm in current research. The QDF is given by[14]

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i \tag{2}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]$ represents a d-dimensional ECG feature vector, y_i is an observed response, ε_i is the QDF error, $\boldsymbol{\beta}$ is a $(d(d+3)/2+1)$-dimensional column vector. $\boldsymbol{X}_i$ is a $(d(d+3)/2+1)$-dimensional row vector, that is

$$\boldsymbol{X}_i = [1, x_1, x_2, \ldots, x_d, x_1^2, x_2^2, \ldots, x_d^2, 2x_1x_2, 2x_1x_3, \ldots, 2x_1x_d, 2x_2x_3, 2x_2x_4, \ldots, 2x_2x_d, \ldots, 2x_{d-1}x_d]$$

The ECG feature vector of a particular ECG segment was mapped to a response (1 or –1). Assume the total number of the ECG segments used for classification at a particular stage is D. The following equation can be given

$$\tilde{\boldsymbol{Y}} = A\boldsymbol{\beta} + \boldsymbol{E} \tag{3}$$

where $\tilde{\boldsymbol{Y}} = [y_1, y_2, \ldots, y_D]^{\mathrm{T}}$ is a D-dimensional column vector of the observed responses, and made up of "1" and "-1", which correspond to different classes respectively, $A = [X_1, X_2, \ldots, X_D]^{\mathrm{T}}$ is a $D \times (d(d+3)1/2+1)$ matrix, $\boldsymbol{E} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_D]^{\mathrm{T}}$ is a D-dimensional column vector of the errors.

The least squares estimator is

$$\boldsymbol{\beta} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\tilde{\boldsymbol{Y}} \tag{4}$$

The quadratic discriminant function of the classifier is

$$Y_I = \boldsymbol{X}_i\boldsymbol{\beta} \tag{5}$$

Table 1 shows the classification algorithm for the MAR and K-L MAR coefficients. The similar classification algorithm can be constructed for the scalar AR coefficients. The criterion based on standard deviation and Euclidean center distance (SDECD) was used to measure the separability between two classes. Associated value of SDECD was computed to determine the groupings of the classes at each stage in order to perform the stage-by-stage classification. The SDECD can be expressed as[14]

$$J = \frac{\sqrt{\sum_{i=1}^{d}(\mu_{1i} - \mu_{2i})^2}}{3(\frac{1}{d}\sum_{i=1}^{d}\sigma_{1ii} + \frac{1}{d}\sum_{i=1}^{d}\sigma_{2ii})} \tag{6}$$

where σ_{1ii} and σ_{2ii} $(i = 1, 2, 3, \ldots, d)$ represent the standard deviations of variables, $\boldsymbol{\mu}_1 = [\mu_{11}, \mu_{12}, \ldots, \mu_{1d}]^{\mathrm{T}}$ and $\boldsymbol{\mu}_2 = [\mu_{21}, \mu_{22}, \ldots, \mu_{2d}]^{\mathrm{T}}$ are the expected vectors.

During the training phase, the estimator $\boldsymbol{\beta}$ was computed by equation (4) using the selected training sets at each stage of the classification. During the testing phase, the output response at each stage of the classification was computed using the feature vectors and the previously estimated $\boldsymbol{\beta}$ by equation (5). A threshold value of zero was used to classify the output response at a particular stage. The average sensitivity and specificity were computed for all the classes for measuring the performance of the classification[15].

Table 1 Classification algorithm for MAR and K-L MAR coefficients

Stage 1			Stage 2			Stage 3			Stage 4		
Groups	Member ship	Decision-making	Groups	Member ship	Decision-making	Groups	Member ship	Decision-making	Groups	Member ship	Decision-making
NSR	1	$Y_1 > 0$	NSR	-1	$Y_2 < 0$	APC/NSR	-1	$Y_3 < 0$	NSR	1	$Y_5 > 0$
APC	1	$Y_1 > 0$	APC	-1	$Y_2 < 0$	PVC/NSR	1	$Y_3 > 0$	PVC	-1	$Y_5 < 0$
PVC	1	$Y_1 > 0$	PVC	-1	$Y_2 < 0$	VT	-1	$Y_4 < 0$	NSR	1	$Y_6 > 0$
VT/VF	1	$Y_1 > 0$	VT/VF	1	$Y_2 > 0$	VF	1	$Y_4 > 0$	APC	-1	$Y_6 < 0$
SVT	-1	$Y_1 < 0$									

3 Results

3.1 MAR and scalar AR modeling Results

In order to evaluate the performance of the MAR modeling, the SSE was computed over all estimates in the length of modeled ECG signals. The results showed that the SSE decreased initially with the model order P, but remained almost constant for model order greater than or equal to three. However, MAR model of order four was selected for extracting the features. This is because more details can be incorporated into the model order, which might be missing from a lower-order model. On the other hand, the number of the MAR coefficients and computation for higher orders would increase rapidly. So the MAR model of order 4 is a fitter selection.

Scalar AR modeling has been performed for the purpose of classification. A fitter scalar AR model of order 4 was found to model the ECG using SSE criterion calculated from single-lead ECGs and over all estimates in the 225-point window. This result was consistent with the other researches on the scalar AR model order selection[16].

3.2 Classification results

A MAR model of order 4 and a scalar AR model of order 4 were selected to model the ECG signals in the current research. The MAR coefficients computed with order 4, the K-L MAR coefficients and the scalar AR coefficients estimated with order 4 were used for QDF-based classification.

3.2.1 Classification results based on MAR and K-L MAR coefficients

The ECG features were extracted by applying MAR process of order 4 to the two-lead ECG signals. This resulted in the 16 MAR coefficients to represent a two-lead ECG segment in this research. Table 1 shows the classification algorithm for this case. The values of SDECD between these classes were computed for determining the groupings of classes at each stage. Table 2 shows the values of SDECD based on the MAR coefficients. One can see that APC/NSR/PVC, VT/VF and SVT form one group respectively due to small values of SDECD within the same group and large values between different groups. Therefore, SVT was separated from APC/NSR/PVC and VT/VF in stage one (Y_1). The membership of SVT was defined as "-1", and the membership of APC/NSR/PVC and VT/VF was defined as "+1". The least squares estimator $\boldsymbol{\beta}$ was computed as equation (4). The output response Y_1 was computed as equation (5). The value of Y_1 was used to determine the classes. Similarly, VT/VF and APC/NSR/PVC were distinguished between each other in the second stage (Y_2). Stage three (Y_3 and Y_4), four (Y_5 and Y_6) were used to differentiate between APC, NSR, PVC, VT and VF as shown in Table 1.

One hundred and fifty cases each from the six classes were selected at random to estimate $\boldsymbol{\beta}$ in training phase, and the remaining were used for testing in testing phase. The classification results based on the MAR coefficients on testing data are given in Tables 3 and 4. Table 3 shows a classification results based on the MAR coefficients for a sample training set. Table 4 shows the performance of classification based on the MAR coefficients for the various classes, which were averaged over 20 runs, each run with different training and testing data sets.

Table 2 Values of SDECD based on MAR coefficients between the different classes

Classes	SVT	APC	PVC	NSR	VT	VF
SVT	0	1.6587	1.3775	1.5718	1.6287	2.8733
APC	1.6587	0	0.9669	1.2325	1.5397	2.8077
PVC	1.3775	0.9669	0	1.1739	1.4374	2.2122
NSR	1.5718	1.2325	1.1739	0	1.9631	2.2082
VT	1.6287	1.5397	1.4374	1.9631	0	1.0671
VF	2.8733	2.8077	2.2122	2.2082	1.0671	0

Table 3 Classification results based on MAR coefficients for a sample training set

Classes	SVT	APC	NSR	PVC	VT	VF
SVT	148	0	0	2	0	0
APC	0	147	3	0	0	0
NSR	0	1	149	0	0	0
PVC	0	0	0	149	1	0
VT	0	0	0	0	150	0
VF	0	0	0	0	0	150

Table 4 Performance of the classification based on MAR coefficients

Classes	SVT	NSR	APC	PVC	VF	VT
Sensitivity	98.6%	99.3%	98.0%	99.3%	100%	100%
Specificity	100%	98.0%	99.3%	98.6%	100%	99.3%

The number of the eigenvectors was chosen to be 10 according to the choice criterion of eigenvectors described in section 2. Thus, 10-dimensional feature vectors based on K-L MAR coefficients were obtained after K-L transformation. The 10-dimensional feature vectors were trained and tested the same way as in the MAR coefficients based classification experiments, the classification results based on the K-L MAR coefficients on the testing data are given in Table 5.

Table 5 Performance of the classification based on K-L MAR coefficients

Classes	SVT	NSR	APC	PVC	VF	VT
Sensitivity	97.3%	99.3%	96.6%	95.3%	98.6%	96.6%
Specificity	96.6%	93.3%	99.3%	98.0%	99.3%	97.3%

3.2.2 Classification results based on scalar AR coefficients

A scalar AR process of order 4 was performed on each ECG lead from the six classes. Thus, the number of the scalar AR coefficients to represent a two-lead ECG segment was 8. A similar analysis method was employed for the scalar AR coefficient classification. The classification results based on the scalar AR coefficients and two-lead ECG segments are shown in Table 6.

Table 6 Performance of the classification based on scalar AR coefficients and two-lead ECG segments

Classes	SVT	NSR	APC	PVC	VF	VT
Sensitivity	96%	99.3%	96.6%	98.0%	98.6%	97.3%
Specificity	99.3%	94.6%	99.3%	96.0%	99.3%	98.0%

The classification results based on scalar AR coefficients and single-lead ECG are given in Table 7. It is for the purpose of comparison between one-lead ECG signal and two-lead ECG signal based classification.

Table 7 Performance of the classification based on single-lead ECG signals

Classes	SVT	NSR	APC	PVC	VF	VT
Sensitivity	90.0%	98.6%	94.6%	92.6%	99.3%	92.0%
Specificity	95.3%	86.0%	99.3%	92.0%	97.3%	98.0%

4 Discussions

The main objective of this study was to model two-lead ECG signals for extracting features in order to explore the feasibility to classify more types of cardiac arrhythmias using MAR and AR modeling. The modeling results showed that the MAR order of 4 was sufficient to model the ECG signals for the purpose of the classification, scalar AR order of 4 was also sufficient for the same purpose. It was reported that the sufficient MAR model order was 25 for modeling HR, BP, and RESP for the purpose of assessment of interaction between them in [8].

Extra calculation was involved in calculating K-L transform of MAR coefficients. This representation may not be worth considering for a real-time system. The classification of the scalar AR coefficients extracted from two-lead ECGs produced the similar percentages of the accuracy compared to classification of the K-L MAR coefficients. The classification of the scalar AR coefficients extracted from signal-lead ECGs gave the lowest classification accuracy. Thus, the MAR coefficients would be the most efficient ECG signal representation. Using two-lead ECG signals can improve the classification accuracy significantly compared with single-lead ECG signals.

The current study classifies six types of ECG arrhythmias, and some of the proposed techniques use only a smaller number of arrhythmias than the current study. For example, two AR coefficients and the mean-square value of QRS complex segment were utilized as features for classifying PVC and NSR using a fuzzy ARTMAP classifier, sensitivity of 97% and specificity of 99% were achieved in [12], the total least sequare-based Prony modeling technique was used for detecting SVT, VT and VF, accuracy of SVT, VT and VF were 95.24%, 96% and 97.78% in [5]. The classification algorithms based on MAR modeling are easy to implement. In this study, the sample size of the various segments was 0.9 seconds only, and it was 3 to 7 seconds and 5 to 9 seconds for the complexity measure-based technique in [3] and the Prony modeling technique in [5], respectively.

The MAR model might not be suited to ECG signals under all conditions since MAR model is a linear modeling technique, nonlinear parametric modeling might improve the results. Future work would involve real-time data collection in order to test our hypothesis and determine the precision of our methodology.

5 Conclusions

MAR coefficients extracted by fusing two ECG leads could be used as features to classify certain cardiac arrhythmias effectively in critical ill patients for real-time automatic diagnosis purpose.

References

1 Chen S, Thakor N V, Mover M M. Ventricular fibrillation detection by a regression test on the autocorrelation function. *Medical and Biological Engineering and Computing*, 1987, **25**(3): 241∼249

2 Afonso V X, Tompkins W J. Detecting ventricular fibrillation: Selecting the appropriate time frequency analysis tool for the application. *IEEE Engineering in Medicine and Biology Magazine*, 1995, **14**(2): 152∼159

3 Zhang X S, Zhu Y S, Thakor N V, Wang Z Z. Detecting ventricular tachycardia and fibrillation by complexity measure. *IEEE Transactions on Biomedical Engineering*, 1999, **46**(5): 548∼555

4 Jekova I. Comparison of five algorithms for the detection of ventricular fibrillation from the surface ECG. *Physiological Measurement*, 2000, **21**(4): 429∼439

5 Chen S W. Two stage discrimination of cardiac arrhythmias using a total least squares based prony modeling algorithm. *IEEE Transactions on Biomedical Engineering*, 2000, **47**(10): 1317∼1326

6 Caswell S A, Kluge K S, Chiang C M J, Jenkins J M, Carlo L A. Pattern recognition of cardiac arrhythmias using two intracardiac channels. In: Proceedings of Computers in Cardiology. London, UK, IEEE, 1993. 181∼184

7 Melo S L, Caloba L P, Nadal J. Arrhythmia analysis using artificial neural network and decimated electrocardio graphic data. In: Proceedings of Computers in Cardiology. Piscataway, USA, IEEE, 2000. **27**: 73∼76

8 Jimenez J C, Biscay R, Montoto O. Modelling the electroencephalogram by means of spatial spline smoothing and temporal auto regression. *Biological Cybernetics*, 1995, **72**(3): 249∼259

9 Keirn Z A, Aunon J I. A new method of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering*, 1990, **37**(12): 1209∼1214

10 Arnold M, Miltner W H R, Witte H. Adaptive AR modeling of nonstationary time series by means of Kalman filtering. *IEEE Transactions on Biomedical Engineering*, 1998, **45**(5): 553∼562

11 Mainardi L T, Bianchi A M, Baselli G, Cerutti S. Pole tracking algorithms for the extraction of time variant heart rate variability spectral parameters. *IEEE Transactions on Biomedical Engineering*, 1995, **42**(3): 250∼258

12 Ham F M, Han S. Classification of cardiac arrhythmias using fuzzy ARTMAP. *IEEE Transactions on Biomedical Engineering*, 1996, **43**(4): 425∼430

13 Tompkins W J. *Biomedical Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice Hall, 1993, 246∼261

14 Fukunaga K. *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1990, 153∼154 and 400∼409

15 Barro S, Ruiz R, Cabello D, Mira J. Algorithmic sequential decision making in the frequency domain for life threatening ventricular arrhythmias and imitative artefacts: a diagnostic system. *Journal of Biomedical Engineering*, 1989, **11**(4): 320∼328

16 Ge Ding-Fei, Xia Shun-Ren. Application of AR model in telediagnosis of cardiac arrhythmias. *Chinese Journal of Biomedical Engineering*, 2004, **23**(3): 222∼229 (in Chinese)

GE Ding-Fei Received his master degrees from Nanyang Technological University, Singapore in 2003. Now he is an associate professor in Zhejiang University of Science and Technology. His research interest covers pattern recognition and data mining. Corresponding author of this paper. E-mail: gedingfei@hotmail.com

HOU Bei-Ping Received his Ph. D. degree from Zhejiang University in 2005. His research interest covers machine vision, image processing, and pattern recognition.

XIANG Xin-Jian Professor of Zhejiang University of Science and Technology. His research interest covers intelligent control and application of pattern recognition.