Editors

A. Charalambopoulos
D. I. Fotiadis
D. Polyzos

ADVANCED TOPICS IN SCATTERING AND BIOMEDICAL ENGINEERING

Proceedings of the Eighth International Workshop on Mathematical Methods in Scattering Theory and Biomedical Engineering

World Scientific

SVM CLASSIFICATION OF HOLTER ECG BEATS USING WAVELET FEATURES*

MICHAL HUPTYCH[1], GEORGE GEORGOULAS[2], VACLAV CHUDACEK[1], CHRYSOSTOMOS STYLIOS[3], LENKA LHOTSKA[1]

[1]*Department of Cybernetics, Czech Technical University in Prague, Czech Republic*
[2]*Intelligent Control Systems Laboratory, Georgia Tech, Atlanta, USA*
[3]*Dept. of Informatics and Communications Technology, TEI of Epirus, Artas, Greece*

In this work we investigate the use of discrete wavelet transform as a means to extract features capable of providing the information needed by a nonlinear classifier to discriminate between Ventricular (V) and Normal (N) beats. Different mother wavelets have been used with different number of vanishing moments producing features to be fed to a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernels. For testing the proposed approach the MIT Arrhythmia database is employed in a local training scheme. By local training we mean that each patient is used as each own control. This approach is selected in order to encompass individual's variations in a personalized treatment.

1. Introduction

Noninvasive investigation of the heart action is one of the most important methods for early diagnosis of heart disease. Long term holter monitoring is widely applied to patients with heart problems such as arrhythmias. The primary task of computer-aided systems in holter ECG evaluation is to distinguish between different beat types.

Among the various beats encountered, the discrimination between Ventricular (V) and Normal (N) is of paramount importance. Therefore in this work we have employed the wavelet transform to process each beat and extract a number of features to subsequently be used to characterize the corresponding beat. Since the feature extraction stage is more of an art than a science, we have extracted a plethora of features, which are then linearly combined using Principal Component Analysis (PCA) reducing at the same time the dimension of the input vector. Subsequently the reduced vector is fed to a Support Vector Machine (SVM).

In order to test the proposed approach we have used MIT-BIH arrhythmia database [6]. The MIT database includes 48 ECG records, which are half an hour long at a sampling frequency of 360 Hz. Since we have focused only on the discrimination between ventricular and normal beats, only beats labeled 'V' or 'N' have been selected for the classification purpose.

The rest of the paper is structured as follows: section 2 summarizes the basics of wavelet theory and section 3 presents the way wavelets are used to extract features-indices from the N and V beats. Section 4 presents the main concepts utilized by SVMs and Section 5 concludes the paper presenting the results of our approach and some hints for future research.

2. Wavelet Transform

2.1. *Continuous Wavelet Transform*

The continuous wavelet transform (CWT) is a time-frequency analysis method which differs from the short time Fourier transform by the way it allows localization of the information in the time frequency plane. The continuous wavelet transform of function a *f(t)* with respect to a mother wavelet is defined as [1]:

$$(T^{wav} f)(a,b) = \int f(x) a^{-1/2} \psi\left(\frac{x-b}{a}\right) \quad (1)$$

where $a,b \in R$, $a \neq 0$ and a scales and b is translates the mother wavelet ψ. Thus, if we define $\psi_{a,b}(t)$ as the translated and scaled $\psi(t)$, then $\psi_{a,b}(t)$ is given by:

$$\psi_{a,b}(t) = a^{-1/2} \psi\left(\frac{t-b}{a}\right) \quad (2)$$

2.2. *Discrete Wavelet Transform*

Most of computer implementations use a Discrete version of the Wavelet Transform (DWT). In DWT, a and b are discretized. The scaling parameter a is discretized using $a = p^j, j \in Z, p \neq 0$. The equation for wavelet is given by [1]:

$$\psi_{j,b}(t) = p^{-j/2} \psi\left(\frac{t - nbp^j}{p^j}\right) = p^{-j/2} \psi\left(p^{-j} t - nb\right) \quad (3)$$

If $p = 2$ and $n = 2^{-j}$, the sampling of time-frequency space is called dyadic, which is the most common and the analyzing wavelet has the form:

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t - 2^{-j}b\right) = 2^{-j/2}\psi\left(2^{-j}t - k\right) \quad \text{where } k = 2^{-j}b \tag{4}$$

2.3. *Multi-Resolution Analysis*

The Multi-resolution analysis (MRA) is in the heart part of the wavelet theory. The idea of MRA has been developed by Mallat and Meyer [2]. In MRA a function is decomposed into an approximation (representing the "slow" variations of the function) and a detail (representing the "rapid" variations of the function), on a level by level basis, by a scaling and a wavelet function.

Thus, let the signal *f(t)* be described by the scaling $\varphi_k(t)$ and wavelet function $\Psi_{jk}(t)$ [1] as:

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \varphi_k(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d_{jk}\Psi_{jk}(t) \tag{5}$$

where c_k is called the approximation of the signal and d_{jk} is called the detail.

2.4. *Filtering Representation*

In practice often the filter bank implementaion (quadrature mirror filters - QMF) and Mallat's algorithm are used for the DWT. The impulse response of a low pass filter corresponds to the scaling function, meaning that the output of the low pass filter is the approximation of the signal. In the same way the impulse response of a high pass filter corresponds to the wavelet function (the output of the high pass filter consists the detail signal). For the filters consisting the bank, the following relations must hold [1]:

$$\begin{aligned} \tilde{a}^0(z)a^0(z) + \tilde{a}^1(z)a^1(-z) &= 0 \\ a^1(z) &= a^0(-z) \\ \tilde{a}^0(z) &= a^0(z) \\ \tilde{a}^1(z) &= -a^0(-z) \end{aligned} \tag{6}$$

From Eq. (6) it is obvious that we need only the coefficients of a^0 in order to calculate all the coefficients of the filters in the filter bank. However we want to get coefficients of impulse response:

$$h(z^{-1}) = a^0(z), \text{so} \quad h(z) = \tilde{a}^0(z) \tag{7}$$

Between low pass and high pass coefficients the following relation holds:

$$g(L-1-n) = (-1)^n h(n), \text{ resp.} \quad g(n) = (-1)^n h(-n+1) \tag{8}$$

From these equations we can get coefficients for reconstruction filters. For decomposition filters the following equation is suitable:

$$\overline{h}(n) = \overline{h(-n)}, \text{ and } \overline{g}(n) = \overline{g(-n)} \tag{9}$$

The reconstruction filter *g(n)* is calculated as the complement of filter *h(n)*, and decomposition filters are defined as the time reverse sequence of *h(n), g(n)*.

3. Feature Extraction

Each record of MIT database has been divided into single beats according to location of the R wave peak. For feature extraction we have used three approaches based on the DWT coefficients and indices calculated from them. The DWT has been implemented using QMFs [3]. Three families of orthonormal wavelets were tested namely Daubechies, coiflet and symmlet wavelets with different number of vanishing moments [1]. Feature extraction has been performed using both approximations and details from the first up to the sixth level of decomposition.

The first set of features have been extracted using the approximation and detail coefficients of the undecimated DWT. The output signal of the undecimated DWT has the same number of samples as input signal. A summary of these features is given in Table 1.

Table 1. Summary of features derived from the undecimated DWT

Amplitude and areas features	Features of position	Statistical features
maximum	position of maximum	average
minimum	position of minimum	median
sum of positive area		variance
sum of negative area		standard deviation

Furthermore the coefficients of the approximations and details produced by the decimated DWT have been used as additional features [4]. We have used coefficients from the third and fourth level (both approximation and detail). Therefore we have obtained 37 coefficients from third level and 18 coefficients from fourth level. Figure 1 depicts the undecimated and decimated DWT of a

beat using a coiflet wavelet of order 2 for the third and the fourth level of decomposition.

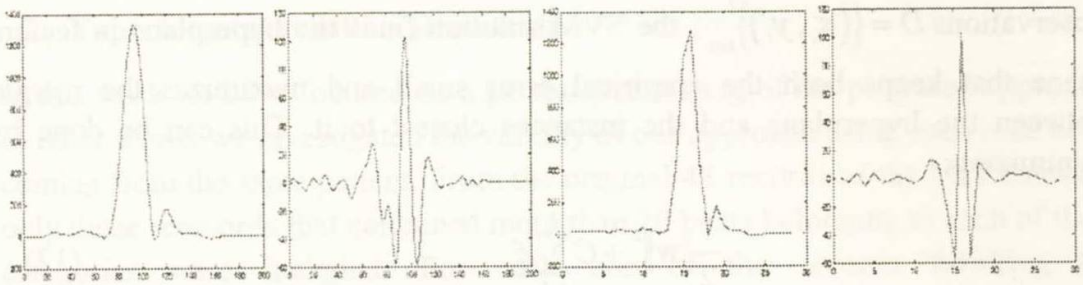

Figure 1. Example of undecimated DWT approximation and detail (left) and decimated DWT approximation and detail (right)

Moreover the coefficients of a filter are used as additional features. In general, for the transfer function of a linear system, the following relation holds [5]:

$$y(k) = \sum_{k-\infty}^{\infty} h(n)x(k-n), \tag{10}$$

where *h(n)* is the intrinsic function of the system, *x(k)* is the input signal and *y(k)* is output signal.

Let Q_J be the Fourier transform of impulse response q_J. Then for the impulse response of filter for the decomposition up to J level by Mallat's algorithm we have [3]:

$$Q_J(\omega) = G(2^{j-1}\omega)H(2^{j-2}\omega)...H(\omega),\ j = 1,2,...J, \tag{11}$$

where $G(\omega)$ is the Fourier transform of the high-pass filter *g(n)* and $H(\omega)$ is Fourier transform of low-pass filter *h(n)*. This means that we can obtain an impulse response for all decompositions at any level. Our idea is to use the impulse response coefficients *h(n)* as an expression of correlation between a normal ECG beat template and an unknown ECG beat. The template beat is obtained as the median of 20 randomly selected N beats. The unknown beat is decomposed to second level by the given wavelet. Using Eq. (10) we identify the intrinsic function *h(n)* of the system which has as input, *x(n)*, the template, and as output, *y(n)*, the second approximation of a given unknown ECG beat. .

4. Support Vector Machine

Support Vector Machines are learning systems that are trained using an algorithm based on optimization theory [7]. For real life problems, given l observations $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, the SVM solution finds the hyperplane in feature space that keeps both the empirical error small and maximizes the margin between the hyperplane and the instances closest to it. This can be done by minimizing:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \tag{12}$$

subject to

$$y_i\left(\langle \mathbf{w}\cdot\boldsymbol{\varphi}(\mathbf{x}_i)\rangle + b\right) \geq 1-\xi_i \quad \text{with } \xi_i \geq 0 \;,\; i = 1,2,\ldots,l$$

where ξ_i are slack variables, which are introduced to allow the margin constraints to be violated, and $\boldsymbol{\varphi}(\cdot)$ is the nonlinear mapping from the input space to the feature space. Parameter C controls the trade off between maximizing the margin and minimizing the error and it is usually determined through a cross-validation scheme [7].

The class prediction for an instance $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l} y_i\alpha_i \langle \boldsymbol{\varphi}(\mathbf{x}_i)\cdot\boldsymbol{\varphi}(\mathbf{x})\rangle + b\right) \tag{13}$$

where the coefficients a_i are calculated by maximizing the Lagrangian:

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \langle \boldsymbol{\varphi}(\mathbf{x}_i)\cdot\boldsymbol{\varphi}(\mathbf{x}_j)\rangle \tag{14}$$

subject to $\sum_{i=1}^{l} y_i\alpha_i = 0$ and $0 \leq \alpha_i \leq C$, $i = 1,2,\ldots,l$

The points for which a_i>0, are called Support Vectors and are the points lying closest to the hyperplane. If the nonlinear mapping function is chosen properly, the inner product in the feature space can be written in the following form:

$$\langle \boldsymbol{\varphi}(\mathbf{x}_i)\cdot\boldsymbol{\varphi}(\mathbf{x}_j)\rangle = K(\mathbf{x}_i, \mathbf{x}_j) \tag{15}$$

where K is called the inner-product kernel [7].

Among others the most popular are the polynomial learning machines, the radial basis function networks and the two-layer perceptrons. In our experimental procedure we have employed the radial basis function kernels.

5. Experimental results – Discussion

In this work we have focused on a personalize testing of the proposed approach. In other words we investigated the validity of our approach using each time beats coming from the same patient. From the original 48 records of the MIT database only those rerecords that contained more than 20 beats belonging to each of the 2 categories were included. For each one of the records fulfilling the aforementioned requirement (N, V>20) we have selected 20 beats at random belonging to the N class and 20 beats belonging to the V class. Due to the stochastic approach we have repeated each experiment 10 times and averaged the results.

Using the wavelet transform 235 features have been extracted. With such a large input vector even for an SVM the course of dimensionality can be a serious issue. Therefore we have employed a stage of feature reduction based on PCA (retaining only 20 Principal Components) before feeding the SVM.

For the selection of the appropriate SVM we tested different configurations for the C and σ parameters. The results for the best configuration and for 9 different mother wavelets are summarized in Table 2. Due to the imbalance nature of the problem the overall accuracy is not the best metric in order to assess the performance of the classifier. Therefore we have used the notion of sensitivity and specificity selecting the best classifier in terms of its geometric mean ($gmean = \sqrt{sensitivity * specificity}$)

Table 2. Performance of the proposed method for different mother wavelets

Wavelet family	Order	sensitivity	specificity	accuracy	gmean
coiflet	2	0.9815	0.9726	0.9803	0.9770
coiflet	3	0.9805	0.9713	0.9793	0.9758
coiflet	4	0.9807	0.9758	0.9801	0.9783
symmlet	2	0.9741	0.9769	0.9745	0.9755
symmlet	4	0.9800	0.9679	0.9784	0.9739
symmlet	6	0.9776	0.9795	0.9779	0.9785
daubechies	2	0.9785	0.9792	0.9786	0.9789
daubechies	4	0.9792	0.9825	0.9796	0.9808
daubechies	6	0.9797	0.9782	0.9795	0.9789

As it can be seen from the above Table 2, the proposed approach performs very satisfactory when applied to each subject alone, showing that there is a

consistent pattern for each one the patients. Therefore by having an expert annotating the first few beats an automatic system could take care the marking of the rest of the (usually very long) recording. Moreover it seems that the selection of neither the wavelet family nor the number of vanishing moments plays a crucial role on the performance of the method.

In future work we will try to substitute the dimensionality reduction stage (PCA stage) with a feature selection one, based on evolutionary approaches, because by retaining most of the variance of the data does not necessarily guarantee that we can have a superior classification performance. Moreover we will investigate the performance of this method when applied across different patients to test how the inter-variance can effect it.

Acknowledgments

This work has been supported by the research program "Information Society" under grant No. 1ET201210527 "Knowledge-based support of diagnostics and prediction in cardiology".

This research work has partially supported by the joint research project "Intelligent System for Medical Data Analysis and Evaluation Based On Computational Intelligence Techniques" under the Greek-Czech bilateral cooperation agreement, funded by Greek GSRT

References

1. I. Daubechies, *CBMS-NSF, SIAM*, **ISBN 0-8987-274-2**, 61 (1992).
2. S. Mallat, *Academic Press, An Imprint of Elsevier*, **ISBN 10 0-12-466606-XD**, (1997)
3. S. Mallat, *IEEE Trans. Pattern Anal. Machine Intelligence*, **ISSN 0162-8828,** 14 (1992).
4. S. Z. Mahmoodabadfi, A. Ahmadian, D. Abolhasani, M. Eslami and J. H. Bidgoli, *Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, (2005)
5. D. G. Manolakis, V K. Ingle and S. M. Kogon, *Artech House*, Norwood, MA, **ISBN 9780070400511** (2005)
6. G. B. Moody and R. G. Mark, *CD-ROM and software for use with it. In Computers in Cardiology*, pp 185–188, IEEE Computer Society Press, (1990).
7. K. R. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf, *IEEE Trans. Neural Networks*, vol. 12, no. 2, 181-201, (2001)