Asian Journal of Information Technology 5(6): 584-590, 2006

ECG Beats Recognition Using Normalized Ellipsoidal Basis Function Network

Djemil Messadeg, Messaoud Ramdani, Mouldi Bedda and [1]Herman Akdag
Department of Electronics, Faculty of Engineering, Lasa Laboratory,
University of Annaba, BP. 12, 23000, Annaba, Algeria
[1]Crestic-Leri Reims University, France

Abstract: In this study, we propose a neural network model for the electrocardiogram (ECG) beat recognition. The description of the ECG signals consists of a multi-domain features which contain a set of meaningful and non redundant parameters. The construction of the system is accomplished by a data-driven learning scheme based on a clustering process to find an initial or coarse neuronal structure and a fine tuning hybrid learning algorithm, including gradient descent nonlinear optimization procedure and a least squares optimization step. The salient features of the system are an effective mechanism for variable learning rates and an adaptive metric norm for the distance. The results of experiments show the good efficiency of the proposed solution.

Key words: ECG beat recognition, multi-features, Normalized Ellipsoidal Basis Function network (NEBF), adaptive learning rates

INTRODUCTION

With the advances of data acquisition systems, it has become possible to collect and store huge amounts of data in the biomedical domain. Moreover, due to the lack of explicit relations among data and to the unstructured nature of medical knowledge; the decision on a patient's diagnosis based on heuristic and analytical symptoms of an examination, requires more and more elaborated tools. Among all the biomedical techniques, the Electrocardiogram (ECG) signal analysis remains one of the most relatively inexpensive and easily accessible investigational tools in clinical cardiology.

For several years, computer-assisted ECG interpretation is playing an increasing role in assisting cardiac diagnosis. Generally, automatic diagnosis can be viewed as a sequential process involving two steps: The symptom extraction and the diagnosis task. Symptom extraction is mainly required to reduce data and to find some qualitative and quantitative features. Nevertheless, the lack of proper diagnostic criteria, which are usually expressed in a natural language, leads to the difficulty of formalizing medical knowledge in a computer program.

Artificial neural networks and fuzzy logic systems have been studied extensively and applied in this field. Neural networks are well-known for their powerful computational and learning abilities but they do not implement a transparent decision process to the user and lack the ability of dealing with expert knowledge, while fuzzy systems are by now famous for their easy interpretability. In this context, neuro-fuzzy systems are defined in the form of IF-then rules trained by a learning algorithm of data driven type derived from neural networks theory.

In this study, we present a neuronal system of ECG beat recognition using various features that are less sensitive to morphological variation of the ECG. More specifically, instead of real waveform, two different types of feature sets are defined, namely, Auto-Regressive (AR) model coefficients and Discrete Wavelet Transform (DWT) based features of the related ECG beats. It will be shown that the derived features are less sensitive to the morphological variation of the ECG.

ECG WAVEFORM DESCRIPTION

Automatic ECG beat recognition can be performed using decision-tree like approaches, artificial neural networks and fuzzy systems based on various features extracted from ECG beat, mainly in the temporal domain, such as the width and height of QRS complex, RR interval, QRS area, etc. The main difficulty is that these features are very susceptible to variations in morphology and temporal characteristics. Thus, it is necessary to define some characteristic features in different domains that are more robust to variations of ECG morphology. In the present study, only two different features describing isolated ECG beats are proposed as candidates to form a compact representation.

Corresponding Author: Djemil Messadeg, Department of Electronics, Faculty of Engineering, Lasa Laboratory, University of Annaba, BP. 12, 23000, Annaba, Algeria

Linear prediction coefficients: A linear auto-regressive model can be used in time series analysis to predict the value of the next sample of a signal. The latter is taken as a linear combination of the previous samples. The next sample of the time series, $\bar{S}_k$ is predicted as the weighted sum of the p previous samples S_{k-1}, S_{k-2}, ..., S_{k-p} and can be given by the following expression:

$$\bar{s}_k = a_1 s_{k-1} + a_2 s_{k-2} + \cdots + a_p s_{k-p} = \sum_{i=1}^{p} a_i s_{k-i} \qquad (1)$$

The transfer function of the model is given by,

$$H(z) = \frac{\bar{S}(z)}{S(z)} = \sum_{i=1}^{p} a_i z^{-i} \qquad (2)$$

Where, a_1, a_2, ..., a_p represent the model coefficients and p its order. The residual error, e_k, is defined as the difference between the actual and the predicted values of the next sample and can be expressed as,

$$e_k = s_k - \bar{s}_k = s_k - \sum_{i=1}^{p} a_i s_{k-i} \qquad (3)$$

The weights can be computed by minimizing the mean square value of the residual errors over an analysis window.

DWT based features: The Continuous Wavelet Transform (CWT) is defined as the integral of the signal s(t) multiplied by scaled, shifted versions of a basic wavelet functionψ(t):

$$c(a,t) = \int_R s(t) \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) dt, a \in R^+ - \{0\}, b \in R \qquad (4)$$

where a is the so-called scaling parameter, b is the time localisation parameter. Associated with wavelet ψ, which is used to define the details in the decomposition, a scaling function $\phi$is used to define the approximations. To avoid intractable computations of the CWT, scales and positions can be chosen based on a power of two, i.e dyadic scales and positions. The Discrete Wavelet Transform (DWT) analysis is more efficient and accurate[1]. In this scheme, the parameters a and b are given by:

$$(j,k) \in Z^2 : a = 2^j, b = k2^j, Z = \{0, \pm 1, \pm 2, \cdots\}$$

This allows us to define

$$\psi_{j,k}(t) = 2^{-j/2} \psi\left(2^{-j} t - k\right),\ \phi_{j,k}(t) = 2^{-j/2} \phi\left(2^{-j} t - k\right) \qquad (5)$$

A wavelet filter with impulse response g, plays the role of the wavelet ψ and a scaling filter with impulse response h, plays the role of scaling function ϕ. Thus, the DWT can be described mathematically as:

$$c(j,k) = \sum_{n \in Z} s(n) g_{j,k}(n)$$
$$a = 2^j,\ b = k2^j,\ (j,k) \in Z^2$$

The detail at level j is defined as:

$$D_j(t) = \sum_{k \in Z} c(j,k) \psi_{j,k}(t) \qquad (6)$$

In practice, the decomposition can be determined iteratively, with successive approximations being computed, such that the analysed signal is decomposed into many lower-resolution components. In the present study, a five level DWT is defined and the normalized variances of the details coefficients are used as features.

FEEDFORWARD NEURAL NETWORKS (FNN) FOR PATTERN RECOGNITION

Feedforward neural networks have been increasingly used in many areas to solve real-word problems. This is mainly due to their universal approximation capabilities, i.e to the property that any continuous function can be approximated within an arbitrary accuracy by means of a neural network, provided that its topology includes a sufficient number of hidden units[2-4]. Basically, a typical FNN is a nonlinear regression technique, which is determined by performing a training process where the goal is to find the parameters that minimize a suitable error function. This is achieved by using a given number of pattern-target pairs that are samples of the input-output relationships to be approximated.

In the study under consideration, we limit ourselves to address classification problems of different ECG beats via a class of RBF network. One reason of our choice is that they form a unifying link between function approximation, regularization, noisy data interpolation, classification and density estimation. It is also the case that training RBF networks is usually faster than multi-layer perceptron networks.

The construction of RBF networks is usually solved in two steps. First, the basis function parameters (positions and widths or spreads of the basis functions) may be determined by unsupervised clustering algorithms; they can also be obtained through growing and pruning procedures[6,7]; or they can be evolved using evolutionary algorithms[5,8]. Second, the final layer weights are determined by least squares optimization which reduces to solving linear system. It is well known that the problem of selecting the appropriate number of basis functions remains a critical issue because it controls the complexity and hence the generalization ability of RBF networks. For instance, few basis functions give poor prediction on unseen data, i.e. poor generalization, since the model has limited flexibility. On the other hand, an RBF network with too many basis functions yields poor generalization since it is too flexible and fits the noise in the training data. Thus, the well-known trade-off bias-variance which highlights the importance of optimizing the complexity of the model in order to achieve the best generalization via a compromise between the conflicting requirements of reducing bias while reducing variance at the same time.

On the other hand, there is a clear evidence that the classification/regression error made the RBF networks depend strongly on the shape of the kernel functions constituting the hidden layer[9,10]. More specifically, it seems more reasonable and beneficial if diagonal covariance matrices could be incorporated into the basis functions so that complex data distributions could be represented efficiently without the need of having to use a large number of basis functions. In this way, the range of spreads of a hidden unit is ellipsoidal and the traditional RBF network is extended into an Ellipsoidal Basis Function (EBF). Moreover, taking a useful normalization variation gives rise to a generalized or Normalized Ellipsoidal Basis Function (NEBF) representation. To make the NEBF network more effective in handling complex classification/regression tasks, an effective data-driven hybrid learning scheme is proposed. The identification of an initial neuronal structure is accomplished by an unsupervised maximum-entropy clustering process. The initial neuronal model can be fine tuned by an efficient two stage Hybrid Learning Algorithm (HLA) combining Gradient-Descent (GD) optimization with least squares based on singular decomposition (SVD). The SVD is a very stable algorithm to handle numerical computation problems associated with the inversion of ill-conditioned matrices. In order to speed up the learning process while avoiding local minima, the proposed GD optimization uses a variable learning rate for every parameter, depending on the progress of the cost function to be minimized.

Assume we have a complex nonlinear multi-input and multi-output (MIMO) relationship where $[x_1, x_2, ..., x_n]^T \in X \subset \Re^n$ is the vector of input variables and $y \in Y \subset \Re^m$ is the vector of output variables. In the multi-input and multi-output NEBF network given in Fig. 1., the overall output is defined as:

$$\hat{y}_j(x) = \sum_{i=1}^{H} f_{ij}\phi_i \Big/ \sum_{i=1}^{H} \phi_i \tag{7}$$

Where j = 1,2,..., m; I = 1,2,..., H and l = 1,2,..., n;

$$\phi_i = \prod_{l=1}^{n} \exp\left\{-(x_l - c_{il})^2 / (\sigma_{il})^2\right\} \tag{8}$$

Here, we assume that $c_{il} \in X_i$, $б_{il} > 0$ and $f_{ij} \in Y_i$ with X_i and Y_i are the variation domains of the input x_i and output y_j, respectively. It is important to notice that since each basis function is described by the vectors $c_i = [c_{i1},...,c_{in}]^T$ and $б_i = [б_{i1},...,б_{in}]^T$ of centre and width, it is able to match the local shapes of the underlying clusters that can exist in nonlinear relationships. In this way, the NEBF network becomes efficient because it can develop local adaptive metric norms.

HYBRID LEARNING SCHEME

The learning process is performed in two phases. Firstly, a clustering algorithm is used to find a coarse model that roughly approximates the underlying input-output relationship. Secondly, parameter optimization procedure is performed for a better tuning of the initial structure. In principle, once an appropriate structure is identified, the learning task can be acomplished by any suitable training algorithm such as the standard Backpropagation Algorithm (BPA). However, because of slow convergence speed of pure BPA, in the following a more efficient training method, namely the combination of gradient descent with least squares optimization procedure will be used.

Structure identification by clustering: From the available training data that contain N input-output samples, a regression matrix X and an output matrix Y are constructed

$$X = [x_1, \cdots, x_N]^T, \ Y = [y_1, \cdots, y_N]^T \tag{9}$$

Since the study under consideration deals with a *classification* task, the clustering process uses only the input portion of the data is used to discover an underlying structure generating the data. Clustering is a technique to partition a set of samples into exactly *c* disjoint subsets. Samples in the same cluster are somehow similar than other samples in other clusters. One way to

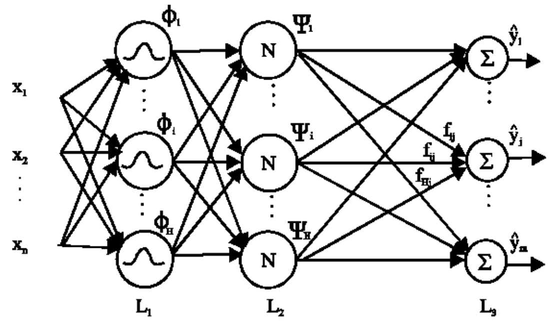

Fig 1: Architecture of the NEBF network

make this problem a well defined one is to define a criterion function that measures the clustering quality of any partition of the data.

Among the clustering methods, the Fuzzy c-means algorithm[11] is one of the most popular. In FCM method, the loss (objective) function is defined as follows:

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2} \tag{10}$$

with $d_{ik} = \|x_k\text{-}v\|_i$ and u_{ik} denotes the grade of membership input vector k to fuzzy cluster i, v_i is interpreted as prototype of cluster i defined by $\{u_{ik}\}$ and weighting exponent m controls the extent of membership sharing between fuzzy clusters. For m=1, FCM converges in theory to the traditional c-means solution. To minimize Eq. 10 subject to normalized condition:

$$\sum_{i=1}^{c} u_{ik} = 1 \tag{11}$$

For each input vector k, using the Lagrangian multiplier method, for m > 1, local minimum of Eq. 10 was demonstrated[11] if and only if

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}} \tag{12}$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^{m} x_k}{\sum_{k=1}^{N} u_{ik}^{m}} \quad \forall i \tag{13}$$

The FCM algorithm is characterized the parameter m that determines the behaviour of the clustering algorithm. The larger m is, the fuzzier is the partition. In this research work, we have used m = 2.

Parameter optimization procedure: The parameters obtained by the identification procedure can be optimized or fine tuned by a variant of gradient descent optimization techniques. This is achieved by an iterative two stage forward-backward optimization algorithm. In the forward stage, with the ellipsoidal basis functions being constant, the weights of the last layer, i.e the functional models f_{ij}; i = 1,..., H and j = 1,...,m are identified by solving a least squares problem. Then, in the backward stage, the functional models are fixed and the parameters of the ellipsoidal functions c_{il}, $б_{il}$, I = 1,..., H; l = 1,...,n are updated by an effective nonlinear Gradient-Descent (GD) optimization technique, which requires the computation of the derivatives of the objective function to be minimized with respect to the parameters c_{il} and $б_{il}$.

The optimization algorithm uses a variable step learning rates. Given a set $D = \{(x^p,d^p)\}^{N}_{p=1}$, such that $x^p \in X \subset \Re^n, d^p \in Y \subset \Re^m$the objective is to find sub-systems $\acute{y}_j\,(x^p)$ in the form of (7), such that the Mean Squared Error (MSE) function

$$E = \frac{1}{2}\sum_{\substack{j=1 \\ x^p \in D}}^{m}\left(\hat{y}_j - d_j^p\right)^2 \tag{14}$$

is minimized. The problem is reduced to the adjustment of the f_{ij} and the mean c_{il} and variance $б_{il}$ of the ellipsoidal functions, so that the MSE is minimized.

Now it can be seen that the network output $\acute{y}_j$ and hence E, depends on c_{il} and $б_{il}$ only through $\phi_i$where $\acute{y}_j$, f_{ij}, b and ψ_i are represented by the following equations:

$$\hat{y}_j = \sum_{i=1}^{H} f_{ij}\psi_i \tag{15}$$

$$\psi_i = (\phi_i/b) \text{ and } b = \sum_{i=1}^{H}\phi_i \tag{16}$$

The Derivatives of E w.r.t c_{il} and $б_{il}$ are given by:

$$\frac{\partial E}{\partial c_{il}} = \frac{\partial E}{\partial \phi_i}\frac{\partial \phi_i}{\partial c_{il}} = \sum_{j=1}^{m}\left(\frac{\partial E}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial \phi_i}\right)\frac{\partial \phi_i}{\partial c_{il}} \tag{17}$$

$$\frac{\partial E}{\partial \sigma_{il}} = \frac{\partial E}{\partial \phi_i}\frac{\partial \phi_i}{\partial \sigma_{il}} = \sum_{j=1}^{m}\left(\frac{\partial E}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial \phi_i}\right)\frac{\partial \phi_i}{\partial \sigma_{il}} \tag{18}$$

Finally, the results of the chain rules are written as follows:

$$\frac{\partial E}{\partial c_{il}} = A.\left\{2.\phi_1.(x_1 - c_{il})/(\sigma_{il})^2\right\} \tag{19}$$

$$\frac{\partial E}{\partial \sigma_{il}} = A.\left\{2.\phi_1.(x_1 - c_{il})^2/(\sigma_{il})^3\right\} \tag{20}$$

with $$A = \left(\sum_{j=1}^{m}(\hat{y}_j - d_j)\cdot(f_{ij} - \hat{y}_j)/b\right)\cdot$$

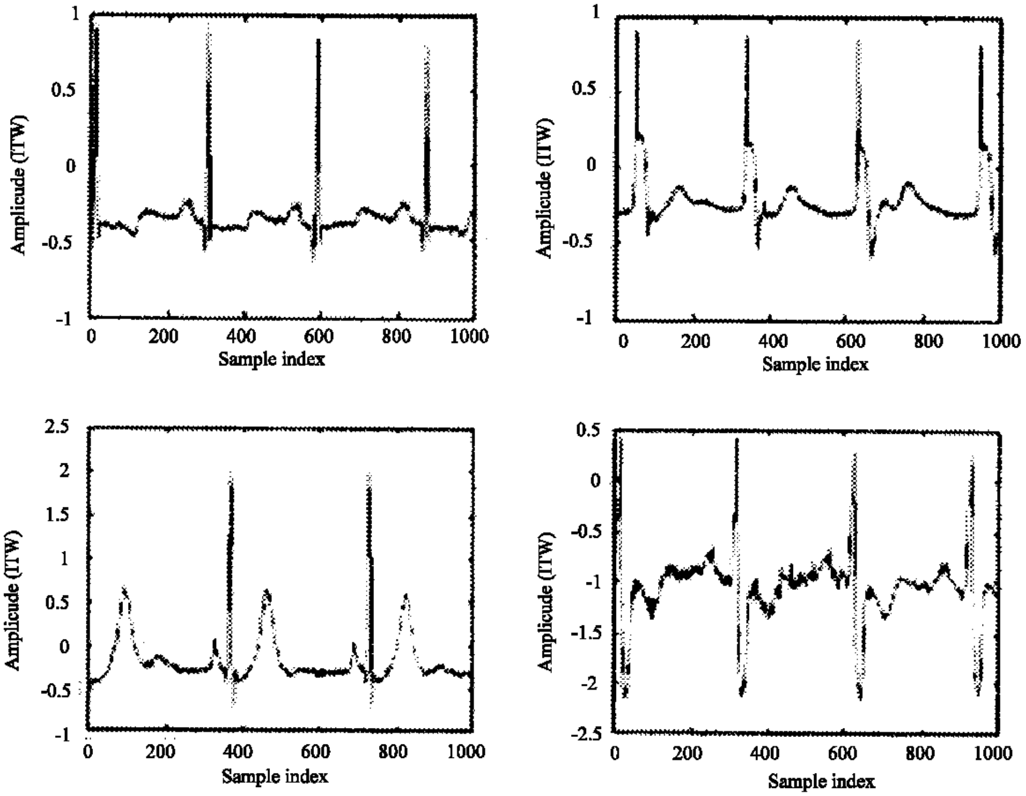

Fig. 2: ECG signals of four classes: (a) Normal sinus rhythm beats; (b) Non-conducted P-wave; © Premature ventricular contraction beats; (d) Right bundle brach block beats

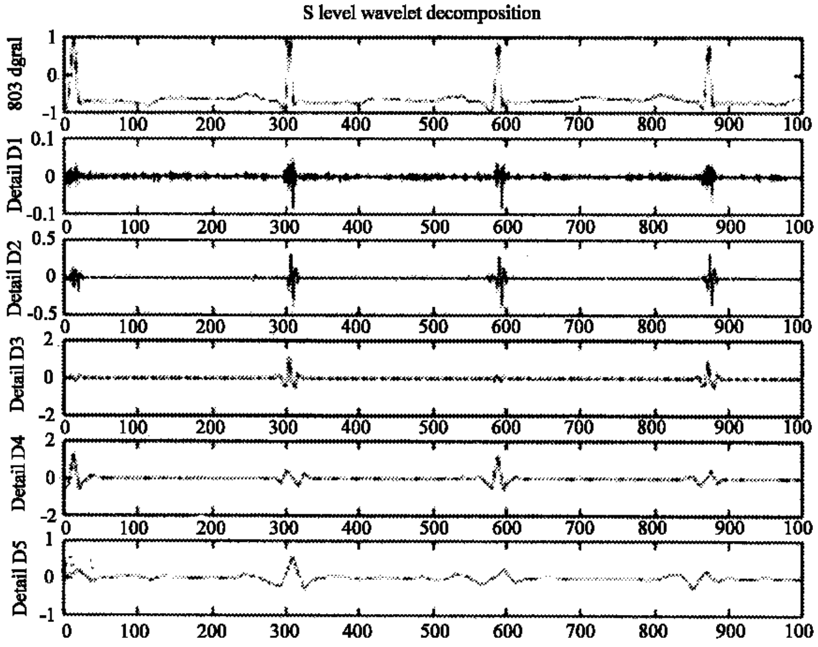

Fig. 3: Five level discrete wavelet decomposition of an ECG signal

Let us assume that the task is to find an optimal vector of parameters w which minimizes some objective function J(w) In the case of the neural system under consideration, all the parameters defining the ellipsoidal functions are stacked in a single vector w. The optimization algorithm is a variant of gradient descent in which each parameter w_j has its own step size and the step sizes are adapted during the optimization process, depending on the learning performance and more specifically on the progress of the objective function and on the sign of its derivatives at successive iterations. Let t be the iteration index. Then, if the objective function has decreased between iteration t-1 and t, the following rule is applied to update each step size η_j

$$\eta_j(t)=\begin{cases}\beta\eta_j(t-1),\text{if } \dfrac{\partial J}{\partial w_j}(t-1)\dfrac{\partial J}{\partial w_j}(t)>0\\ \gamma\eta_j(t-1),\text{ otherwise}\end{cases} \qquad (21)$$

where $\beta>0$ and $\gamma<1$ are two coefficients. Hence, the step size is increased if the derivatives have kept the same sign during two iterations and it is increased if the sign of the derivative has changed, because a jump over a minimum has occurred. The parameters are then updated by

$$w_j(t+1)=w_j(t)-\eta_j(t)\frac{\partial J}{\partial w_j}(t) \qquad (22)$$

If now the objective function has increased between iterations \$t-1\$ and \$t\$, all step sizes are decreased simultaneously

$$\begin{aligned}&\eta_j(t)=\delta\eta_j(t-1)\ \forall j\\ &w_j(t+1)=w_j(t-1)-\eta_j(t)\frac{\partial J}{\partial w_j}(t)\end{aligned} \qquad (23)$$

For the study under consideration, the following numerical empirical values of the coefficients are used:

$$\beta=1.2,\ \gamma=0.8,\delta=0.5 \qquad (24)$$

THE RESULTS OF NUMERICAL EXPERIMENTS

The MIT/BIH arrhythmia database[12] has been considered in the experiments because is widely accepted as a standard in the evaluation of methods for the automatic recognition of the ECG signals. The MIT-BIH ECG records are two-channels, 30 minutes duration, sampled at 360 samples per second and annotated by trained cardiologists. Four different types of ECG classes including Normal (N), Non-conducted P wave (P), Premature Ventricular Contraction (PVC) and Right bundle branch bock (R) beats are selected for this work from a subset of four ECG records (files numbered 100, 102, 106 and 118). An example for each beat type is shown in Fig. 2. Table 1 shows the records selected from the MIT-BIH arrhythmia database. Since most beats belong to the normal sinus rhythm and the number of some arrhythmia types is very small, we have deliberately limited the number of patients to provide approximate proportions of different arrhythmia cases taking part in experiments.

Table 1: MIH-BIH arrhythmia data base selected beats

Class	Record No.	Description
0	100	Normal Sinus rhythm beats
1	102	Non-Conducted P-wave
2	106	Premature Ventricular contraction beats
3	118	Right bundle branch block beats

Table 2: Rates of misclassification

Class	Learning	Test
0	1.5	2
1	0.5	1
2	1.5	1
3	1	2
Total	1.125	1.5

Table 3: Comparative results of ECG beat classifiers

	No. of beat types	Efficiency
MLP1	13	84.5
MLP2	12	92.0
MLP-Fourier	3	98.0
SOM-SVD	4	92.2
Proposed NEBF	4	98.5

The selected ECG beats being classified have been divided in two groups: one used for the learning purposes and the other for testing the performance of the classifier. The total number of beats used in learning is equal to 800 with 200 for each class. The testing was performed on 400 beats with 100 for each class. The feature vector is of size ten: 5 AR coefficients and the variances of detail signals of a 5 level DWT decomposition that corresponds to the 'Daubechies-2' wavelet. An example of decomposition is illustrated in Fig. 3. For the FCM algorithm, we have usedc = H = 8 clusters, which are equal to the number of ellipsoidal basis functions. Table 2 shows the average misclassification rates for the training and testing datasets. The average misclassification rate for both learning and test sets is very small and the recognition rate in the test mode is approximately 98.5%. In order to compare the obtained results with other techniques, Table 3 summarizes the comparative results between the following classifiers: Multistage systems using MLP (MLP)[11] Fourier and MLP MLP-Fourier[13]; Self-organizing maps and singular value decomposition (SOM-SVD)[12]. However, it is interesting to mention that patients and rhythms selected in all compared experiments were different. Hence fair comparison of the classifiers and their results is very difficult. Moreover, since different numbers of beat types have been used in the above methods, the second column of 2 gives the number of these beat types or classes and column 3 shows the overall recognition rate.

CONCLUSION

This study points out the ECG beat recognition in a systematic way using the NEBF network in order to improve the quality of the diagnosis. The construction of the network system is solved in two steps: The structure identification step and the parameter adaptation step. The learning algorithm uses a variable learning rate depending on the progress of the cost function. The recognition of the normal and different beats representing the arrhythmias has been done with good performance. In order to recognize other beat types, the integration of various features is under way. To fully automate the heartbeat recognition method presented here, an automatic heartbeat detection module is also required. This will be the object of our future work.

REFERENCES

1. Daubechies, I., 1998. Orthonormal bases of compactely wavelets, 1998. Commun. Pure Applied Math., 41: 909-996.
2. Hornik, K., M. Stinchcombe and H. White, 1989. Multilayer feedforward networks are universal approximators, Neural Networks, 2: 359-366.
3. Hartman, E.J., J.D. Keeler and J.M. Kowalski, 1990. Layered neural networks with Gaussian hidden units as universal approximators, Neural Comput., 2: 210-215.
4. Poggio, T. and F. Girosi, 1990. Networks for approximation and learning, Proceedings of the IEEE, 78: 481-1497.
5. Whitehead, B.A. and T.D. Choate, 1996. Cooperative-competitive genetic evolution of radial basis function centers and widths for time series prediction, IEEE Trans. on Neural Networks., 7: 869-880.
6. Karayiannis, N.B. and G.W. Mi, 1997. Growing Radial Basis Neural Networks: Merging Supervised and unsupervised Learning with network Growth Techniques, IEEE Trans. Neural Networks., 8: 1492-1506.
7. Yingwei, L., N. Sundararajan and P. Saratchandran, 1997. A sequential learning scheme for function approximation, Neural Computation., 9: 461-478.
8. Burdsall, B. and C. Giraud-Carrier, 1997. GA-RBF: A Self-Optimising RBF network, Proc. of the third international conference on artificial neural networks and genetic algorithms. Norwich, UK., pp: 346-349.
9. Gomm, J.B. and D. Yu, 2000. Selectiong Radial Basis Function Network Centers with Recursive Orthogonal Least Squares Training, IEEE Trans. on Neural Networks, 11: 306-314.
10. Chen, T. and H. Chen, 1995. Approximation capability to functions of several variables, nonlinear functions and operators by radial basis function neural networks, IEEE Trans. on Neural Networks, 6: 904-910.
11. Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function, Plenum Press, New York.
12. Mark, R. and G. Moody, 1988. MIT-BIH Arrhythmia Database Directory, Cambridge, MA: MIT.
13. Hu, Y.H., W. Tompkins, J.L. Urrusti and V.X. Alfonso, 1994. Applications of artificial neural networks for ECG signal detection and classification, J. Electrocardiology.
14. Hu, Y.H., S. Palreddy and W. Tompkins, 1997. A patient adaptable ECG beat classifier using a mixture of experts approach, IEEE Trans. Biomed. Eng., 44: 891-900.
15. Minami, K., H. Nakajima and T. Toyoshima, 1999. Real-time discrimination of ventricular tachyarrhythmia with Fourier transform neural network, IEEE Trans. Biomed. Eng., 46: 179-185.