Umut ORHAN

Cukurova University

Detection of Arrhythmia from ECG Signals by a Robust Approach to Outliers

Abstract. *The study focuses on arrhythmia detection from ECG signals, and for this aim it uses Fuzzy C-means (FCM) and Single Neuron Perceptron (SNP). FCM clustering adapted to time-series transforms ECG signals into useful features, and then SNP classifies them. We use MIT-BIH Arrhythmia database. The database is utilized for two experiments in the study. In the first experiment, RR intervals trimmed from the database are prepared for training the model, and in the second one ECG segments are used for real time simulation. Obtained results are compared with some other studies. According to the results, the proposed approach is good at arrhythmia detection as well as at least the studies in the literature. Lastly we interpret the results and present some studies for the future.*

Streszczenie. *W artykule skoncentrowano się na detekcji arytmii na podstawie sygnału ECG przy wykorzystaniu pojedynczego perceptronu i algorytmu FCM. Do badań wykorzystano bazę danych MIT-BIH Arrhythmia. W artykule oceniono zastosowaną metodę, przedstawiono interpretację wyników i dalsze propozycje. (**Detekcja arytmii na podstawie sygnału ECG przy wykorzystaniu sieci neuronowych**)*

Keywords: *Electrocardiography, arrhythmia, fuzzy c-means, single neuron perceptron, classification, real-time detection.*
Słowa kluczowe: elektrokardiogram, arytmia, sieci neuronowe

I. Introduction

Cardiovascular diseases are the most important problems all around the world that threaten the human body. In the detection of diseases about cardiovascular system, electrocardiogram (ECG) is the most common source. Especially in coronary intensive care services, the ECG recordings of the patients are inspected for long duration in real time uninterruptedly. Certainly the evaluation by the expert physicians is necessary to transform ECG signals into useful information. But a skilled doctor may not be everywhere, in every time. Therefore a computer based detection system would be very useful for humanity. Nowadays partly, some computer aided systems are in use, and they serve important services. Many researchers have still studied on the computer based detection of all diseases able to be diagnosed from ECG.

In fact, ECG signals are in the form of the combination of six main waves commonly recognized in the medical literature. A simple RR interval includes all types of these waves at once, and is accepted as unit element of a long-term ECG signal. In the detection of cardiac illnesses, the most useful features should be extracted from RR intervals, and then these features must be interpreted for making a decision. By the aim of disease detection, there are lots of algorithms proposed for feature extraction from ECG signals. ECG signals are usually seen as non-stationary signals, and thus they are analyzed in time-frequency domain [1-5]. But some sources tell the ECG signals partially stationary, since ECG signals are comprehended as the repetition of RR intervals, and for this reason they can be examined in frequency domain [6-11]. The transporting a signal into frequency or time-frequency domain in a real time system means high computational time and cost. The computational time in real time systems is an important detail for both feature extraction and classification steps. Therefore the chosen methods for designed system should not only be preferred on account of its high classification success, but also they must include the features like having less computation and making fast decision. Because of its properties of low computational cost, it is a more effective approach to try to improve only classification success in time domain.

The detection of R peak of an ECG segment is important for dividing a long-term ECG signal into its RR intervals. In this matter, the most used method is based on the centering the maximum amplitude in a window with enough length [12]. But sometimes because of unusual waves in the signal, this approach can not catch the real R peak value, and thus it causes many mistakes in other steps coming after the preparing RR intervals. Although there are many algorithms to detect R peak, this topic has still maintained its importance. On the other hand, if an undesired RR interval due to incorrect detected R peak is injected the training of decision making system, the system can become ill, and it may produce wrong results. For this reason, the designing a system having strong immunity against to the little infection like outlier samples, artifacts caused by muscle acts and electrical power line is a necessity.

Many researchers have focused on the diseases able to be detected from ECG signals, and one of the diseases is arrhythmia. There are several causes of ECG arrhythmia, and all of them are easily detectable by help of some characteristics of RR intervals. The ECG signals with arrhythmia can primitively be defined as ECG signals with some spoiled RR intervals. The most of the studies on the detection of arrhythmia in literature focused on classification of RR intervals with arrhythmia determined from some ECG records [1-2, 12-16]. Since those studies used RR intervals before classified as arrhythmia or normal, they have some disadvantages on account of their naïve approach to the matter. For example these two ECG segments may not similar to each other: a RR interval from ECG record of a completely healthy subject and another RR interval from an arrhythmia patient but classified as normal. Consequently, by help of the computer aided systems, the comparison of normal RR intervals from arrhythmia patients and RR intervals from healthy subjects may reveal some hidden characteristics of arrhythmia disease.

This study focuses on an algorithmic approach to real time detection of cardiac arrhythmia being easy to diagnose for expert physicians. Although the method used in the study for trimming RR intervals might choose irregular RR intervals due to determining R peaks incorrectly, the proposed system can reach very high successes by means of its robust to outliers. The preparation of ECG data used in the study described in second section. These signals were transformed into numerical features, and then the extracted features were classified by fuzzy C-means (FCM) method and a single neuron perceptron (SNP) defined in third and fourth section respectively. In fifth section, the implemented experiments in the study were presented, and lastly conclusions were drawn.

II. Dataset Used

The ECG signals used as data for the most of studies focused cardiac disease detection need to prepare before by experts. Thus we use MIT-BIH Arrhythmia database from Physiobank web page [17]. Arrhythmia database composed of 48 long-term ECG records taken randomly from a big database with recordings of 4000 subjects. In this database, 23 records of all were chosen from ambulatory ECG recordings and other 25 records from clinically significant arrhythmias.

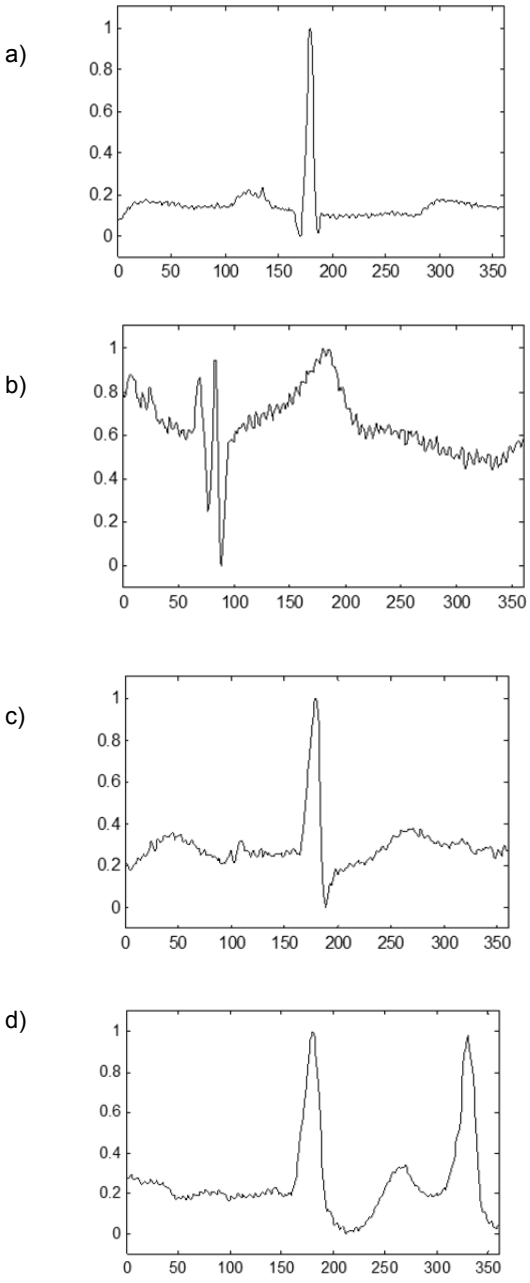

Fig.1. RR samples trimmed from database: (a)normal, (b)incorrect normal, (c)arrhythmia, (d)incorrect arrhythmia.

The database is utilized for two experiments in the study. In the first experiment, RR intervals with 1s length chosen from the ECG database are prepared for training the methods, and ECG segments with 10s length are used for real time simulation in the second one. All segments are selected from random positions. Every RR intervals of ECG records in database are classified by expert doctors, and that class information is written as the label of related RR interval. But these labels were not utilized for datasets used in the study. Instead of that, all RR intervals from healthy subjects and the other RR intervals (from arrhythmia patients) are labeled as “normal” and “arrhythmia” respectively. Therefore it can be also tested whether or not RR intervals from arrhythmia patients hide some secrets.

During the test process, long-term ECG segments are scanned by a window with 1s length to make the application look like a real time simulation. If peak value of short (1s) segment in the window is located exactly in the middle of the window, this short segment is assumed as a RR interval and added into the dataset to be used in experiments. Because of this primitive technique used to detect the peak (R) value, the trimmed RR intervals for dataset may not always be correct and thus it can cause bad classification. Four different RR intervals chosen as the samples are shown in the following figure.

Since some R waves can’t detect correctly as seen in Fig.1, the system tries to run with wrong RR intervals for both normal and arrhythmia. In addition every RR intervals are normalized into [0 1] for focusing on pattern recognition in the study. Therefore the studying on the patterns independent of amplitude axis could be possible.

III. Feature Extraction by Fuzzy C-means

In the second step of the study, RR intervals in the prepared dataset are transformed into useful features by Fuzzy C-means (FCM) clustering method reorganized to be applicable to time series signals. The feature extraction method represents the signals with fuzzy membership values according to the clusters found by FCM. Consequently each signal is symbolized the features as much as the number of clusters given to FCM. The describing a signal with the less features probably decreases system complexity and also computational time. But achieving high classification accuracy would be very hard by such a system. As important as the number of separation on amplitude axis, the form of separations must also be correctly determined for successful result. When the clustering process finishes, the membership values of RR intervals would have been already determined by FCM method. In the matter of separating places on amplitude axis, FCM is a quite skillful method. The most appropriate number of clusters can only be decided by try-and-error method.

FCM method, as a short description, is a fuzzificated form of K-means which is the most popular clustering method. As a difference from K-means algorithm in which each data point belongs only to one cluster, each data point is member of every cluster with different ratio in FCM. Although the cost functions of two algorithms look generally similar, they have also important differences. The cost function of the FCM algorithm is as follows [18].

$$(1) \qquad J = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} \left\| x_j - c_i \right\|$$

where: c_i - ith cluster centers, x_j - jth data points, and u_{ij} - memberhip value of jth data point to ith cluster.

To be able to minimize cost function (J), FCM algorithm uses the equations below.

$$(2) \qquad c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}$$

FCM algorithm as in K-means also starts with random cluster centers, and then it updates the coordinates of cluster centers and membership values. When computed cost or total iteration number reaches a specific value, the algorithm is stopped. Although there are many researches about using clustering algorithms in transforming a

continuous signal into discrete one in the literature, the clustering a non-continuous signal to extract practical features like in this study is not a common approach yet. For this aim, FCM algorithm is adapted to the matter by some special viewpoints. The process of transforming RR intervals into the features has follow steps:

- Chose RR intervals with 1s so that the peak value in the window becomes right in the middle.
- Normalize each RR interval, thus the amplitudes of all R waves is 1.
- Assume every RR interval composed of two part (before and after R wave), and so divide each RR interval into two segments. This also causes that the dataset is divided into two.
- The values of the signals in these new two datasets are regarded as independent numbers, and each dataset is thought as a long number array.
- The numbers in these two number arrays are sorted from small to large.
- For each number array, random cluster centers, as much as given c number, are chosen, and then these centers are optimized by FCM.
- By using the membership values computed by FCM, the fuzzy membership of each RR interval to each cluster center is calculated.

As a result, each RR interval can be represented by 2c number of features. While the some of these features may be more useful, some others may damage the classification accuracy. But determination of both optimal c value and selection of useful features are the procedures required a long computational time. Because of concerns about the computational time, all of c features are used in classification process by skipping the feature selection step. Optimal c value is determined by trial and error.

IV. Classification by Single Neuron Perceptron

Classification is a such procedure that it provides the capability to interpret a new sample by help of data of a problem in which its cause and result known. In this study, Single Neuron Perceptron (SNP) known by its success in several areas is used as a classifcation method [19]. This model has only one neuron with non-linear activation function. A general appearance of SNP model used in the study is shown in Figure 2.

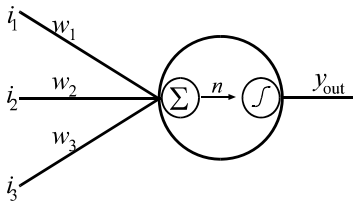

Fig. 2. SNP model used in the study.

The neuron shown in Figure 2 has three inputs and one output. Each dimension of data (i_1, i_2, i_3) reaches the neuron by multiplying its weight (w_1, w_2, w_3), and then computed result is sent to output (y_{out}). The neuron includes two computational units. The first of them is linear unit which calculates the summation of weighted inputs, and it sends the summation result to the second unit. The second unit is the unit of the neuron that contains non-linear activation function. The summation from the first unit turns into activated result, and this value is the final prediction produced by the SNP model. The output (y_{out}) of SNP is computed by following equation.

$$y_{out} = f(\sum_k i_k w_k) \tag{3}$$

where: f – preferred activation function.

Although there are many activation functions in the literature, the most popular activation functions are sigmoid and hyperbolic tangent. An activation function to be more successful than other one depends entirely problem.

As a training algorithm in SNP, the derivations of back-propagation algorithm are generally used. The weights of inputs are determined optimally by means of training algorithm. If the variable y is the output given in data and y^* is the prediction value computed by SNP, then the weights are updated by distributing the error of the model in each iteration of training procedure. Gradient descent method, known by its convergence concept, is mostly used for the updating process. Following equation provides to update the weights of inputs of a SNP model by using back-propagation algorithm.

$$\Delta w = -\eta \frac{\partial\left(\frac{1}{2}(y^*-y)^2\right)}{\partial w} \tag{4}$$

In fact, there are three simple equations here as in nested. At first, the deviation amount of the output of SNP model is calculated after a feed-forward. Then the error energy is computed, and lastly the distribution ratio of the error energy into each weight is determined. This updating amount (Δw) is decreased by a learning rate (η) parameter. This procedure is repeated for each data point, thus the average of all solutions would be counted up by help of η parameter. For SNP model used in the study, sigmoid and back-propagation are preferred as activation function and training algorithm respectively. The application is implemented in Matlab environment, and other parameters of SNP model are tuned by the software tool utilized.

V. Experimental Work

Since the study aims arrhythmia detection, arrhythmia patterns in real ECG signals are analyzed, and RR intervals of both normal subjects and arrhythmia patients are compared. Almost all of the researches related to this matter in the literature focuses on the classification with RR intervals labeled before. This approximation to the matter has some drawbacks. For example, a RR interval from a full healthy human and a healthy RR interval from an arrhythmia patient may not be the similar. They can even include entirely different features. The technique to detect this difference is the proposed system. For this reason, the study concentrate, instead of the recognition of types of arrhythmia, the availability of arrhythmia pattern by comparing RR intervals from both healthy subjects and arrhythmia patients.

On the other hand, to generalize the success in classification experiments implemented on restricted data sets, the detection model must be supplied by a cross validation method. The main principle of cross validation methods is based on division of the data into two parts as train and test. A validation method basically aims that after training the classification method by a part of data, it is tested by another part of data. Thus the success of the classifier is measured objectively. In this study, we use leave-one-out method for the validation. Leave-one-out splits the data with N samples into train set with N-1 samples and test set with only 1 sample, and it repeats this procedure N times by changing the test sample in each repetition. Since each test set includes only 1 sample, the computed accuracies would be either 0% or 100%. Then the accuracies computed on test sets are averaged for the final evaluation. Although in case the data set comprises of too many samples, leave-one-out method is not useful in point of computational time, it is the most reliable cross validation method. Besides it makes easy to analyze outliers in the data.

For the first experiment implemented in the study, a data set is prepared by trimming 10 RR intervals from each long-term recording in arrhythmia database. Therefore the data set is comprises of 480 RR intervals. To get rid of amplitude values, all RR intervals are normalized. Assuming each RR interval as an array variable, all of them are combined at a large array. Then Fuzzy C-means (FCM) clusters the large array for different cluster numbers in [2 99]. During the clustering, some membership values are found for every number in the array. By averaging the memberships of every number belonging to each RR interval, membership vector of RR intervals are calculated. Transforming two RR intervals into membership vectors is shown in Figure 3.

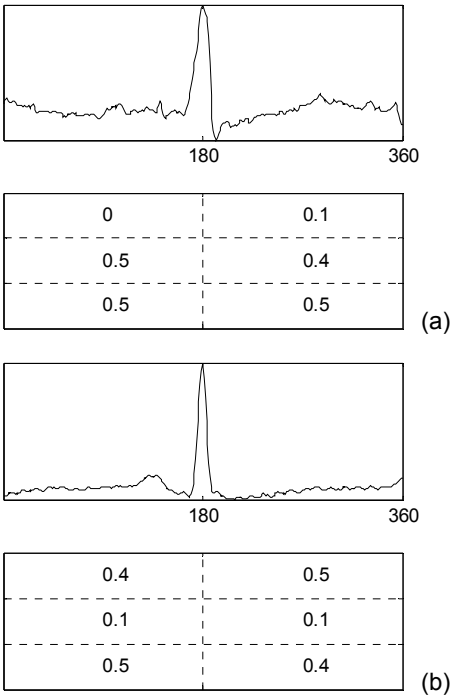

Fig. 3. Transforming a normal and an arrhythmia RR intervals into membership vectors for *c*=3.

As seen in Fig.3, the dissimilarity between two membership vectors is so obvious that the vectors can be classified easily by any prediction method. This procedure is applied to all RR intervals to extract feature vectors, and then the obtained vectors are used to train a Single Neuron Perceptron (SNP) classifier. According to the validation method mentioned above, the best accuracy determined is compared with some researches on the detection of arrhythmia from ECG signal. The comparison is presented in Table I.

Table I. The comparison of the proposed approach with some researches in literature.

Studies	Success
Osowski and Linh [20]	95.91
Engin [12]	98.00
Dokur and Ölmez [21]	96.70
This study	98.13

As in Table I, it can be told that the approach proposed in the study has an acceptable success. Additionally, this trained system with RR intervals is also tried in a real time arrhythmia detection simulation. For this aim, a new data set comprised of 10s-long ECG segments is prepared at random from arrhythmia database. Each segment in this data set is regarded as if it is measured in real time. Then a window with 1s-long is defined in order to recognize RR intervals on a real time ECG signal. When ECG signal passes through this window, the maximum value in the window is determined as the peak value of R wave. If the peak value is just in the middle position of the window, the ECG segment in the window is regarded as a RR interval, and it is entered the detection system. During 10s-long, the detection system recognizes many RR intervals, and lastly it makes a final decision according to the majority. If the majority of the recognized RR intervals is detected as arrhythmia, then the system says 10s-long signal being diseased. The decision values obtained for each segment in the data set are presented in Table II.

Table II. The results of the proposed model in real time simulation performed.

Classes	Normal	Arrhythmia
The Number of Segments	23	25
The Number of RRs	239	267
Correct Classified Segments	22	25
Correct Classified RRs	179	211

According to Table II, the accuracy is 77.08% in point of the classification of RR intervals and the general accuracy is 97.92% with regard to ECG segments. The classification accuracy of RR intervals seems less than the result of the first experiment. The reason is the primitive method used to choose RR intervals in 1s-long window. Because of the primitive methods determined the R wave, incorrect segments are regarded as RR intervals, and thus the final success of the model may be less than the expected value. Nevertheless, the proposed model is very successful in point of the segments.

VI. Conclusions

In the study, we focused on arrhythmia detection from ECG signals, and for this aim Fuzzy C-means (FCM) and Single Neuron Perceptron (SNP) methods were used. Thanks to FCM clustering adapted to time-series, ECG signals represented by the powerful parameters were easily classified with SNP. The model trained by RR intervals was also used in a real time simulation of arrhythmia detection from long-term ECG signals. Although we are sure in point of applicability of the model onto all kind on time series signals, acceptability of the model depends on its generalized successes. For this reason, it needs more test in several signals. Nevertheless, it can be said that the proposed approach is good at arrhythmia detection as well as at least the studies in the literature. On the other hand, despite a primitive R detection method, such a high classification accuracy whispers in our ears that RR intervals of a healthy and an arrhythmia patient can sometimes seem similar to each other. In order to reach a definitive judgment, some special tests focused on this issue must be performed. Algorithms covered by the approach proposed in this paper are like the meteors selected from a space containing many different methods. Without disturbing the main idea of the proposed approach; a more accurate windowing method for R determination, feature extraction methods alternative to FCM, and different classifiers should be researched. Therefore the novel methods can be adapted to proposed approach to reach higher accuracies. Finally, in determining all the problems able to be understood from ECG signals and other biomedical signals; how the method to be useful must be examined.

REFERENCES

[1] M. R. Homaeinezhad, S. A. Atyabi, E. Tavakkoli, H. N. Toosi, A. Ghaffari, R. Ebrahimpour, ECG arrhythmia recognition via a neuro-SVM–KNN hybrid classifier with virtual QRS image-based geometrical features, Expert Systems with Applications, 39 (2012) 2047–2058.

[2] C. P. Shen, W.C. Kao, Y. Y. Yang, M. C. Hsu, Y. T. Wuc, F. Lai, Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines, Expert Systems with Applications, 39 (2012) 7845–7852.
[3] A. Mousa, R. Saleem, Using Reduced Interference Distribution to Analyze Abnormal Cardiac Signal, Journal of Electrical Engineering, 62(3) (2011) 168–172.
[4] C. H. Lin, Y. C. Du, T. Chen, Adaptive wavelet network for multiple cardiac arrhythmias recognition, Expert Systems with Applications, 34 (2008) 2601–2611.
[5] J. C. Wood, D. T. Barry, Time-frequency analysis of skeletal muscle and cardiac vibrations, Proceedings of the IEE, 84 (9) (1996), 1281–1294.
[6] C. H. Lin, Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier, Computers and Mathematics with Applications, 55 (2008) 680–690.
[7] S. Kar, M. Okandan, Atrial fibrillation classification with artificial neural networks. Pattern Recognition, 40 (2007) 2967-2973.
[8] I. Christov, G. Gomez-Herrero, V. Krasteva, I. Jekova, A. Gotchev, K. Egiazarian, Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification, Medical Engineering and Physics, 28 (2006) 876–887.
[9] M. Stridh, L. Sörnmo, C. J. Meurling, S. B. Olsson, Sequential characterization of atrial tachyarrhythmias based on ECG time-frequency analysis, IEEE Transactions on Biomedical Engineering, 51 (1) (2004).
[10] D. Benitez, P. A. Gaydecki, A. Zaidi, A. P. Fitzpatrick, The use of the Hilbert transform in ECG signal analysis, Computers in Biology and Medicine, 31 (2001) 399–406.
[11] R. H. Clayton, A. Murray, Estimation of the ECG signal spectrum during ventricular fibrillation using the fast Fourier transform and maximum entropy methods, Proceedings of the Computers in Cardiology, (1993) 867–870.
[12] M. Engin, ECG beat classification using neuro-fuzzy network, Pattern Recognition Letters 25 (2004) 1715–1722.
[13] M. Llamedo, A. Khawaja, J. P. Martinez, Cross-Database Evaluation of a Multilead Heartbeat Classifier, IEEE Transactions on Information Technology in Biomedicine, 16 (4) (2012) 658-664.
[14] F. Yaghouby, A. Ayatollahi, R. Bahramali, M. Yaghouby, Robust genetic programming-based detection of atrial fibrillation using RR intervals, Expert Systems, 29 (2) (2012) 183-199.
[15] L. Hong-wei, S. Ying, L. Min, L. Pi-ding, Z. Zheng, A probability density function method for detecting atrial fibrillation using R–R intervals, Medical Engineering & Physics, 31 (2009) 116–123.
[16] M. G. Tsipouras, D. I. Fotiadis, Automatic arrhythmia detection based on time and time-frequency analysis of heart rate variability, Computer Methods and Programs in Biomedicine, 74 (2004), 95-108.
[17] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation, 101 (23) (2000) e215–e220.
[18] J. C. Bezdek, R. Ehrlich, W. Full, FCM: The Fuzzy C-means Clustering Algorithm, Computers & Geosciences, 10 (2-3)(1984) 191-203.
[19] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, 65 (1958) 386–408.
[20] S. Osowski, T. H. Linh, ECG beat recognition using fuzzy hybrid neural network, IEEE Transactions on Biomedical Engineering, 48 (2001) 1265–1271.
[21] Z. Dokur, T. Olmez, ECG beat classification by a hybrid neural network, Computer Methods and Programs in Biomedicine, 66 (2001) 167–181.

Author:
Asst. Prof. Dr. Umut Orhan, Cukurova University, Computer Engineering Department, 01330 Adana, Turkey, umutorhan@hotmail.com