Available online at www.sciencedirect.com
ScienceDirect
Digital Signal Processing 17 (2007) 702–710
Digital Signal Processing
www.elsevier.com/locate/dsp

An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease

Kemal Polat *, Salih Güneş

Electrical and Electronics Engineering Department, Selcuk University, 42035 Konya, Turkey

Available online 20 October 2006

Abstract

Diabetes occurs when a body is unable to produce or respond properly to insulin which is needed to regulate glucose (sugar). Besides contributing to heart disease, diabetes also increases the risks of developing kidney disease, blindness, nerve damage, and blood vessel damage. In this paper, we have detected on diabetes disease, which is a very common and important disease using principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS). The aim of this study is to improve the diagnostic accuracy of diabetes disease combining PCA and ANFIS. The proposed system has two stages. In the first stage, dimension of diabetes disease dataset that has 8 features is reduced to 4 features using principal component analysis. In the second stage, diagnosis of diabetes disease is conducted via adaptive neuro-fuzzy inference system classifier. We took the diabetes disease dataset used in our study from the UCI (from Department of Information and Computer Science, University of California) Machine Learning Database. The obtained classification accuracy of our system was 89.47% and it was very promising with regard to the other classification applications in literature for this problem.

Keywords: PCA; ANFIS; Diabetes disease; Expert system; Medical diagnosis

1. Introduction

While diabetes is not actually a form of heart disease, it often contributes to heart disease. Diabetes occurs when a body is unable to produce or respond properly to insulin which is needed to regulate glucose (sugar). Besides contributing to heart disease, diabetes also increases the risks of developing kidney disease, blindness, nerve damage, and blood vessel damage. More than 80 percent of people with diabetes die from some form of heart or blood vessel disease. There are two forms of diabetes: juvenile diabetes and adult-onset diabetes. Adult-onset diabetes is associated with obesity and can be delayed or controlled with proper diet and exercise. There is no cure for diabetes. However, diabetes can be controlled through changing eating habits and exercise programs. Drugs are also available. However, even if diabetes is under control, it still contributes to heart disease [1].

* Corresponding author. Fax: +90 332 241 0635.
E-mail address: kpolat@selcuk.edu.tr (K. Polat).

1051-2004/$ – see front matter
doi:10.1016/j.dsp.2006.09.005

One of the central problems of the information age is dealing with the enormous amount of raw information that is available. More and more data is being collected and stored in databases or spreadsheets. As the volume increases, the gap between generating and collecting the data and actually being able to understand it is widening. In order to bridge this knowledge gap, a variety of techniques known as data mining or knowledge discovery is being developed. Knowledge discovery can be defined as the extraction of implicit, previously unknown, and potentially useful information from real world data, and communicating the discovered knowledge to people in an understandable way [2–4].

The use of classifier systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. But, expert systems and different artificial intelligence techniques for classification also help experts in a great deal.

In this study, we have proposed the expert system that has two stages. First, dimension of diabetes disease dataset that has 8 features is reduced to 4 features using principal component analysis. Then, we used adaptive neuro-fuzzy inference system classifier diagnosis diabetes disease. The obtained classification accuracy of our system was 89.47% and it was very promising with regard to the other classification applications in literature for this problem.

The rest of the paper is organized as follows. Section 2 gives the material and method. In each subsection of this section, the detailed information was given. The results obtained in applications are given in Section 3. This section also includes the discussion of these results in specific and general manner. Consequently in Section 4, we conclude the paper with summarization of results by emphasizing the importance of this study and mentioning about some future work.

2. Material and method

2.1. Used diabetes disease dataset

This data set was obtained from Ref. [5]. The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. The binary response variable takes the values '0' or '1,' where '1' means a positive test for diabetes and '0' is a negative test for diabetes. There are 268 (34.9%) cases in class '1' and 500 (65.1%) cases in class '0.' There are eight clinical findings: (1) Number of times pregnant. (2) Plasma glucose concentration a 2 h in an oral glucose tolerance test. (3) Diastolic blood pressure (mm Hg). (4) Triceps skin fold thickness (mm). (5) 2-h serum insulin (mu U/ml). (6) Body mass index. (7) Diabetes pedigree function. (8) Age (years). A brief statistical analyze is given in Table 1 [6].

2.2. Proposed expert system

We have proposed expert system, which has two stages. In the first stage, dimension of diabetes disease dataset that has 8 features is reduced to 4 features using principal component analysis. In the second stage, we used adaptive neuro-fuzzy inference system classifier to diagnosis of diabetes disease. The block diagram of proposed system is shown in Fig. 1.

Table 1
Brief statistical analyze of diabetes disease dataset

Attribute number	Mean	Standard deviation	Min/max
1	3.8	3.4	0/17
2	120.9	32.0	0/199
3	69.1	19.4	0/122
4	20.5	16.0	0/99
5	79.8	115.2	0/846
6	32.0	7.9	0/67.1
7	0.5	0.3	0.078/2.42
8	33.2	11.8	21/81

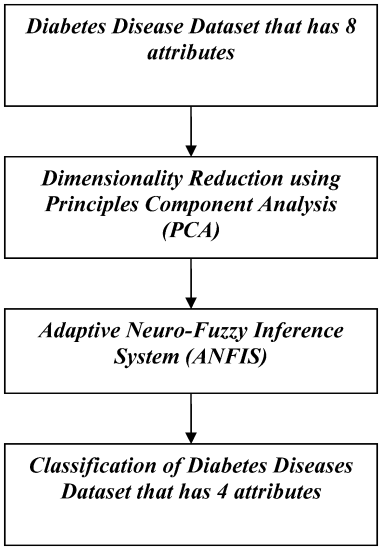

Fig. 1. The block diagram of proposed system.

2.2.1. Principal component analysis

PCA was used to make a classifier system more effective. For this aim, before classifying, PCA method was used for dimensionality reduction of diabetes disease dataset. Therefore, diabetes disease dataset was represented a vector consists of 8 attributes. PCA is based on the assumption that most information about classes is contained in the directions along which the variations are the largest. The most common derivation of PCA is in terms of a standardized linear projection, which maximizes the variance in the projected space [7]. For a given p-dimensional data set X, the m principal axes $T_1, T_2, \ldots, T_m$, where $1 \leqslant m \leqslant p$, are orthonormal axes onto which the retained variance is maximum in the projected space. Generally, $T_1, T_2, \ldots, T_m$ can be given by the m leading eigenvectors of the sample covariance matrix $S = (1/N)\sum_{i=1}^{N}(x_i - \mu)^T(x_i - \mu)$, where $x_i \in X$, μ is the sample mean, and N is the number of samples, so that

$$ST_i = \lambda_i T_i, \quad i \in 1, \ldots, m, \tag{1}$$

where λ_i is the ith largest eigenvalue of S. The m principal components of a given observation vector $x_i \in X$ are given by

$$y = [y_1, y_2, \ldots, y_m] = \left[T_1^T x, T_2^T, \ldots, T_m^T\right] = T^T x. \tag{2}$$

The m principal components of x are decorrelated in the projected space. In multi-class problems, the variations of data are determined on a global basis, that is, the principal axes are derived from a global covariance matrix:

$$\hat{S} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}\left(x_j - \hat{\mu}\right)\left(x_j - \hat{\mu}\right)^T, \tag{3}$$

where $\hat{\mu}$ is the global mean of all the samples, K is the number of classes, N_j is the number of samples in class j; $N = \sum_{j=1}^{K} N_j$ and x_{ji} represents the ith observation from class j. The principal axes $T_1, T_2, \ldots, T_m$ are therefore the m leading eigenvectors of $\hat{S}$:

$$\hat{S}T_i = \hat{\lambda}_i T_i, \quad i \in 1, \ldots, m, \tag{4}$$

where $\hat{\lambda}_i$ is the ith largest eigenvalue of $\hat{S}$. An assumption made for feature extraction and dimensionality reduction by PCA is that most information of the observation vectors is contained in the subspace spanned by the first m principal axes, where $m < p$. Therefore, each original data vector can be represented by its principal component vector with dimensionality m [8].

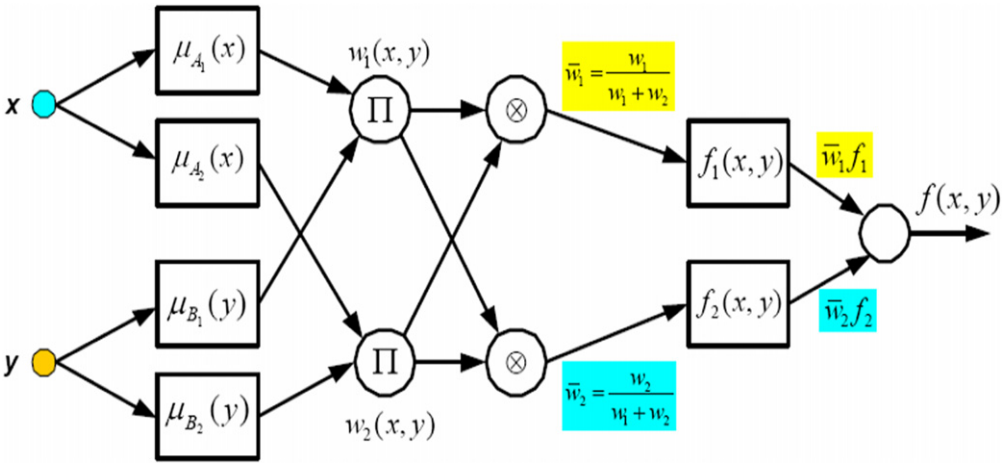

Fig. 2. ANFIS architecture.

2.2.2. Adaptive neuro-fuzzy inference system

The ANFIS is a fuzzy Sugeno model put in the framework of adaptive systems to facilitate learning and adaptation [9,10]. Such framework makes the ANFIS modeling more systematic and less reliant on expert knowledge. To present the ANFIS architecture, two fuzzy if–then rules based on a first order Sugeno model are considered:

Rule 1. If (x is A_1) and (y is B_1) then ($f_1 = p_1 x + q_1 y + r_1$);
Rule 2. If (x is A_2) and (y is B_2) then ($f_2 = p_2 x + q_2 y + r_2$),

where x and y are the inputs, A_i and B_i are the fuzzy sets, f_i are the outputs within the fuzzy region specified by the fuzzy rule, p_i, q_i, and r_i are the design parameters that are determined during the training process. The ANFIS architecture to implement these two rules is shown in Fig. 2, in which a circle indicates a fixed node, whereas a square indicates an adaptive node.

In the first layer of ANFIS, all the nodes are adaptive nodes. The outputs of layer 1 are the fuzzy membership grade of the inputs, which are given by

$$O_i^1 = \mu_{A_i}(x), \quad i = 1, 2, \tag{5}$$

$$O_i^1 = \mu_{B_{i-2}}(y), \quad i = 3, 4, \tag{6}$$

where μ is an obtained weight according to related fuzzy membership function, $\mu_{A_i}(x)$, $\mu_{B_{i-2}}(y)$ can adopt any fuzzy membership function. For example, if the bell shaped membership function is employed; $\mu_{A_i}(x)$ is given by

$$\mu_{A_i}(x) = \frac{1}{1 + \{((x - c_i)/a_i)^2\}^{b_i}}, \tag{7}$$

where a_i, b_i, and c_i are the parameters of the membership function, governing the bell shaped functions accordingly.

In the second layer, the nodes are fixed nodes. They are labeled with M, indicating that they perform as a simple multiplier. The outputs of this layer can be represented as

$$O_i^2 = w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1, 2, \tag{8}$$

which are the so-called firing strengths of the rules.

In the third layer, the nodes are also fixed nodes. They are labeled with N, indicating that they play a normalization role to the firing strengths from the previous layer.

The outputs of this layer can be represented as

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2, \tag{9}$$

which are the so-called normalized firing strengths.

In the fourth layer, the nodes are adaptive nodes. The output of each node in this layer is simply the product of the normalized firing strength and a first order polynomial (for a first order Sugeno model). Thus, the outputs of this layer are given by

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i), \quad i = 1, 2. \tag{10}$$

In the fifth layer, there is only one single fixed node labeled with S. This node performs the summation of all incoming signals. Hence, the overall output of the model is given by

$$O_i^5 = \sum_{i=1}^{2} \bar{w}_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{w_1 + w_2}. \tag{11}$$

It can be observed that there are two adaptive layers in this ANFIS architecture, namely the first layer and the fourth layer. In the first layer, there are three modifiable parameters $\{a_i, b_i, c_i\}$, which are related to the input membership functions. These parameters are the so-called premise parameters. In the fourth layer, there are also three modifiable parameters $\{p_i, q_i, r_i\}$, pertaining to the first order polynomial. These parameters are so-called consequent parameters [9,10].

The task of the learning algorithm for this architecture is to tune all the modifiable parameters, namely $\{a_i, b_i, c_i\}$ and $\{p_i, q_i, r_i\}$, to make the ANFIS output match the training data. When the premise parameters a_i, b_i, and c_i of the membership function are fixed, the output of the ANFIS model can be written as

$$f = \frac{w_1}{w_1 + w_2} f_1 + \frac{w_2}{w_1 + w_2} f_2. \tag{12}$$

Substituting Eq. (9) into Eq. (12) yields

$$f = \bar{w}_1 f_1 + \bar{w}_2 f_2. \tag{13}$$

Substituting the fuzzy if–then rules into Eq. (14), it becomes

$$f = \bar{w}_1 (p_1 x + q_1 y + r_1) + \bar{w}_2 (p_2 x + q_2 y + r_2). \tag{14}$$

After rearrangement, the output can be expressed as

$$f = (\bar{w}_1 x) p_1 + (\bar{w}_1 y) q_1 + (\bar{w}_1) r_1 + (\bar{w}_2 x) p_2 + (\bar{w}_2 y) q_2 + (\bar{w}_2) r_2, \tag{15}$$

which is a linear combination of the modifiable consequent parameters p_1, q_1, r_1, p_2, q_2, and r_2. The least-squares method can be used to identify the optimal values of these parameters easily. When the premise parameters are not fixed, the search space becomes larger and the convergence of the training becomes slower. A hybrid algorithm combining the least-squares method and the gradient descent method is adopted to solve this problem. The hybrid algorithm is composed of a forward pass and a backward pass. The least-squares method (forward pass) is used to optimize the consequent parameters with the premise parameters fixed. Once the optimal consequent parameters are found, the backward pass starts immediately. The gradient descent method (backward pass) is used to adjust optimally the premise parameters corresponding to the fuzzy sets in the input domain. The output of the ANFIS is calculated by employing the consequent parameters found in the forward pass. The output error is used to adapt the premise parameters by means of a standard back propagation algorithm. It has been proven that this hybrid algorithm is highly efficient in training the ANFIS [9,10].

2.3. Performance evaluation of proposed expert system

2.3.1. Measuring error

Given a random set of initial weights, the outputs of the network will be very different from the desired classifications. As the network is trained, the weights of the system are continually adjusted to reduce the difference between the output of the system and the desired response. The difference is referred to as the error and can be measured in different ways. The most common measurement is the MSE. The MSE is the average of the squares of the difference between each output and the desired output [10].

Table 2
Representation of confusion matrix

Actual	Predicted	
	Negative	Positive
Negative	*a*	*b*
Positive	*c*	*d*

2.3.2. Classification accuracy

In this study, the classification accuracies for the datasets were measured according to the equation

$$\begin{aligned} &\text{accuracy}(T) = \frac{\sum_{i=1}^{|T|} \text{assess}(t_i)}{|T|}, \quad t_i \in T, \\ &\text{assess}(t) = \begin{cases} 1, & \text{if classify}(t) = t.c, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{16}$$

where T is the set of data items to be classified (the test set), $t \in T$, $t.c$ is the class of the item t, and classify(t) returns the classification of t by ANFIS.

2.3.3. Sensitivity and specificity

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \ (\%), \tag{17}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \ (\%). \tag{18}$$

2.3.4. k-Fold cross validation

k-Fold cross validation is one way to improve the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it is not important how the data is divided. Every data point appears in a test set exactly once, and appears in a training set $k-1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm must be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of this method is that we can independently choose the size of the each test and the number of trials [11]. We used this method as 10-fold cross validation in our applications. But we also conducted our experiments with there run for each training-test configuration. The average of these three test results gave us the test result for each fold. So we obtained 30 results in total to average.

2.3.5. Confusion matrix

A confusion matrix [12] contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Table 2 shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect of predictions that an instance negative, and
- d is the number of correct predictions that an instance is positive.

3. Results and discussion

The ANFIS classification with PCA of diabetes disease was classified due to training and test of all the diabetes disease dataset. The obtained test classification accuracy was 89.47% by using the 10-fold cross validation. The

Table 3
Classification accuracies obtained with our proposed system and other classifiers from literature

Author (year)	Accuracy (%)	Method
Logdisc [13]	77.7	Statlog
IncNet [13]	77.6	Norbert Jankowski
DIPOL92 [13]	77.6	Statlog
Linear discr. anal. [13]	77.5–77.2	Statlog; Ster and Dobnikar
SMART [13]	76.8	Statlog
GTO DT (5 × CV) [13]	76.8	Bennet and Blue
kNN, $k = 23$, Manh, raw, W [13]	76.7 ± 4.0	WD-GM, feature weighting 3CV
kNN, $k = 1{:}25$, Manh, raw [13]	76.6 ± 3.4	WD-GM, most cases $k = 23$
ASI [13]	76.6	Ster and Dobnikar
Fisher discr. analysis [13]	76.5	Ster and Dobnikar
MLP + BP [13]	76.4	Ster and Dobnikar
MLP + BP [13]	75.8 ± 6.2	Zarndt
LVQ [13]	75.8	Ster and Dobnikar
LFC [13]	75.8	Ster and Dobnikar
RBF [13]	75.7	Statlog
kNN, $k = 22$, Manh [13]	75.5	Karol Grudziński
MML [13]	75.5 ± 6.3	Zarndt
SNB [13]	75.4	Ster and Dobnikar
BP [13]	75.2	Statlog
SSV DT [13]	75.0 ± 3.6	WD-GM, SSV BS, node 5CV MC
kNN, $k = 18$, Euclid, raw [13]	74.8 ± 4.8	WD-GM
CART DT [13]	74.7 ± 5.4	Zarndt
DB-CART [13]	74.4	Shang and Breiman
ASR [13]	74.3	Ster and Dobnikar
SSV DT [13]	73.7 ± 4.7	WD-GM, SSV BS, node 10CV strat
C4.5 DT [13]	73.0	Stalog
Bayes [13]	72.2 ± 6.9	Zarndt
C4.5 (5 × CV) [13]	72.0	Bennet and Blue
CART [13]	72.8	Ster and Dobnikar
Kohonen [13]	72.7	Statlog
kNN [13]	71.9	Ster and Dobnikar
IB3 [13]	71.7 ± 5.0	Zarndt
IB1 [13]	70.4 ± 6.2	Zarndt
kNN, $k = 1$, Euclid, raw [13]	69.4 ± 4.4	WD-GM
kNN [13]	67.6	Statlog
C4.5 rules [13]	67.0 ± 2.9	Zarndt
OCN2 [13]	65.1 ± 1.1	Zarndt
Default [13]	65.1	
QDA [13]	59.5	Ster and Dobnikar
Logdisc [13]	77.7	Statlog
IncNet [13]	77.6	Norbert Jankowski
DIPOL92 [13]	77.6	Statlog
Linear discr. anal. [13]	77.5–77.2	Statlog; Ster and Dobnikar
SMART [13]	76.8	Statlog
GTO DT (5 × CV) [13]	76.8	Bennet and Blue
Yildirim et al.	77.08	BFGS quasi Newton
Yildirim et al.	77.60	Gradient descent
Yildirim et al.	77.08	Levenberg–Marquardt
Yildirim et al.	68.23	RBF
Yildirim et al.	80.21	GRNN
Our study (2006) (10 × FC)	89.47	PCA-ANFIS

obtained classification accuracy by ANFIS classifier with PCA for diabetes disease is the highest classifier among classifier report from literature. In view of classification accuracy, Table 3 is shown to compare these classifiers with our method.

Table 4
Confusion matrix

Output/desired	Result (normal)	Result (patient)
Result (normal)	24	4
Result (patient)	4	46

In this study, there were two classes as healthy and patient which were indicating situation of subject's diabetes disease. Classification results of the expert system were displayed by using a confusion matrix. In a confusion matrix, each cell contains the raw number of exemplars classified for the corresponding combination of desired and actual network outputs. The confusion matrix showing the classification results of this network is given in Table 4.

The obtained sensitivity, specificity and MSE values by PCA-ANFIS for diabetes disease are 85.71%, 92.0%, and 0.262.

A new hybrid system designed by PCA and ANFIS that we have built gave very promising results in classifying the healthy and patient subjects. We are proposing a complimentary system that can be coupled to software of the medical decision making devices. The benefit of the system is to assist the physician to make the final decision without hesitation.

The proposed method was arrived to the highest classification accuracy among classifiers in Table 3. It is shown from these results that our improved method was shown to be a corresponding and safe system to medical diagnostic decision-making.

4. Conclusion

With the improvements in expert systems and ML tools, the effects of these innovations are entering to more application domains day-by-day and medical field is one of them. Decision-making in medical field can sometimes be a trouble. Classification systems that are used in medical decision-making provide medical data to be examined in shorter time and more detailed.

Classification systems that are used in medical decision-making provide medical data to be examined in shorter time and more detailed. In this study, for the diagnosis of diabetes disease, a novel hybrid system based on PCA and ANFIS is proposed.

In the research reported in this paper, ANFIS with PCA was applied on the task of diagnosing diabetes disease and the most accurate learning methods was evaluated. Experiments were conducted on the diabetes disease dataset to diagnose diabetes disease in a fully automatic manner using ANFIS with PCA.

The results strongly suggest that ANFIS with PCA can aid in the diagnosis of diabetes disease. It is hoped that more interesting results will follow on further exploration of data.

Acknowledgment

This study has been supported by Scientific Research Project of Selcuk University (Project No. 05401069).

References

[1] http://www.fi.edu/biosci/healthy/diabetes.html, last accessed September 2006.
[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R.G.R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996.
[3] G. Piatetsky-Shapiro, W.J. Frawley, Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1991.
[4] D. Michie, Methodologies from machine learning in data analysis and software, Comp. J. 34 (6) (1991) 559–565.
[5] C.L. Blake, C.J. Merz, UJI Repository of Machine Learning Databases, 1996, http://www.ics.uci.edu/~mlearn/MLRepository.html.
[6] K. Kayaer, T. Yıldırım, Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), Istanbul, Turkey, June 26–29, 2003, pp. 181–184.
[7] X. Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, J. Pattern Recogn. Soc. 36 (2003) 2429–2439.
[8] I.S. Lindsay, A Tutorial on Principal Components Analysis, 2002, http://kybele.psych.cornell.edu/~edelman/Psych-465Spring-2003/PCA-tutorial.

[9] İ. Güler, D.E. Übeyli, Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients, J. Neurosci. Methods 148 (2005) 113–121.
[10] J.-S.R. Jang, Self-learning fuzzy controllers based on temporal backpropagation, IEEE Trans. Neural Network 3 (5) (1992) 714–723.
[11] R. Kohavi, F. Provost, Glossary of terms, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, no. 2–3, 1998.
[12] Jeff Schneider's home page, http://www.cs.cmu.edu/~schneide/tut5/node42.html, last accessed September 2006.
[13] Datasets used for classification comparison of results, http://www.phys.uni.torun.pl/kmk/projects/datasets.html.

Kemal Polat graduated from Electrical and Electronics Engineering Department of Selcuk University with B.Sc. degree in 2000 and from Electrical and Electronics Engineering Department of Selcuk University with M.Sc. degree in 2004. Subsequently, he is pursuing the Ph.D. degree at the Electrical and Electronics Engineering Department of Selcuk University. His current research interests are artificial immune systems, biomedical signals and digital signal processing, pattern recognition and classification.

Salih Günes graduated from Erciyes University in 1989. He took his M.Sc. degree in electrical and electronic engineering at Erciyes University in 1993. He took the Ph.D. degree in electrical and electronic engineering at Selcuk University in 2000. He is an Assistant Professor Dr. at the Department of Electrical and Electronics Engineering at Selcuk University. His interest areas are biomedical signal processing, artificial immune system, logic circuits, and artificial intelligence.