Available online at www.sciencedirect.com
ScienceDirect
Applied Mathematics and Computation 186 (2007) 898–906
APPLIED MATHEMATICS AND COMPUTATION
www.elsevier.com/locate/amc

Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine

Kemal Polat *, Salih Güneş

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

Abstract

Changes in the normal rhythm of a human heart may result in different cardiac arrhythmias, which may be immediately fatal or cause irreparable damage to the heart sustained over long periods of time. The ability to automatically identify arrhythmias from ECG recordings is important for clinical diagnosis and treatment. In this study, we have detected on ECG Arrhythmias using principal component analysis (PCA) and least square support vector machine (LS-SVM). The approach system has two stages. In the first stage, dimension of ECG Arrhythmias dataset that has 279 features is reduced to 15 features using principal component analysis. In the second stage, diagnosis of ECG Arrhythmias was conducted by using LS-SVM classifier. We took the ECG Arrhythmias dataset used in our study from the UCI (from University of California, Department of Information and Computer Science) machine learning database. Classifier system consists of three stages: 50–50% of training-test dataset, 70–30% of training-test dataset and 80–20% of training-test dataset, subsequently, the obtained classification accuracies; 96.86%, 100% ve 100%. The end benefit would be to assist the physician to make the final decision without hesitation. This result is for ECG Arrhythmias disease but it states that this method can be used confidently for other medical diseases diagnosis problems, too.

Keywords: ECG Arrhythmia; Principal component analysis (PCA); Least square support vector machine (LSSVM); ROC curves

1. Introduction

With improvements in medical knowledge systems in medical institutes and hospitals, determining useful knowledge is becoming more difficult. Especially, because the conventional manual data analysis techniques are not effective in diagnosis, using computer based analyses are becoming inevitable in disease diagnosis. So, it is the time to develop modern, effective and efficient computer based systems for decision support. There are a number of data analysis techniques: statistical, machine learning and data abstraction. Medical analysis

* Corresponding author.
E-mail addresses: kpolat@selcuk.edu.tr (K. Polat), sgunes@selcuk.edu.tr (S. Güneş).

0096-3003/$ - see front matter
doi:10.1016/j.amc.2006.08.020

using machine learning techniques has begun to be conducted for last 20 years. The advantages of using machine learning schemes in medical analysis have caused human support and costs to decrease and caused diagnosis accuracy to increase [1].

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 motivation behind the research reported in this paper is the results obtained from extensions of an ongoing research effort. The research reported in [6,5] is on developing a non-invasive ECG hardware and embedded software for capturing, analysing, diagnosing, and recommending remedies for homecare patients with heart conditions. In the effort, we focused on the (hardware) acquisition and (software) analysis of ECG signals for early diagnosis of Tachycardia heart disease. The work reported here builds on the initial work by, first, using machine learning techniques to study and understand the accurate prediction of arrhythmic diseases and suggestive remedies based on the classification schemes or models [6,5].

The used data source is taken from the University of California at Irvine (UCI) machine learning repository [7]. This dataset is commonly used among researchers who use machine learning (ML) methods for ECG Arrhythmia classification and so it provides us to compare the performance of our system with other conducted studies related with this problem.

In this study, we have proposed the system that has two stages. Firstly, dimension of ECG Arrhythmia dataset that has 279 features is reduced to 15 features using principal component analysis. Then, we used LS-SVM diagnosis ECG Arrhythmia. The obtained classification accuracy of our system 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 background information including ECG Arrhythmia classification problem and previous research in corresponding area. Also, we explained the method in Section 2 with subtitles of proposed a new medical diagnosis method and measures for performance evaluation. In each subsection of that section, the detailed information is 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. Materials and method

2.1. ECG Arrhythmia dataset

The dataset used in this study was obtained from the archives of machine learning datasets at the University of California, Irvine [7]. The datasets are grouped into three broad classes to facilitate their use in experimentally determining the presence or absence of arrhythmia, and for identifying the type of arrhythmia. In the set, Class 01 refers to 'normal' ECG. Classes 02–15 refer to different classes of arrhythmia and Class 16 refers to the rest of unclassified data. The arrhythmia dataset has 279 attributes, 206 of which are linear valued and the rest are nominal. There are 452 instances, and as indicated above, 16 classes. But we have used to two classes as the presence or absence of arrhythmia in our experiments. There are missing values in the dataset. In such cases, probabilistic values were assigned according to the distribution of the known values for the attributes.

2.2. Previous research in ECG Arrhythmia dataset

There has been much work in the field of classification and most work has been based on neural networks, Markov chain models and support vector machines (SVMs). The datasets used to train these methods are often small. In [5], direct-kernel methods and support vector machines are used for pattern recognition in

magnetocardiography. In [8], self-organizing maps (SOM) are used for analysis of ECG signals. The SOMs helps discover a structure in a set of ECG patterns and visualize a topology of the data. In [9] machine learning methods like artificial neural networks (ANNs) and logically weighted regression (LWR) methods are used for automated morphological galaxy classification. The focus of the investigation described in this paper is to evaluate three standard machine-learning algorithms applied to classify cardiac arrhythmias. All related previous research cited in this paper use classes, features, and machine learning methods and related software, which we used. Therefore, our comparisons are in the context of the predictability, accuracy, and ease of – learning of these algorithms. The former two capabilities are significant in diagnosing and treating ECG abnormalities while the latter facilitates the practical use of our ECG diagnostic device.

2.3. Proposed approach

We have proposed system, which has two stages. In the first stage, dimension of ECG Arrhythmia dataset that has 279 features is reduced to 4 features using principal component analysis. In the second stage, we used LS-SVM classifier to diagnosis of ECG Arrhythmia. The block diagram of proposed system is shown in Fig. 1.

2.3.1. Principal component analysis (PCA)

PCA was used to make a classifier system more effective. For this aim, before classifying, PCA method was used for dimensionality reduction of ECG Arrhythmia dataset. Therefore, ECG Arrhythmia dataset was represented a vector consists of 279 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 [10]. 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 $S = (1/N)\sum_{i=1}^{N}(x_i - \mu)^{\mathrm{T}}(x_i - \mu)$, where $x_i \in X$, μ is the sample mean and N is the number of samples, so that

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

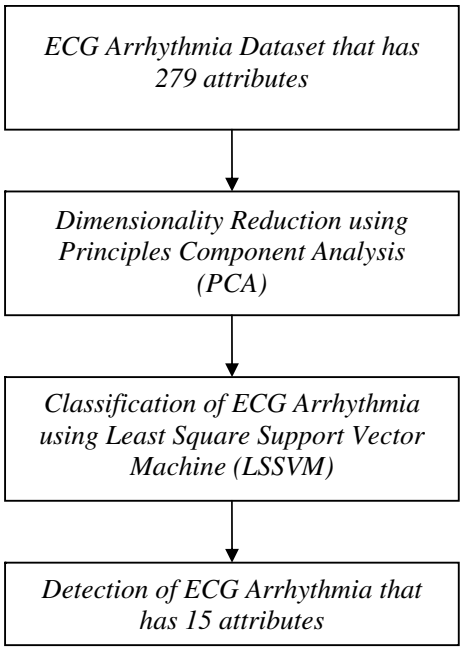

Fig. 1. The block diagram of proposed system.

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

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

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:

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

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 $\widehat{S}$:

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

where $\hat{\lambda}_i$ is the ith largest eigenvalue of $\widehat{S}$. 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 [11].

2.3.2. Least square support vector machine (LSSVM)

In this section we firstly mention about SVM classifier after that LSSVM related to SVM.

2.3.2.1. Support vector machines (SVMs). SVM is a reliable classification technique, which is based on the statistical learning theory. This technique was firstly proposed for classification and regression tasks by [12].

As shown in Fig. 2, a linear SVM was developed to classify the data set which contains two separable classes such as $\{+1, -1\}$. Let the training data consist of n datum $(x_1, y_1), \ldots, (x_n, y_n)$, $x \in R^n$ and $y \in \{+1, -1\}$. To separate these classes, SVMs have to find the optimal (with maximum margin) separating hyperplane so that SVM has good generalization ability. All of the separating hyperplanes are formed with

$$D(x) = (w * x) + w_0 \tag{5}$$

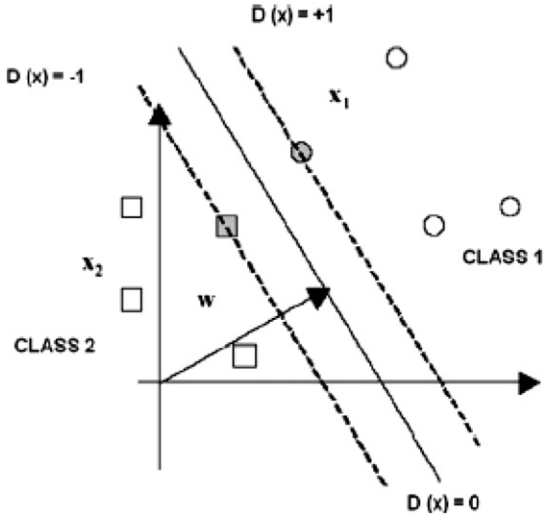

Fig. 2. The structure of a simple SVM.

and provide following inequality for both $y = +1$ and $y = -1$.

$$y_i[(w * x_i) + w_0] \geqslant 1, \quad i = 1, \ldots, n. \tag{6}$$

The data points which provide above formula in case of equality are called the support vectors. The classification task in SVMs is implemented by using of these support vectors.

Margins of hyperplanes obey following inequality:

$$\frac{yk \times D(xk)}{\|w\|} \geqslant \Gamma, \quad k = 1, \ldots, n. \tag{7}$$

To maximize this margin (Γ), norm of w is minimized. To reduce the number of solutions for norm of w, following equation is determined.

$$\Gamma \times \|w\| = 1. \tag{8}$$

Then formula (9) is minimized subject to constraint (6).

$$1/2\|w\|^2. \tag{9}$$

When we study on the non-separable data, slack variables ξ_i, are added into formula (6) and (9). Instead of formulas (6) and (9), new formulas (10) and (11) are used.

$$y_i[(w \cdot x_i) + w_0] \geqslant 1 - \xi_i, \tag{10}$$

$$C \sum_{i=1}^{n} \xi_i + 1/2\|w\|^2. \tag{11}$$

Since originally SVMs classify the data in linear case, in the nonlinear case SVMs do not achieve the classification tasks. To overcome this limitation on SVMs, kernel approaches are developed. Nonlinear input data set is converted into high dimensional linear feature space via kernels. In SVMs, following kernels are most commonly used.

- dot product kernels: $K(x, x') = x \cdot x'$;
- polynomial kernels: $K(x, x') = (x \cdot x' + 1)^d$; where d is the degree of kernel and positive integer number;
- RBF kernels: $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$; where σ is a positive real number.

In our experiments σ is selected 10,000.

2.3.3. LSSVM (least squares support vector machines)

LSSVMs are proposed by [13]. The most important difference between SVMs and LSSVMs is that LSSVMs use a set of linear equations for training while SVMs use a quadratic optimization problem [14]. While formula (11) is minimized subject to formula (10) in Vapnik's standard SVMs, in LSSVMs formula (13) is minimized subject to formula (12).

$$y_i[(w \cdot x_i) + w_0] = 1 - \xi_i, \quad i = 1, \ldots, n. \tag{12}$$

$$1/2\|w\|^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2. \tag{13}$$

According to these formulas, their dual problems are built as following:

$$Q(w, b, \alpha, \xi) = 1/2\|w\|^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i \{y_i[(w \cdot x_i) + w_0] - 1 + \xi_i\}. \tag{14}$$

Another difference between SVMs and LSSVMs is that α_i (Lagrange multipliers) are positive or negative in LSSVMs but they must be positive in SVMs. Information in detailed is found in [13,14].

2.4. The performance evaluation of LSSVM on the diagnosis of ECG Arrhythmia

2.4.1. Classification accuracy

In this study, the classification accuracy for the dataset was measured according to Eq. (15):

$$\text{accuracy}(T) = \frac{\sum_{i=1}^{|T|} \text{assess}(t_i)}{|T|}, \quad t_i \in T, \tag{15}$$

$$\text{assess}(t) = \begin{cases} 1 & \text{if classify}(t) \equiv t.c, \\ 0 & \text{otherwise}, \end{cases}$$

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

2.4.2. Sensitivity, specificity, TP rate, FP rate, accuracy and F-measure

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}(\%), \tag{16}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}(\%), \tag{17}$$

$$\text{FP}_{\text{rate}} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{18}$$

$$\text{TP}_{\text{rate}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{19}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \tag{20}$$

$$F\text{-measure} = \frac{2}{1/\text{specificity} + 1/\text{sensitivity}}, \tag{21}$$

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives respectively.

True positive (TP): An input is detected as a patient with atherosclerosis diagnosed by the expert clinicians.
True negative (TN): An input is detected as normal that was labeled as a healthy by the expert clinicians.
False positive (FP): An input is detected as a patient that was labeled as a healthy by the expert clinicians.
False negative (FN): An input is detected as normal with atherosclerosis diagnosed by the expert clinicians.

2.4.3. Receiver operating characteristic (ROC) curves

A receiver operating characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance. ROC graphs are commonly used in medical decision making, in recent years have been used increasingly in machine learning and data mining research. Although ROC graphs are apparently simple, there are some common misconceptions and pitfalls when using them in research [15].

ROC graphs are two-dimensional graphs in which tp (true positive) rate is plotted on the Y axis and fp (false positive) rate is plotted on the X axis. An ROC depicts relative tradeoffs between benefits (true positives) and costs (false positives) [15].

3. Results and discussion

The PCA-LSSVM classification of ECG Arrhythmia was classified as 50–50%, 70–30%, and 80–20%, respectively due to training and test of all the ECG Arrhythmia dataset. The obtained test classification accuracies were 96.89%, 100%, and 100% respectively. Our experimental study, ECG Arrhythmia that has normal heart and patient heart are classified by LSSVM classifier.

The obtained classification accuracy, sensitivity, specificity, TP rate, FP rate, accuracy and F-measure values by PCA-LSSVM classifier for diagnosis of ECG Arrhythmia with 50–50% training-test dataset, 70–30% training-test dataset and 80–20% training-test dataset were shown in Table 1.

To compare the classification performances of PCA-LSSVM with 50–50% training-test dataset, 70–30% training-test dataset and 80–20% training-test dataset, ROC (Receiver Operator Characteristic) curves method is preferred. According to this method, ROC curves and area under these curves are computed for all datasets. While ROC curve of PCA-LSSVM with 50–50% training-test dataset was shown in Fig. 3, ROC curve of

Table 1
The experimental results obtained by PCA-LSSVM classifier for diagnosis of ECG arrhythmia

Number of dataset	Datasets	Measures	PCA-LSSVM
Number of normal ECG	50–50% of training-test partition	Sensitivity (%)	100
Arrhythmia: **245**		Specificity (%)	100
Training: **123**		TP rate (%)	100
Test: **122**			
Number of diseased ECG		FP rate (%)	0
Arrhythmia: **207**		Accuracy (%)	100
Training: **103**		*F*-measure (%)	100
Test: **102**			
Number of normal ECG	70–30% of training-test partition	Sensitivity (%)	100
Arrhythmia: **245**		Specificity (%)	100
Training: **172**		TP rate (%)	100
Test: **73**			
Number of diseased ECG		FP rate (%)	0
Arrhythmia: **207**		Accuracy (%)	100
Training: **145**		*F*-measure (%)	100
Test: **62**			
Number of normal ECG	80–20% of training-test partition	Sensitivity (%)	100
Arrhythmia: **245**		Specificity (%)	100
Training: **196**		TP rate (%)	100
Test: **49**			
Number of diseased ECG		FP rate (%)	0
Arrhythmia: **207**		Accuracy (%)	100
Training: **166**		*F*-measure (%)	100
Test: **41**			

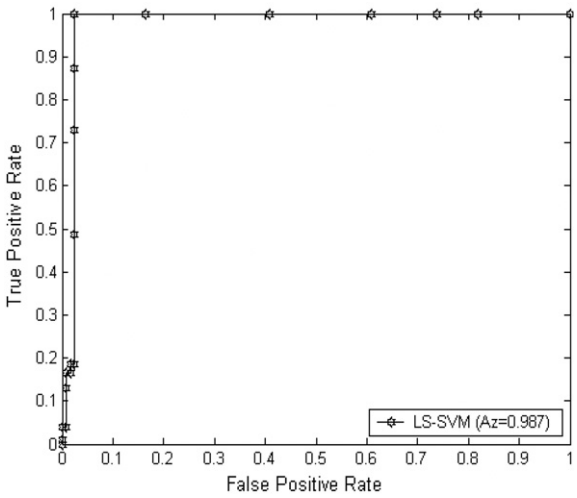

Fig. 3. ROC curve for LSSVM on 50–50% of training-test partition.

PCA-LSSVM with 70–30% training-test dataset was shown in Fig. 4. ROC curve of PCA-LSSVM with 80–20% training-test dataset was shown in Fig. 5.

ROC curves is a statistical comparing method which uses the rates of true positive and false positive. Areas under ROC curves are represented by Az value. This value is related to the accuracies of classifiers. Higher values represent higher classification accuracies, while lower values represent lower classification accuracies [16,17]. ROC curves show that there is a significant difference between computed areas for classifiers. In this

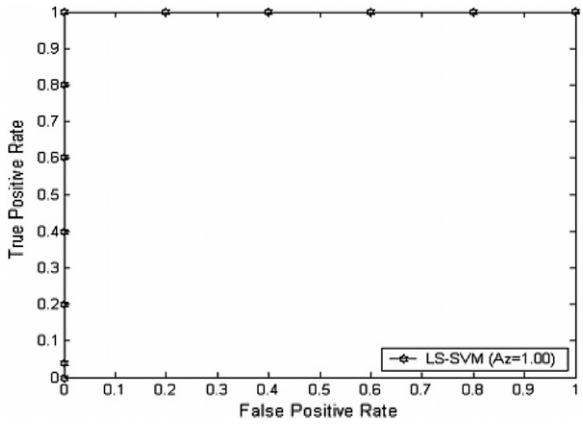

Fig. 4. ROC curve for LSSVM on 70–30% of training-test partition.

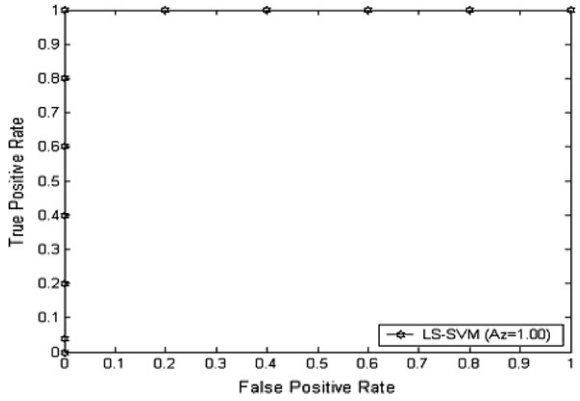

Fig. 5. ROC curve for LSSVM on 80–20% of training-test partition.

study, Az areas have been computed as $Az = 0.987$ for LSSVM with 50–50% training-test dataset, $Az = 1.00$ for LSSVM with 70–30% training-test dataset and $Az = 1.00$ for LSSVM with 80–20% training-test dataset.

Our technique gets around this problem using PCA and least square support vector machine (LSSVM) to decide and assist the physician to make the final judgment in confidence.

4. Conclusion

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 ECG Arrhythmia, a novel medical decision support system based on PCA and LSSVM proposed.

In the research reported in this paper, PCA-LSSVM was applied on the task of diagnosing ECG Arrhythmia and the most accurate learning methods were evaluated. Experiments were conducted on the ECG Arrhythmia dataset to diagnose cardiac arrhythmias in a fully automatic manner using PCA and LSSVM. The results strongly suggest that PCA and LSSVM can aid in the diagnosis of ECG Arrhythmia. It is hoped that more interesting results will follow on further exploration of data.

Although developed method is built as an offline diagnosing system, it can be rebuilt as an online diagnosing system in the future.

Acknowledgements

This study is supported by the Scientific Research Projects of Selcuk University (Project no: 05401069).

References

[1] N. Cheung, Machine learning techniques for medical analysis, School of Information Technology and Electrical Engineering, BSc thesis, University of Queenland, 2001.
[2] G. Piatetsky-Shapiro, W.J. Frawley, Knowledge Discovery in Databases, AAAI Press, Menlo Park, CA, 1991.
[3] D. Michie, Methodologies from Machine Learning in Data Analysis and Software, Computer Journal 34 (6) (1991) 559–565.
[4] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R.G.R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, Menlo Park, CA, 1996.
[5] M. Embrechts, B. Szymanski, K. Sternickel, T. Naenna, R. Bragaspathi, Use of Machine Learning for Classification of Magnetocardiograms, in: Proc. IEEE Conference on System, Man and Cybernetics. Washington, DC, 2003, pp. 1400–1405.
[6] M. Pazzani, D. Kibler, The Utility of Knowledge in Inductive Learning, Machine Learning 9 (1) (1992) 57–94.
[7] ECG Arrhythmia Dataset. UCI Repository of Machine Learning Databases. Available from: <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/> (accessed July 2006).
[8] G. Bortolan, W. Pedrycz, An interactive framework for an analysis of ECG signals, Artificial Intelligence in Medicine 24 (2002) 109–132.
[9] J. de la Calleja, O. Fuentes, Machine learning and image analysis for morphological galaxy classification, Monthly Notices of the Royal Astronomical Society 349 (2004) 87–93.
[10] X. Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, Pattern Recognition. The Journal of The Pattern Recognition Society 36 (2003) 2429–2439.
[11] Lindsay, I. Smith, A tutorial on principal components analysis. Available from: <http://kybele.psych.cornell.edu/~edelman/Psych-465Spring-2003/PCA-tutorial>, 2002.
[12] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[13] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (3) (1999) 293–300.
[14] Tsujinishi Daisuke, Abe Shigeo, Fuzzy Least Squares Support Vector Machines for Multi-Class Problems, Neural Networks Field, 16, Elsevier, 2003, pp. 785–792.
[15] F. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.
[16] A. Osareh, M. Mirmehdi, B. Thomas, R. Markham, Comparative exudate classification using support vector machines and neural networks, in: T. Dohi, R. Kikinis (Eds.), 5th International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS, vol. 2489, Springer, 2002, pp. 413–420.
[17] R.M. Centor, Signal Detectability: The use of ROC Curves and their Analysis, Medical Decision Making 11 (1991) 102–106.