Archives of Control Sciences
Volume 12(XLVIII), 2002
No. 3, *pages 289–299*

Minimum absolute error classifier design with generalization control

JACEK ŁĘSKI

This paper introduces a new classifier design method, that is based on an extension of the classical Ho-Kashyap procedure. The proposed method uses absolute error rather than squared error to design a linear classifier. Additionally, easy control of generalization ability and outliers robustness is obtained. Finally, examples are given to demonstrate the validity of the introduced method.

Key words: classifier design, robust methods, generalization control

1. Introduction

Pattern recognition is concerned with classification of patterns into categories. This field of study was developed in the early 60s, and it has recently played an important role in many engineering fields, such as medical diagnosis, computer vision, character recognition, data mining, communication and so on. Two of the main textbooks on pattern recognition are those by Duda and Hart [5] and Tou and Gonzalez [15].

There are two main types of classification methods: supervised (discrimination) and unsupervised (clustering). In supervised classification we have a set of data called a training set that has class labels associated with each datum. The unsupervised methods divide a set of observations into groups, so that members of the same group are more similar to one another than to members of other groups. The most important feature of a classifier is its generalization ability, which refers to producing reasonable decision for data unseen during the training process (designing). The easiest way to measure the generalization ability is to choose a test set that contains data whose elements do not belong to the training set.

From the statistical learning theory, we know that in order to achieve a good generalization capability, we should select the classifier with the smallest Vapnik-Chervonenkis (VC) dimension (complexity) and the smallest error rate on the training set. This principle is called the principle of structural risk minimization [18], [17].

The author is with Institute of Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland.

Received 7.1.2002, revised 18.9.2002.

In real applications, data from the training set are corrupted by noise and outliers. It follows that classifier design methods need to be robust. According to Huber [10], a robust method should have the following properties: i) it should have a reasonably good accuracy at the assumed model, ii) small deviations from the model assumptions should impair the performance only by a small amount, iii) larger deviations from the model assumptions should not cause a catastrophe. In literature there are many robust loss functions [10]. In this work, an absolute error loss function is of special interest due to its simplicity.

In literature there are many classifiers, including: statistical, linear discriminant, K-nearest neighbor, kernel, neural network, classification tree, and many more [5], [14], [15], [19]. But linear classifiers are of special interest due to its simplicity and easy expansibility to nonlinear classifiers. One of the most powerful classical methods of designing linear classifiers is the least mean-squared error procedure with Ho-Kashyap modification [8], [9]. The two main disadvantages of this approach are: i) the use of the quadratic loss function that leads to a non-robust method, ii) impossibility to minimize the VC dimension of the designed classifier.

The goal of this work is to introduce an extension of the classical Ho-Kashyap procedure. This new method uses the absolute loss function, resulting in robustness to outliers. Additionally, this method minimizes the VC dimension of the designed classifier. The rest of this work is concerned with the two-class problem. The proposed method can be easily generalized to the multi-class problem using the class-rest or the class-class methodology [15].

This paper is organized as follows: Section 2 presents a description of the minimum absolute error classifier design procedure with generalization control. Section 3 presents simulation results and discussion for classification of simple synthetic two-dimensional data and real-world high-dimensional data. Finally, conclusions are drawn in Section 4.

2. The method

The classifier is designed on the basis of a set of data called a training set, $Tr^{(N)} = \{(\mathbf{x_1}, \varphi_1), (\mathbf{x_2}, \varphi_2), ..., (\mathbf{x_N}, \varphi_N)\}$, where N is cardinality of the set, and each independent datum (pattern) $\mathbf{x_i} \in \mathbb{R}^{\mathbf{t}}$ has a corresponding dependent datum $\varphi_i \in \{+1, -1\}$, which indicate the assignment to one of two classes, ω_1 or ω_2

$$\varphi_i = \begin{cases} +1, & \mathbf{x_i} \in \omega_1, \\ -1, & \mathbf{x_i} \in \omega_2. \end{cases} \tag{1}$$

Defining the augmented pattern vector $\mathbf{x_i'} = \left[\mathbf{x_i^T}, \mathbf{1}\right]^{\mathbf{T}}$, we seek a weight vector $\mathbf{w} \in \mathbb{R}^{\mathbf{t+1}}$, such that

$$d(\mathbf{x_i}) \triangleq \mathbf{w^T x_i'} \begin{cases} > 0, & \mathbf{x_i'} \in \omega_1, \\ < 0, & \mathbf{x_i'} \in \omega_2, \end{cases} \tag{2}$$

where $d(\mathbf{x_i})$ is called linear discrimination (or decision) function.

If conditions (2) are satisfied for all members of the training set, then the data are said to be linearly separable. For overlapping classes it is impossible to find a weight vector $\mathbf{w}$, such that conditions (2) are satisfied for all data from the training set. If we multiply by -1 all patterns of the training set who are members of ω_2 class, then (2) can be rewritten in the form: $\varphi_i \mathbf{w^T x_i'} > \mathbf{0}$, for $i = 1, 2, ..., N$. Let $\mathbf{X}$ be the $N \times (t+1)$ matrix

$$\mathbf{X} \triangleq \begin{bmatrix} \varphi_1 \mathbf{x_1'^T} \\ \varphi_2 \mathbf{x_2'^T} \\ \vdots \\ \varphi_N \mathbf{x_N'^T} \end{bmatrix}. \tag{3}$$

Then conditions (2) can be written in a matrix form: $\mathbf{Xw} > \mathbf{0}$. To obtain a solution, the above linear inequalities system is replaced by the following linear equalities system: $\mathbf{Xw} = \mathbf{b}$, where $\mathbf{b}$ is an arbitrary positive vector, $\mathbf{b} > \mathbf{0}$. Now, we seek $\mathbf{w}$ and $\mathbf{b}$ vectors by minimization of the criterion function

$$\min_{\mathbf{w} \in \mathbb{R}^{\mathbf{t+1}}, \mathbf{b} > \mathbf{0}} I(\mathbf{w}, \mathbf{b}) \triangleq (\mathbf{Xw} - \mathbf{b})^T \mathbf{D} (\mathbf{Xw} - \mathbf{b}) + \tau \mathbf{w_n^T w_n}, \tag{4}$$

where $\mathbf{w_n}$ is a narrowed vector $\mathbf{w}$, with the last component of $\mathbf{w}$ excluded. The matrix $\mathbf{D} = \text{diag}(d_1, d_2, ..., d_N)$, where d_i is a weight corresponding to the i-th pattern, can be interpreted as reliability attached to the i-th pattern. The criterion function (4) is the squared error weighted by coefficients d_i with the second term related to minimization of the Vapnik-Chervonenkis dimension (complexity) of classifier. Parameter $\tau > 0$ controls the trade-off between the classifier complexity and the error tolerance.

We get conditions for optimality by differentiating (4) with respect to $\mathbf{w}$, $\mathbf{b}$ and setting the results equal to zero

$$\begin{cases} \mathbf{w} = \left(\mathbf{X^T D X} + \tau \widetilde{\mathbf{I}} \right)^{-1} \mathbf{X^T D b}, \\ \mathbf{e} \triangleq \mathbf{Xw} - \mathbf{b} = \mathbf{0}, \end{cases} \tag{5}$$

where $\widetilde{\mathbf{I}}$ is the identity matrix with the last element on the main diagonal set to zero.

From Equation (5), we see that the vector $\mathbf{w}$ depends on vector $\mathbf{b}$. The vector $\mathbf{b}$ is called margin vector because its components determine the distance from patterns to the separating hyperplane. For fixed $\mathbf{w}$, if a pattern is lying on the right side of the hyperplane, the corresponding margin can be increased to obtain zero error. However, if a pattern is lying on the wrong side of the hyperplane, then the error is negative and we may decrease the error only by decreasing the corresponding margin value. To prevent $\mathbf{b}$ from converging to zero we start with $\mathbf{b} > \mathbf{0}$ and refuse to decrease any of its components. Ho and Kashyap proposed an iterative algorithm for alternately determining $\mathbf{w}$ and $\mathbf{b}$, where the components of $\mathbf{b}$ cannot decrease. Now, this algorithm can be extended to our weighted squared error criterion with regularization. Vector $\mathbf{w}$ is determined on the

basis of the first equation from (5), i.e., $\mathbf{w^{(k)}} = \left(\mathbf{X^T D X} + \tau\widetilde{\mathbf{I}}\right)^{-1}\mathbf{X^T D b^{(k)}}$, where superscript (k) denotes the iteration index. Components of the vector $\mathbf{b}$ are modified by components of the error vector $\mathbf{e}$, but only in the case when it results in increase of components of $\mathbf{b}$. Otherwise, components of $\mathbf{b}$ remain unmodified

$$\mathbf{b^{(k+1)}} = \mathbf{b^{(k)}} + \rho\left(\mathbf{e^{(k)}} + \left|\mathbf{e^{(k)}}\right|\right), \tag{6}$$

where $\rho > 0$ is a parameter. Note that for $\mathbf{D} = \mathbf{I}$ ($d_i = 1$) and $\tau = 0$ the original Ho-Kashyap algorithm is obtained.

Now, another method for the selection of parameters d_i will be shown. Real data has noise and outliers. As a result classifier design methods need to be robust. It is well-known from literature [10], that the minimum squared error procedure does not lead to robust methods. One of the simplest technique to obtain the robust method is to use the minimum absolute error procedure. The absolute error criterion is easy to obtain by taking $d_i = 1/|e_i|$ for $i = 1, 2, ..., N$, where e_i is the i-th component of the error vector. However, the error vector depends on $\mathbf{w}$. So, we use the vector $\mathbf{w}$ from the previous iteration. This procedure is based on the premise that near the optimum solution sequential vectors $\mathbf{w^{(k)}}$ differs imperceptibly. The absolute error minimization procedure of classifier design can be summarized in the following steps:

1. Fix $\tau > 0$, $\rho > 0$ and $\mathbf{D^{(1)}} = \mathbf{I}$. Initialize $\mathbf{b^{(1)}} > \mathbf{0}$. Set the iteration index $k = 1$,
2. $\mathbf{w^{(k)}} = \left(\mathbf{X^T D^{(k)} X} + \tau\widetilde{\mathbf{I}}\right)^{-1}\mathbf{X^T D^{(k)} b^{(k)}}$,
3. $\mathbf{e^{(k)}} = \mathbf{X w^{(k)}} - \mathbf{b^{(k)}}$,
4. $d_i^{(k)} = 1/\left|e_i^{(k)}\right|$, for $i = 1, 2, ..., N$, $\mathbf{D^{(k+1)}} = \mathrm{diag}\left(d_1^{(k)}, ..., d_N^{(k)}\right)$,
5. $\mathbf{b^{(k+1)}} = \mathbf{b^{(k)}} + \rho\left(\mathbf{e^{(k)}} + \left|\mathbf{e^{(k)}}\right|\right)$,
6. if $\left\|\mathbf{b^{(k+1)}} - \mathbf{b^{(k)}}\right\| > \xi$, then $k = k + 1$, go to 2 else stop.

Remarks

The quantity ξ is a pre-set parameter. If step 4 in this algorithm is omitted, then the squared error minimization procedure is obtained. In practice, a divide-by-zero error in step 4 does not occur. It follows from the fact that some components of vector $\mathbf{e}$ go to zero as (k) goes to infinity. But in this case convergence is slow and the condition 6 stops the algorithm.

It is worth noting that the traditional method to solve the Least Absolute Value (LAV) problems leads to the linear programming problem with equality constraints [1], [2]. For large training set standard optimization techniques quickly become intractable in their memory and time requirements. The above presented algorithm is very similar to the Iterative Re-weighting Least Squares (IRLS) [12], [6], [11]. However, in the original IRLS

algorithm during the iteration process the vector $\mathbf{b}$ remains unmodified. The presented approach to solve linear matrix inequalities differs from IRLS by incorporating the modification of the margin vector $\mathbf{b}$ and the term related to minimization of the complexity of the solution.

It should also be noted that the presented algorithm can be viewed as a general method to solve the Matrix Linear Inequalities (LMI). LMI recently plays an important role in many control applications, such as control structure selection, optimal design of experiments, robust filtering and robust controller analysis and design [16], [13]. In [16] it is reported that the currently available software using an interior point algorithm (e.g. Matlab's LMI Control Toolbox) can handle problems with size up to 100×100. Several tests confirm the usefulness of the proposed method for problems with size up to 1000×1000.

3. Numerical experiments and discussion

In all experiments, $\mathbf{b}^{(1)} = \mathbf{10}^{-\mathbf{6}}$ was used. The iterations were stopped as soon as the Euclidean norm in a successive pair of $\mathbf{b}$ vectors was less than 10^{-4}. All computations were run on Pentium IV 1.6 GHz processor running Windows NT4 and MATLAB environment. Benchmark databases were obtained via Internet - *ftp://markov.stats.ox.ac.uk/pub/PRNN* and *http://www.ics.uci.edu/~mlearn*.

3.1. A simple synthetic two-dimensional data

The purpose of this experiment was to compare the classical Ho-Kashyap and the method of classifier design proposed in this work. The simulations were done for data generated by Ripley [14]. These data consist of patterns characterized by two features and they are assigned to two classes. Each class has a bimodal distribution obtained as a mixture of two normal distributions. The class distribution was chosen to allow the best-possible error rate of about 8% (for nonlinear classifier). The training set consists of 250 patterns (125 patterns belong to each class), and the testing set consists of 1000 patterns (500 patterns belong to each class).

The parameter τ was in the range from 0 to 10 with the step 0.1, and the parameter ρ was equal to 0.05, 0.5, 1.0 and 2.0. For each combination of the above parameters values, after the training stage (classifier design on the training set), the generalization ability of classifier was determined as the misclassification error rate on the test set. This error rate is presented in Fig.1 for minimization of the squared error ($\mathbf{D} = \mathbf{I}$) and in Fig.2 for minimization of the absolute error.

In case of, the squared error minimization procedure, the best generalization is obtained for $\rho = 0.05$ and τ from 2.6 to 3.1. For other values of the parameter ρ we also have an optimum of the generalization ability, but it is worse from that obtained for $\rho = 0.05$. The classical Ho-Kashyap method (for $\tau = 0$) independently of the parameter ρ value leads to worse generalization.

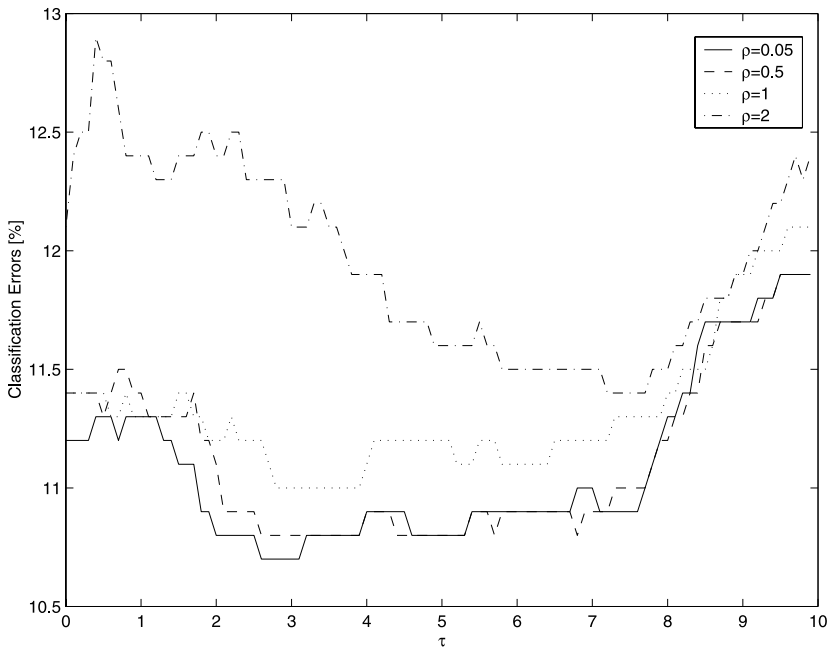

Figure 1. The squared error classifier design results as a function of the "regularization" parameter τ

In case of, the absolute error minimization procedure (see Fig.2), the best generalization is obtained for $\rho = 0.5$, $\tau = 1.0$ and $\rho = 1$, $\tau = 5.1$. We also see that the misclassification error rate is less sensitive to the parameter ρ, and obtained generalization is better for the absolute error minimization method. The classifier with the best generalization ($\rho = 0.5$, $\tau = 1.0$) is presented with the testing set in Fig.3. The error rate in this case is equal to 10.2%.

For comparison the linear support vector machine was used. Computationally effective method called incremental support vector learning was chosen [3]. The Matlab code of this method is available at *http://bach.ece.jhu.edu/pub/gert/svm/incremental*. For this method error rate is equal to 10.2% (for regularization parameter equal to 5.0), but computation time was approximately 10-times longer.

3.2. Real high-dimensional data

The main goal of this experiments was to examine a usefulness of the proposed method to construct a classifier for real-world high-dimensional data.

3.2.1. Pima Indians diabetes

The main goal of this experiments was to examine a usefulness of the proposed method to construct a classifier for real-world high-dimensional data. The data were collected by US National Institute of Diabetes and Kidney Diseases. According to the

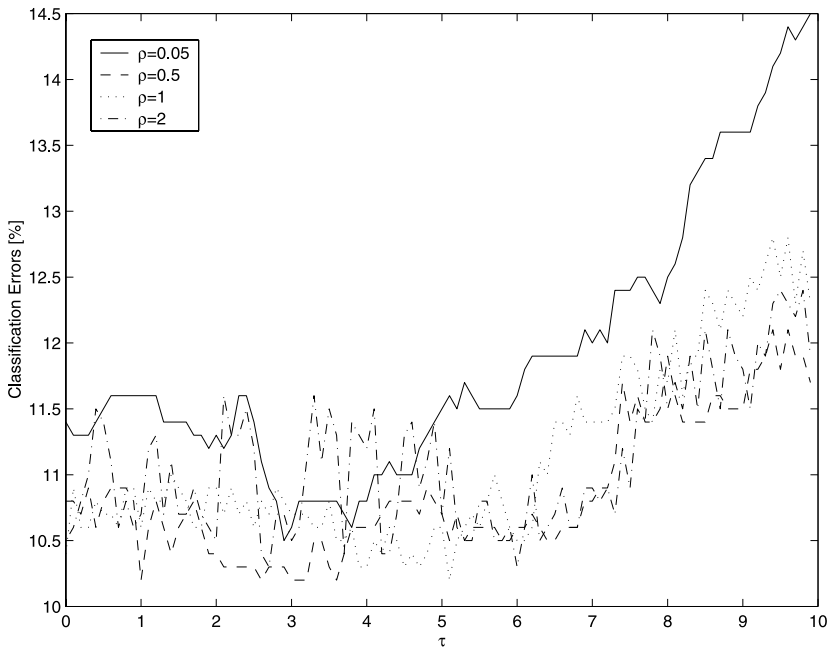

Figure 2. The absolute error classifier design results as a function of the "regularization" parameter τ

criteria of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, a population of women who were at least 21 years old was tested for diabetes. The women are of Pima Indians (living near Phoenix, Arizona). For each woman the following personal data were collected: the number of pregnancies, plasma glucose concentrations in fasting plasma glucose test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), body mass index (weight in kg / (height in m)^2), diabetes pedigree function (the function of the number and location in pedigree tree of common ancestors up to second degree relatives suffering from diabetes mellitus), age in years. In 768 collected records 376 were incomplete. Ripley divided randomly the complete records into a training set of size 200 and a test set of size 332. In the book by Ripley [14] the performance of several classical pattern recognition methods was tested using Pima Indians database. The obtained error rate (in percent) was as follows: linear discrimination - 20.2%, projection pursuit regression - 22.6%, backpropagation neural network - 21.1%, and learning vector quantization - 21.1%.

The parameter τ was in the range from 0 to 10 with step 0.1, and the parameter ρ was equal to 0.05, 0.5, 1.0 and 2.0. Table 1 shows the minimal error rate determined on the testing set for each value of the parameter ρ. From this table we see that the best generalization, equal to 18.67%, is obtained for the absolute error procedure with $\rho = 2.0$ and $\tau = 3.4$. The classical Ho-Kashyap method leads to the error rate equal to 22.5% independently of the parameter ρ value. The error rate for the incremental support

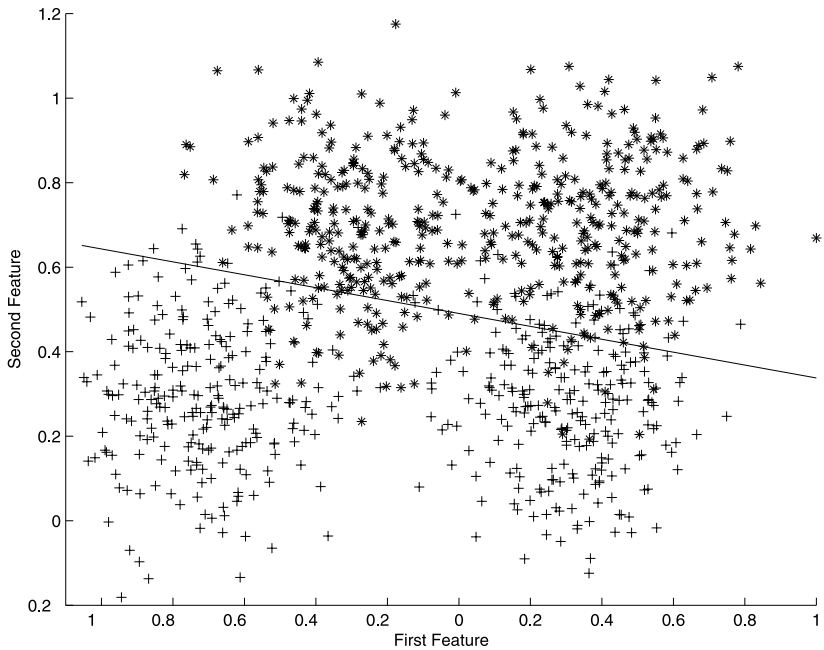

Figure 3. The testing set for Ripley two-class problem and the classifier line with the best generalization ability

Table 1. Simulation results for Pima Indians classification.

ρ	Squared error procedure	Absolute error procedure
0.05	20.18% for $\tau = 1.0$	20.78% for $\tau = 2.0$
0.5	20.18% for $\tau = 1.0$	18.97% for $\tau = 4.3$
1	19.87% for $\tau = 2.8$	19.27% for $\tau = 0.1$
2	18.97% for $\tau = 0.7$	18.67% for $\tau = 3.4$

vector machine is equal to 19.3% for regularization parameter equal to 0.5. In this case, the computation time for this method was also approximately 10-times longer comparing with the absolute error procedure.

3.2.2. Heart Disease database

The heart disease database contains 4 subdatabases concerning heart disease diagnosis. The data was collected from the four following locations: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert De-

Table 2. Simulation results for Heart Disease classification.

Squared error procedure	Absolute error procedure
21.97% for $\tau = 0$	21.66% for $\tau = 0$
21.81% for $\tau = 4.5$	21.04% for $\tau = 1.6$

trano, M.D., Ph.D. Numbers of instances are: Cleveland: 303, Hungarian: 294, Switzerland: 123, Long Beach VA: 200. For each person the following data were collected: 1) age in years, 2) sex (1 $\doteq$ male; 0 $\doteq$ female), 3) chest pain type (1 $\doteq$ typical angina, 2 $\doteq$ atypical angina, 3 $\doteq$ non-anginal pain, 4 $\doteq$ asymptomatic), 4) resting blood pressure (in mm Hg), 5) serum cholesterol (in mg/dl), 6) fasting blood sugar > 120 mg/dl (0 $\doteq$ false, 1 $\doteq$ true), 7) resting electrocardiographic results (0 $\doteq$ normal, 1 $\doteq$ having ST-T wave abnormality - T wave inversions and/or ST elevation or depression of > 0.05 mV, 2 $\doteq$ showing probable or definite left ventricular hypertrophy by Estes' criteria), 8) maximum heart rate achieved, 9) exercise induced angina (0 $\doteq$ no, 1 $\doteq$ yes), 10) ST depression induced by exercise relative to rest, 11) the slope of the peak exercise ST segment (1 $\doteq$ upsloping, 2 $\doteq$ flat, 3 $\doteq$ downsloping), 12) number of major vessels colored by fluoroscopy, 13) thal perfusion scynthygraphy (3 $\doteq$ normal, 6 $\doteq$ fixed defect, 7 $\doteq$ reversible defect), 14) amniography disease status; < 50% major vessel diameter narrowing (0 $\doteq$ false, 1 $\doteq$ true). The "class" field refers to the presence of coronary artery disease in the patient.

The misclassification error rate was estimated using twenty-fold cross-validation. In each realization, the Heart Disease database was randomly divided into a training set of size 250 and a test set of size 670. The parameter τ was in the range from 0 to 10 with step 0.1, and the parameter ρ was equal to 1.0. Table 2 shows the misclassification error rate estimated by the cross-validation. From this table we see that the best generalization, equal to 21.04%, is obtained for approximation to the absolute error procedure with $\tau = 3.5$. The classical Ho-Kashyap method leads to the error rate equal to 21.97%. The error rate for the incremental support vector machine is equal to 21.3% for regularization parameter equal to 0.1. In this case, the computation time for this method also was approximately 10-times longer comparing the absolute error procedure. The Fisher's linear discriminant leads to the error rate equal to 21.96%. The logistic-regression-derived discriminant function [4] leads to the error rate equal to 23% and the Bayes point machine [7] leads to the error rate 22.8%.

4. Conclusions

In this work, a new classifier design method is introduced. This method is an extension of the classical Ho-Kashyap methodology, which uses the absolute loss function

rather than the quadratic one. This results in robustness to outliers. Additionally, the proposed method minimizes the Vapnik-Chervonenkis dimension that results in easy control of generalization ability of the classifier. Three numerical examples are given to illustrate the validity of the presented method. These examples show the usefulness of the minimum absolute error classifier design procedure in the classification of synthetic as well as real-world high-dimensional data.

References

[1] I. BARRONDALE and F.D.K. ROBERTS: An Improved Algorithm for Discrete L_1 Linear Approximation. *SIAM J. Numer. Anal.*, **10**(5), (1973), 839-848.

[2] I. BARRONDALE and A. YOUNG: Algorithms for Best L_1 and L_∞ Linear Approximations on a Discrete Set. *Numerische Mathematik*, **8** (1966), 295-306.

[3] G. CAUWENBERGHS and T. POGGIO: Incremental and Decremental Support Vector Machine Learning. *Adv. Neural Information Processing Systems*, Cambridge MA, MIT Press, **13** (2001).

[4] R. DETRANO, A. JANOSI, W. STEINBRUNN, M. PFISTERER, J. SCHMID, S. SANDHU, K. GUPPY, S. LEE and V. FROELICHER: International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, **64** (1989), 304-310.

[5] R.O. DUDA and P.E. HART: Pattern Classification and Scene Analysis, John Wiley&Sons, New York, 1973.

[6] P.J. GREEN: Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives. *J. Roy. Statist. Soc.*, **B**(46), (1984), 149-192.

[7] R. HERBRICH, T. GRAEPEL and C. CAMPBELL: Bayes Point Machines. *Journal of Machine Learning Research*, **1** (2001), 245-279.

[8] Y.-C. HO and R.L. KASHYAP: An algorithm for linear inequalities and its applications. *IEEE Trans. Elec. Comp.*, **14** (1965), 683-688.

[9] Y.-C. HO and R.L. KASHYAP: A class of iterative procedures for linear inequalities. *J.SIAM Control*, **4** (1966), 112-115.

[10] P.J. HUBER: Robust Statistics. Wiley, New York, 1981.

[11] M.I. JORDAN and R.A. JACOBS: Hierarchical mixture of experts and the EM algorithm. *Neural Computations*, **6**(2), (1994), 181-214.

[12] P. MCCULLAGH and J.A. NELDER: Generalized linear models. Chapman and Hall, London, 1983.

[13] R.M. PALHARES and P.L.D. PERES: Robust Filtering with Guaranteed Energy-to-peak Performance – an LMI approach. *Automatica*, **36** (2000), 851-858.

[14] B.D. RIPLEY: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, 1996.

[15] J.T. TOU and R.C. GONZALEZ: Pattern Recognition Principles. Adison-Wesley, London, 1974.

[16] J.G. VANANTWERP and R.D. BRAATZ: A Tutorial on Linear and Bilinear Matrix Inequalities. *J. Proc. Cont.*, **10** (2000), 363-385.

[17] V. VAPNIK: An Overview of Statistical Learning Theory. *IEEE Trans. Neural Networks*, **10**(5), (1999), 988-999.

[18] V. VAPNIK: Statistical Learning Theory. Wiley, New York, 1998.

[19] A. WEBB: Statistical Pattern Recognition. Arnold, London, 1999.