A COMPARISON OF SOME APPROACHES TO THE RECOGNITION PROBLEMS IN CASE OF TWO CLASSES

Yurii I. Zhuravlev, Yuryi Laptin, Alexander Vinogradov, Aleksey Likhovid

Abstract: *We consider an improved model of the empirical risk minimization problem and its continuous relaxation. The continuous relaxation of the formulated problem is compared with the mathematical model used in the support vectors method. The results of numerical experiments comparing different models for problems with linearly inseparable sets are presented.*

Keywords: *cluster, decision rule, discriminant function, linear and nonlinear programming, nonsmooth optimization*

ACM Classification Keywords: *G.1.6 Optimization - Gradient methods, I.5 Pattern Recognition; I.5.2 Design Methodology - Classifier design and evaluation*

Introduction

Mathematical models of problems of constructing linear and non-linear classifiers and methods of constructing, based on these models, have been considered in many papers (see, e.g. [1-3]). In the present time the method of support vectors machine (SVM) is the most widely used.

For such problems it is convenient to represent mathematical models in the form of convex optimization problems. In [7] the technique using effective methods of non-smooth optimization for solving these problems was considered. The results of computational experiments were given for special large-scale test problems with linearly separable sets. A comparison was carried out with well-known program implementation LIBSVM of the method of support vector machine.

In this paper the models and approaches proposed in [4, 5] are further developed. We formulate an improved model of the empirical risk minimization problem and its continuous relaxation. The possibilities and complexity of the development of approximation algorithms to minimize the empirical risk are discussed. The continuous relaxation of the formulated problem is compared with the mathematical model used in the support vectors method. The results of numerical experiments comparing different models for problems with linearly inseparable sets are presented.

1. A brief description of problems of constructing classifiers

Let there be given a linear function $f(x,W)=\langle w,x\rangle+w_0$, where $x\in R^n$ is a vector of features, $W=(w,w_0)\in R^{n+1}$ is a vector of parameters. Function $a(x,W)$ of the following form is called *linear classifier*:

$$a(x,W)=\begin{cases}1, & \text{if } f(x,W)>0,\\ 2, & \text{if } f(x,W)\le 0.\end{cases} \qquad (2)$$

Classifier $a(x,W)$ refers each point $x\in R^n$ to one of the two classes of $\{1,\ 2\}$.

Consider a set of finite non-overlapping sets (training sample) that consists of points of R^n : $\Omega_i = \left\{ x^t : \ t \in T_i \right\}$, $i = 1,2$, $T = T_1 \bigcup T_2$.

The problem of constructing (training) classifier $a(x,W)$ is to determine the values of the parameters W based on the training sample Ω_i , $i = 1,2$.

It is said that the classifier $a(x,W)$ *correctly separates* the points of Ω_i , $i = 1,2$, if $a(x,W) = i$ for all $x \in \Omega_i$, $i = 1,2$. Classifier *gap* at a point x^t is the following value

$$g^t(W) = \begin{cases} f(x^t,W), \text{ if } t \in T_1, \\ -f(x^t,W), \text{ if } t \in T_2. \end{cases} \tag{4}$$

The value $g(W) = \min\left\{ g^t(W) : t \in T \right\}$ is called a gap of classifier $a(x,W)$ on the collection of sets Ω_i , $i = 1,2$. Classifier $a(x,W)$ correctly separates the points of the sets Ω_i , $i = 1,2$ if $g(W) > 0$.

The sets Ω_i , $i = 1,2$ are called separable in the class of linear classifiers, if there is a linear classifier, correctly separating the points of these sets.

Classifier $a(x,W)$ is invariant with respect to the multiplication function f (vector W) by a positive number, the gap $g(W)$ is linear with respect to this multiplication. The value of $g(W)$ can be used as a quality criterion for classifier $a(x,W)$ (the larger the value of $g(W)$, the more reliable the points of Ω_i , $i = 1,2$ are separeted), but we must also take into account some normalization of the vector W , which we denote $\eta(W)$ and will call the norm of the classifier $a(x,W)$.

We consider the problem of constructing an optimal classifier (determination of the values of parametres W) for the sets, Ω_i , $i = 1,2$, which are separable in the class of linear classifiers, in the following form: to find

$$g^* = \max_W \left\{ g(W) : \eta(W) \le 1, \ W \in R^{n+1} \right\} \tag{5}$$

As the norm of the vector W we use the function $\eta(W) = \sqrt{\sum_{j=1}^{n} (w_j)^2}$.

Problem (5) can be rewritten in the equivalent form

$$\eta^* = \min_V \left\{ \eta(V) : \ g(V) \ge 1, \ V \in R^L \right\} \tag{6}$$

$$\eta^* = \min_V \left\{ \eta(V) : \ g^t(V) \ge 1, t \in T, V \in R^L \right\} \tag{7}$$

This equivalence is understood in the sense that if W^* is an optimal solution of problem (5), then for optimal solutions V^* of (6) or (7) the equalities $V^* = W^* / g^*$, $\eta^* = 1 / g^*$ are satisfied [8]. Note that $g^* > 0$ for the sets which are separable in the class of linear classifiers.

2. Minimization of Empirical Risk

In the case of linearly inseparable samples the natural criterion for the choice of the classifier is the minimization of the empirical risk, i.e. number of points of training sample which the classifier separates incorrectly.

We assume that parameter $\delta > 0$ of the reliability of separating points of training sample Ω_i, $i = 1,2$ is given. The points x^t, $t \in T$ are separated by classifier $a(x,W)$ unreliably if the gap $g^t(W) < \delta$. Empirical risk with the reliability [5], defined by parameter δ, equals to the number of points of training sample, which the classifier separates incorrectly or unreliably.

The problem under consideration is to determine the minimum number of points which should be excluded from the training sample that the remaining points are separated reliably. It is natural to require that after excluding in each class at least one point is remained. This is possible if

$$\delta < \max\left\{\left\|x^\tau - x^s\right\| : \tau \in T_1,\ s \in T_2\right\} \tag{8}$$

Further we will assume that this condition is valid. It can be shown that there are sufficiently large positive numbers B_t, $t \in T$ (in [5] it was assumed that all B_t are the same) for which the empirical risk minimization problem with the reliability can be represented as the following: to find

$$Q^* = \min_{w,y}\left\{\sum_{t\in T} y_t\right\} \tag{9}$$

subject to constraints

$$g^t(W) \geq \delta - B_t \cdot y_t,\ \ t \in T \tag{10}$$

$$\langle w, w\rangle \leq 1 \tag{11}$$

$$\sum_{t\in T_i} y_t \leq |T_i| - 1,\ \ i = 1,2 \tag{12}$$

$$0 \leq y_t \leq 1,\ t \in T \tag{13}$$

$$y_t \in \{0,1\},\ t \in T \tag{14}$$

Variable y_t determines whether a point x^t is taken into account in the formulation of the problem. We say that numbers B_t, $t \in T$ satisfy the **correctness condition** if in case of $y_t = 1$ the point x^t is excluded from the training sammple, i.e. constraints (10) are satisfied for all feasible values of the other variables of the problem. Constraints (12) define the condition that at least one point from each set Ω_i should be included in the problem.

The problem (9) - (14) is NP-complete. In this regard, approximate algorithms for solving such problem must be developed for practical use. For small values of the problem dimension the existing general purpose optimization software can be used (the possibility of such approach will be considered in Section 4).

As approximate algorithms one can consider the algorithms based on the ideas of directed enumeration (sequential analysis of variants, the branch and bound methods), local search methods. Developing such algorithms it is essential to have effective procedures for calculating lower bounds for Q^* and the construction of feasible solutions of the problem (9)-(14). To implement these procedures we will use continuous relaxation of (9)-(14). It is clear that all integer formulations of the problem (9)-(14) for sufficiently large values B_t (satisfying the correctness condition) are equivalent. However, the continuous relaxation of the problem and the value of the lower bound for Q^* essentially depend on the values of B_t, since with increasing B_t the range of feasible solutions of continuous relaxation of the problem (9)-(14) is expanding. To obtain the best estimate for Q^* you must use the lowest possible values for B_t.

Let $t \in T$, $s \in T_1$, $\tau \in T_2$, $s, \tau \neq t$. Consider the problem

$$\beta_t^{s\tau} = \max\left\{\delta - g^t(W)\right\} \tag{15}$$

$$g^j(W) \geq \delta,\ j = s, \tau \tag{16}$$

$$\langle w, w \rangle \leq 1 \tag{17}$$

Denote

$$B_t^* = \max\left\{\beta_t^{s\tau} : s \in T_1, \tau \in T_2,\ s, \tau \neq t\right\},\ t \in T \tag{18}$$

Theorem 1. *Numbers* B_t, $t \in T$ *satisfy the correctness condition for problem (9) - (14) if*

$$B_t \geq B_t^*,\ t \in T \tag{19}$$

Proof. Let an index $t \in T$ be fixed. The point x^t is excluded from training sample in case of $y_t = 1$ when the constraint (10) for this index is valid for any feasible values of the remaining variables.

Denote $y = (y_\tau,\ \tau \in T)$, Y - the set of all y satisfying the constraints (12), (14), $D(y)$ - the set of all vectors W satisfying the constraints (10) and (11) for a given value of vector y. Consider the vector $y \in Y$ such that $y_t = 1$. Let

$$\beta_t(y) = \min\left\{\theta : g^t(W) \geq \delta - \theta,\ W \in D(y)\right\} = \max\left\{\delta - g^t(W) : W \in D(y)\right\}.$$

Denote $\beta_t^* = \max\left\{\beta_t(y) : y \in Y,\ y_t = 1\right\}$. It is evident that the inequality $B_t \geq \beta_t^*$ is the condition of exclusion of the point x^t from the training sample when $y_t = 1$. Let $s \in T_1$, $\tau \in T_2$, $s, \tau \neq t$. Denote $y^{s\tau} = \left(y_t, t \in T,\ y_s = 0,\ y_\tau = 0,\ y_j = 1,\ j \neq s, \tau\right)$. It is easy to see that for any $y \in Y$ such that $y_s = 0,\ y_\tau = 0$ $D(y) \subseteq D(y^{s\tau})$ is performed, i.e. $\beta_t(y) \leq \beta_t(y^{s\tau})$. Hence $\beta_t^* = \max\left\{\beta_t(y^{s\tau}) : s \in T_1, \tau \in T_2,\ s, \tau \neq t\right\}$. Taking into account that $\beta_t(y^{s\tau}) = \beta_t^{s\tau}$, i.e. $B_t^* = \beta_t^*$, we obtain the statement of the theorem.

Let $t \in T_1$. Consider in more detail the problem (15) - (17).Taking into account (4), we can rewrite this problem as

$$\beta_t^{s\tau} = -\min_{w, w_0}\left\{\left\langle w, x^t \right\rangle + w_0 - \delta\right\} \tag{20}$$

$$\left\langle w, x^s \right\rangle + w_0 \geq \delta,\ s \in T_1 \tag{21}$$

$$-\left\langle w, x^\tau \right\rangle - w_0 \geq \delta,\ \tau \in T_2 \tag{22}$$

$$\langle w, w \rangle \leq 1 \tag{23}$$

If the system of constraints (21) - (23) is inconsistent, then $\beta_t^{s\tau} = -\infty$. This occurs if $\delta > \left\| x^s - x^\tau \right\|$. By (8) there is always a pair s, τ such that $\delta \leq \left\| x^s - x^\tau \right\|$.

It is easy to see that in the optimal solution of problem (20) - (23) constraints (21), (23) must be satisfied as

equality, and constraint (22) can be either active or inactive. Consider the case when the constraint (22) is inactive in the optimal solution. Using the Lagrange multiplier rule, we obtain for optimal solutions $w = \frac{x^s - x^t}{\left\| x^s - x^t \right\|}$, $w_0 = \delta - \left\langle w, x^s \right\rangle$, $\beta_t^{s\tau} = \left\| x^s - x^t \right\|$. For the resulting vector $(w,\ w_0)$ constraint (22) should be satisfied. If this constraint is not satisfied, then the optimal solution should constructed on the fact that the constraint (22) is active. The obtained relations allow relatively easy to determine the values of B_t^*, $t \in T$.

Consider the problem (9)-(13) - the continuous relaxation of the problem of minimization of the empirical risk. Denote $d^t(W) = \max\left(0, \frac{1}{B_t}\left(\delta - g^t(W)\right)\right)$ and fix some values of the variables W. It is easy to see that if for these values W a solution of problem (9) - (13) exists, then $y^t = d^t(W)$. Hence we obtain the problem of minimization in the variables W: to find

$$q^* = \min_W \sum_{t \in T} d^t(W) \tag{24}$$

subject to

$$\left\langle w, w \right\rangle \le 1 \tag{25}$$

$$\sum_{t \in T_i} d^t(W) \le \left| T_i \right| - 1,\ i = 1, 2 \tag{26}$$

$$d^t(W) \le 1,\ t \in T \tag{27}$$

Value q^* is a lower bound for the minimum value of the empirical risk Q^* and the vector W obtained by solving the problem (24) - (27) defines an approximate solution of the problem (9) - (14). $d^t(W)$ - convex piecewise-linear functions. To solve the problem (24) - (27) it is appropriate to use effective methods of non-smooth optimization [6].

3. Method of Support Vector

In the method of support vectors (SVM) for the case $m = 2$ the following problem is solved: to find

$$\eta^* = \min_{v, v_0} \left\{ \left\langle v, v \right\rangle + C \sum_{t \in T} \xi^t \right\} \tag{28}$$

subject to

$$\left\langle v, x^t \right\rangle + v_0 \ge 1 - \xi^t,\ t \in T_1 \tag{29}$$

$$-\left\langle v, x^t \right\rangle - v_0 \ge 1 - \xi^t,\ t \in T_2 \tag{30}$$

$$\xi^t \ge 0, t \in T \tag{31}$$

The method of support vector (SVM) is used for finding an optimal classifier for linearly separable classes, and also for the classes which are linearly inseparable.

Note that constraints (29) and (30) correspond to the constraint $g^t(V) \ge 1,\ \ t \in T$. In the case of linearly separable classes it follows from theorems of non-smooth penalties (see, for example, [6]) that for a sufficiently

large value of C the problems (7) and (28) - (31) have the same solution. In the case of linearly inseparable classes the problem (28) - (31) is interpreted as some regularization of the empirical risk minimization problem. We will show that there are certain relationships between the problem (28) - (31) and the continuous relaxation (9) - (13) of empirical risk minimization problem.

Relax the constraints (10), setting $B_t = B := \max\limits_{\tau} B_\tau^*$ and exclude the constraint (12). We obtain the following problem:

$$\overline{q}^* = \min_{w,y} \left\{ \sum_{t \in T} y_t \right\} \tag{32}$$

subject to

$$\left\langle w, x^t \right\rangle + w_0 \geq \delta - B \cdot y_t, \ t \in T_1 \tag{33}$$

$$-\left\langle w, x^t \right\rangle - w_0 \geq \delta - B \cdot y_t, \ t \in T_2 \tag{34}$$

$$\langle w, w \rangle \leq 1 \tag{35}$$

$$y_t > 0, \ t \in T \tag{36}$$

Here, the constraint (11) is replaced by the equivalent pair of constraints (33) and (34). It is clear that $\overline{q}^* \leq q^*$. Let make a change of the variables $w = \delta v, \ w_0 = \delta v_0$, $\xi^t = \dfrac{B y_t}{\delta}$, $t \in T_1 \bigcup T_2$. The problem takes the form

$$\overline{q}^* = \frac{\delta}{B} \cdot \min_{v, v_0, \xi} \left\{ \sum_{t \in T} \xi^t \right\} \tag{37}$$

subject to

$$\left\langle v, x^t \right\rangle + v_0 \geq 1 - \xi^t, \ t \in T_1 \tag{38}$$

$$-\left\langle v, x^t \right\rangle - v_0 \geq 1 - \xi^t, \ t \in T_2 \tag{39}$$

$$\langle v, v \rangle \leq 1 \Big/ \delta^2 \tag{40}$$

$$\xi_t > 0, \ t \in T \tag{41}$$

Let $\alpha \geq 0$ be a dual variable for constraint (40). Consider the Lagrangian function $L(\alpha, \xi, v) = \dfrac{\delta}{B} \sum\limits_{t \in T} \xi^t + \alpha \cdot (\langle v, v \rangle - 1 \Big/ \delta^2)$ and Lagrangian relaxations of the problem (37) - (41): to find

$$\varphi(\alpha) = \min_{v, v_0, \xi} L(\alpha, \xi, u) \tag{42}$$

subject to (38), (39), (41).

Since $\varphi(\alpha)$ is the optimal value of the Lagrangian relaxation of (37) - (41), then $\varphi(\alpha) \leq \overline{q}^*$ for any $\alpha \geq 0$ (see, e.g., [6]). Given a penalty factor C in (28) - (31), choose α from the condition $\dfrac{\delta}{\alpha B} = C$. We obtain

$$L(\alpha, \xi, u) = \alpha \left\{ \langle v, v \rangle + C \cdot \sum_{t \in T} \xi^t \right\} - \frac{\alpha}{\delta^2},$$

i.e. the problem (42), (38), (39), (41) is equivalent to (28) - (31) with accuracy to an additive constant and a fixed factor in the objective function value for the above choice of the dual variable.

Thus, the problem (28) - (31), which is solved by the method of support vectors can be obtained as a result of relaxing constraints of (24) - (27), which in turn is a continuous relaxation of the problem of minimization of the empirical risk.

4. The results of numerical experiments

The quality of solutions obtained by using the empirical risk minimization model (9) - (14), the continuous relaxation of (9) - (13) and the model SVM (28) - (31) is compared in the computational experiments. Quality criterion is the error of classification - the number of training sample points that are classified incorrectly. The well-known software package CPLEX is used to solve the generated problems. Points in the training sample for each class were generated on the basis of a uniform distribution in the unit cube. These cubes are shifted relative to each other in the first coordinate so that the distance between them is 0.1. Family of linearly inseparable sets is formed iteratively, by moving at the current iteration a single point of each class to the opposite one.

Model (9) - (14) is NP-complete, so problems of low dimension were generated for numerical experiments. Fig. 1 shows the results for the case when $|\Omega_i| = 25,\ i = 1, 2$, ($\Omega_i$ - points of the training sample for the class i) n = 5 (n - dimension of the feature space R^n). For 25 iterations all points of a class move to the other, and vice versa. On X-axis the number of moved points of a class is indicated, the vertical axis - the classification error, MER - empirical risk minimization model (9) - (14), RMER - the relaxed model of minimization of the empirical risk (9) - (13). The complexity of the exact solution of the empirical risk minimization problems (9) - (14) for the family, shown in Figure 1, reached 90 min. Solving problems of larger dimension we obtained the messages of the package CPLEX for failure of computing resources.

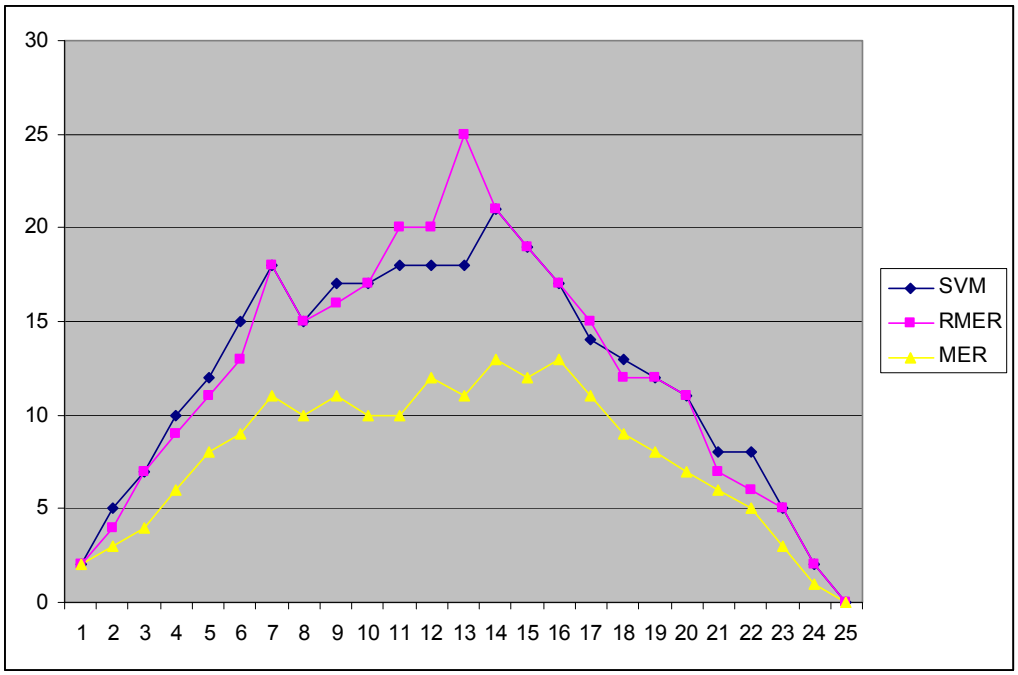

Figure 1. The dependence of the classification error on the number of displaced points n = 5, $|\Omega_i| = 25,\ i = 1, 2$.

In this regard, the comparison for the large-scale problems was realized only for the relaxed model of minimization of the empirical risk RMER and model SVM. Fig. 2 shows the results for the case $|\Omega_i| = 100,\ i = 1,2$, n = 30. It is essential to analyze the possibilities of the different models for the problems in which the value $|\Omega_i|,\ i = 1,2$ are significantly different. For this case it is necessary to estimate the value of the error of classification separately for each class. Fig. 3 shows the results for the case $|\Omega_1| = 30,\ |\Omega_2| = 200$, n = 30. The number of iterations for constructing a family of problems is 30.

Conclusion

The paper discusses various approaches to solving the problems of classification in the case of two classes. For linearly inseparable sets a mixed-integer model of the problem of minimization of the empirical risk and the continuous relaxation of the model are considered. It is shown that at weakened constraints of the proposed continuous relaxation the mathematical model used in the method of support vectors can be obtained.

The results of numerical experiments comparing approaches considered for the case of linearly inseparable sets are given. Classification error obtained by using the model of minimization of the empirical risk is much smaller than the error obtained when using continuous relaxation of this model and SVM method. This comparison was made for the problems of low dimension due to NP-completeness of the first model. Comparison of the second and the third models was also performed for large-scale problems. The resulting classification errors were about the same.

From the obtained results one can make a conclusion that it is appropriate to develop the approximate algorithms for solving the problem of minimization of the empirical risk based on the ideas of directed enumeration (sequential analysis of variants, branch-and-bound methods), local search methods, to improve the quality of generated classifiers.

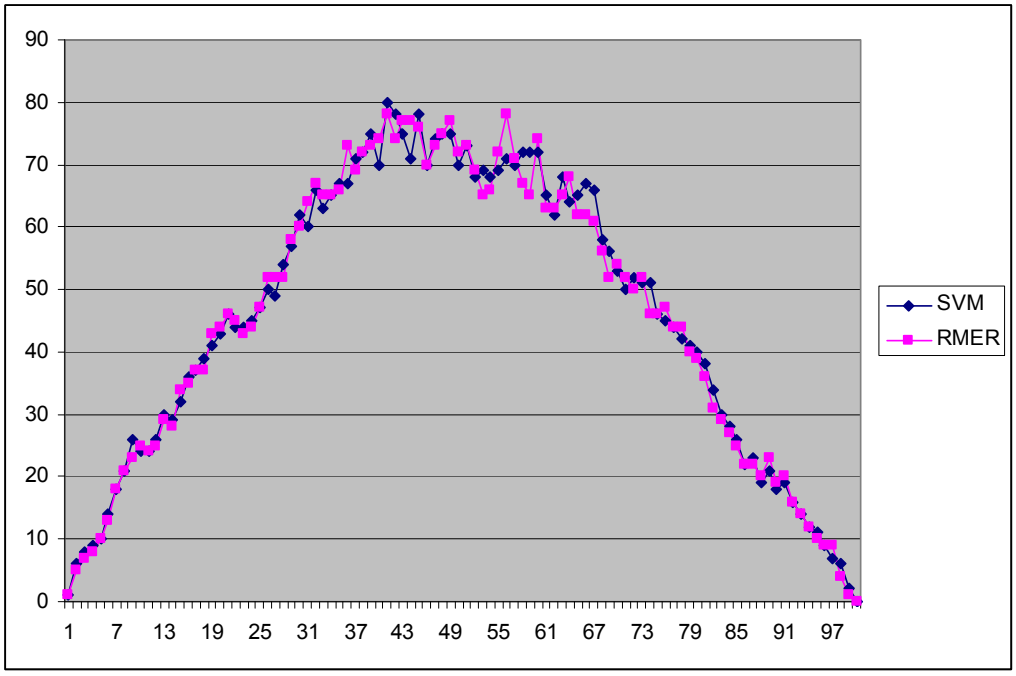

Figure 2. The dependence of the classification error on the number of displaced points n = 30, $|\Omega_i| = 100,\ i = 1,2$.

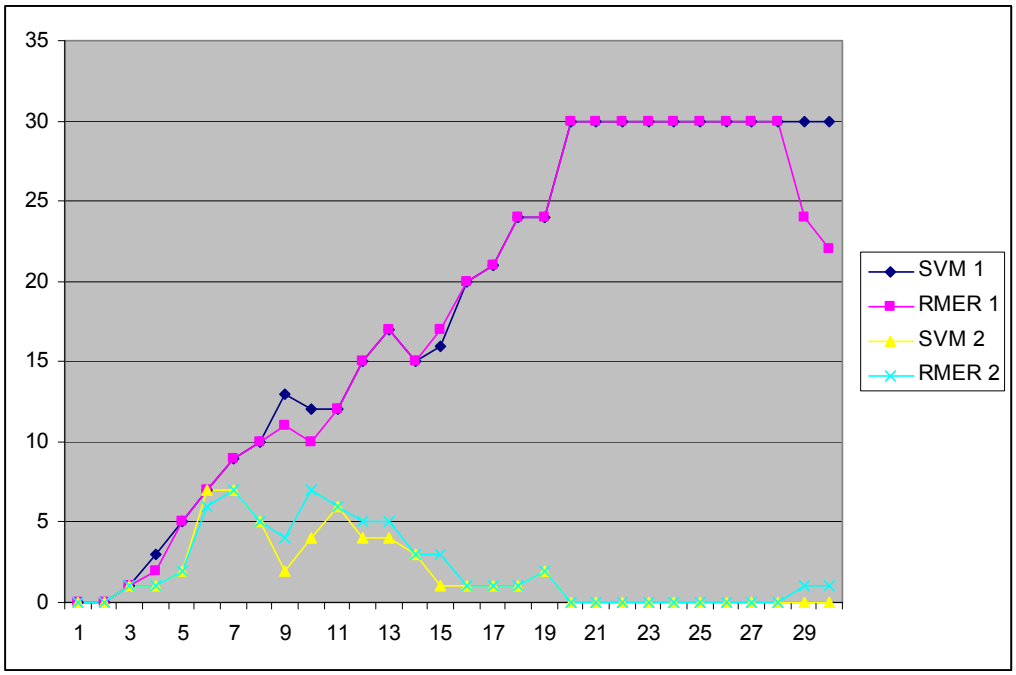

Figure 3. The dependence of classification error on the number of displaced points for each class, n = 30 $|\Omega_1| = 30$, $|\Omega_2| = 200$, SVM 1 - a model of SVM, set Ω_1, SVM 2 - a model of SVM, set Ω_2 , RMER 1 - the relaxed model of minimization of the empirical risk, set Ω_1 , RMER 2 model RMER, set Ω_2.

Bibliography

1. Vapnik V. Statistical Learning Theory. New York: Wiley, 1998.
2. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.
3. Thorsten Joachims. Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer, 2002
4. Zhuravlev Yu., Laptin Yu., A.Vinogradov Minimization of empirical risk in linear classifier problem // New Trends in Classification and Data Mining, ITHEA, Sofia, Bulgaria, 2010. – Pages 9-15
5. Журавлев Ю.И., Лаптин Ю.П., Виноградов А.П. Минимизация эмпирического риска и задачи построения линейных классификаторов // Кибернетика и системный анализ. 2011, № 4.- С. 155 – 164.
6. Shor N. Z. Nondifferentiable Optimization and Polynomial Problems. – Amsterdam / Dordrecht / London: Kluwer Academic Publishers, 1998. – 381 p.
7. Yurii I. Zhuravlev, Yuryi Laptin, Alexander Vinogradov, Nikolay Zhurbenko, Aleksey Likhovid. Nonsmooth optimization methods in the problems of constructing a linear classifier // Int Journal Information Models & Analyses (ISSN 1314-6416) 2012 Volume 1 Number 2 pp 103-111.
8. Laptin Yu. P., Likhovid A. P., and Vinogradov A. P. Approaches to Construction of Linear Classifiers in the Case of Many Classes // Pattern Recognition and Image Analysis, Vol. 20, No. 2, 2010, p. 137-145.

Authors' Information

Yurii I. Zhuravlev - *Academician of the RAS, Deputy Director, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: zhuravlev@ccas.ru*

Yuryi Laptin - *Senior Researcher, VMGlushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: laptin_yu_p@mail.ru*

Alexander Vinogradov - *Senior Researcher, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: vngrccas@mail.ru*

Aleksey Likhovid - *Researcher, VMGlushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: o.lykhovyd @ gmail.com*