Transactions on Machine Learning
and Data Mining
Vol. 2, No. 2 (2009) 65-79
© ISSN: 1865-6781 (Journal),
ISBN: 978-3-940501-12-7,
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Selection of Subsets of Ordered Features in Machine Learning

O. Seredin[1], A. Kopylov[1], V. Mottl[2]

[1] Tula State University, 300600, Tula, pr. Lenina, 92, Russia
oseredin@yandex.ru, And.Kopylov@gmail.com
[2] Computing Centre of the Russian Academy of Science,
117967, Moscow, Vavilova str., 40, Russia
vmottl@yandex.ru

Abstract. The new approach of relevant feature selection in machine learning is proposed for the case of ordered features. Feature selection and regularization of decision rule are combined in a single procedure. The selection of features is realized by introducing weight coefficients, characterizing degree of relevance of respective feature. *A priori* information about feature ordering is taken into account in the form of quadratic penalty or in the form of absolute value penalty on the difference of weight coefficients of neighboring features. Study of a penalty function in the form of absolute value shows computational complexity of such formulation. The effective method of solution is proposed. The brief survey of author's early papers, the mathematical frameworks, and experimental results are provided.

Keywords: machine learning, feature selection, ordered features, regularization of training, support vector machines, parametric dynamic programming

1 Introduction

The pattern recognition problem in the presence of a large amount of features (in comparing with training set size) known as the "curse of dimensionality". There are two standard approaches to tackle the case, namely, by *a priori* restrictions impositions (decision rule regularization) or dimensionality reduction by most informative features selection. The approach of joining these two techniques is proposed in the paper. The selection of informative features in pattern recognition

problem in the case of their ordering is considered. Feature ordering is typical for tasks of signal and image learning. Only one-dimensional ordering is accented in this work. Indeed, most of techniques for feature selection consider feature vector as non-ordered set of numbers, moreover a lot of methods accept hypothesis that features are independent. However, there is exists a number of tasks where features are consecutive measurements along the axis of some argument, for example, observation of some signal along time axis, components of a spectrum, etc.

In previous articles the authors already proposed methods of decision rule regularization and methods of feature selection. In the early papers [1,2] the method of regularization which takes into account *a priori* information about feature interrelation was described. At the same time the research of modality combination in data mining was developed, and actually the effective technique for informative feature selection was suggested [3-5]. Taking into account *a priori* information about one-dimensional ordering of features directly for a selection method requires development of modified procedures. Such attempt was done in [6] where the model of feature interrelation was represented as quadratic penalty on difference between the informative weights of neighbor features. In this paper we will investigate the new penalty criterion in the form of the absolute value function.

It should be noted, that the method of potential functions is chosen as a theoretical background for suggested algorithms. The reason for such selection is high popularity of the method as the basis for support vector machine learning [7].

The paper has the following structure – in the second section the idea of the support vector machine learning will be briefly reminded. In the third section the effective feature selection procedure regardless to their relationship will be described. The fourth section focuses on the idea of the learning regularization for the case of ordered features. The next three sections are devoted to algorithms of regularized order-based feature selection. Experimental results are presented in the eighth section.

2 SVM – the Basis for Modifications

We will develop methods of learning with respect to structural relations between features by insertion of additional penalties into existing well known criteria. The incorporated into the model coefficient of regularization will define the balance between "classical" approach and regularization based on *a priory* information.

Let $(\mathbf{x}_j, g_j)$, $j = 1,...,N$ – will be a training set, where $\mathbf{x} = (x_i, i = 1,...,n) \in \mathrm{R}^n$ – real-valued feature vector of recognition object, $g = \{\pm 1\}$ – index of classification; $\mathbf{a} = \{a_i, i = 1,...,n\} \in \mathrm{R}^n$ – directional vector of the optimal separable hyperplane and $b \in \mathrm{R}$ is its shift defined as decision of well-known criterion [7]:

$$\begin{cases} \sum_{i=1}^{n} a_i^2 + C\sum_{j=1}^{N} \delta_j \to \min(a_1,...,a_n, b, \delta_1,...,\delta_N), \\ g_j\left(\sum_{i=1}^{n} a_i x_{ij} + b\right) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1,...,N. \end{cases} \quad (1)$$

Here the non-negative parameter C and auxiliary variables δ_j, $j = 1,...,N$, introduced for the case of linear non-separability of objects of two classes. Usually the task (1) is solved in the dual form:

$$\begin{cases} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \left(g_j g_k \sum_{i=1}^{n} x_{ij} x_{ik} \right) \lambda_j \lambda_k \to \max(\lambda_1,...,\lambda_N), \\ \sum_{j=1}^{N} \lambda_j g_j = 0, \ 0 \le \lambda_j \le C/2, \ j = 1, \ldots, N, \end{cases} \quad (2)$$

as the task of quadratic programming related to the non-negative Lagrange multipliers λ_j, $j = 1,...,N$. The relation of coefficients of the directional vector of the optimal separable hyperplane and Lagrange multipliers is defined as follows:

$$a_i = \sum_{j:\lambda_j > 0} g_j \lambda_j x_{ij}, \ i = 1,...,n. \quad (3)$$

The simplicity of algorithmic and numerical realization of this method, the evident reference to the so-called support elements in the training set (only they, in fact, form the separable hyperplane), and good experimental results have made this formulation of the pattern recognition problem the most popular in recent times. These are the reasons for such criterion to be the basis for constructing our method of feature selection in the case of feature ordering or interrelation.

It is necessary to note that formulation (1) is most simple one, so to speak the academic one from the number of criteria, joined by the common title of support vector machines. In this form the solution based on inner products between feature vectors of objects. The decision rule is linear in the initial feature space. There are exist formulations with another type of kernels, another kind of penalties for non-separable cases. There are discussions in literature about relationship of SVM and method of potential functions [3], methods of featureless pattern recognition [5]. For clarity of our reasoning we will rely on the canonical formulation of the problem (1).

3 Feature Selection Based on Potential Functions Combining

The method of the potential functions (or kernels) combining in featureless pattern recognition and regression estimation was published in [3,4]. It has been shown that this technique can be transferred on the case where objects are presented by their features and can be efficiently applied as a non-iterative informative feature selection. The non-negative weights $r_i \ge 0, i = 1,...,n$, each of which is corresponds to component of the directional vector of the sought for separable hyperplane, are incorporated into the “classical” Vapnik’s (1) SVM criterion as it was proposed in [8]:

$$\sum_{i=1}^{n}\frac{a_i^2+1/\mu}{r_i}+\sum_{i=1}^{n}\left(\frac{1}{\mu}+1+\mu\right)\ln r_i+ \\ +C\sum_{j=1}^{N}\delta_j\to\min(r_1,...,r_n,a_1,...,a_n,b,\delta_1,...,\delta_N), \tag{4}$$

here μ is non-negative parameter of selectivity [9].

We propose to solve the problem of minimizing criterion (4) using the method of Gauss-Seidel by separating the variables into two groups: first – $a_i, i=1,...,n,\ b,\ \delta_j\geq 0,\ j=1,...,N$, and second $r_i\geq 0, i=1,...,n$, and implement step-by-step minimization criteria for one group of variables, with a fixed second one.

While coefficients $r_i, i=1,...,n$ are fixed the initial problem is actually reduced to the classical training problem by support vectors. The dual form of criterion related to non-negative Lagrange multipliers $\lambda_j,\ j=1,...,N$ almost coincide with criterion (2) of SVM:

$$\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(g_j g_k\sum_{i=1}^{n}r_i x_{ij}x_{ik}\right)\lambda_j\lambda_k\to\max(\lambda_1,...,\lambda_N). \tag{5}$$

The difference between the last criterion and "classical" formulation is the presence of additional coefficients $r_i, i=1,...,n$ in the matrix of quadratic form. The restrictions remain unchanged, while coefficients of the directional vector of the separable hyperplane are calculated using the rule: $a_i=r_i\sum_{j=1}^{N}g_j\lambda_j x_{ij},\ i=1,...,n$.

While parameters $a_i, i=1,...,n,\ b,\ \delta_j\geq 0,\ j=1,...,N$ are fixed the calculation of weighted coefficients is utterly simple:

$$r_i=\frac{a_i^2+(1/\mu)}{(1/\mu)+1+\mu},\ i=1,...,n. \tag{6}$$

The stopping rule of the iterative process of learning can be defined, for example, on the condition of convergence of sequences $r_i, i=1,...,n$: $\frac{1}{n}\sum_{i=1}^{n}\left|r_i^{step+1}-r_i^{step}\right|<\varepsilon$, $\varepsilon>0$.

It is necessary to make a reservation that introducing the notion of "informative feature" we do not have in mind the actual informational characteristic of feature like it was introduced for Akaike informational criterion or Shannon entropy criterion. We only suggest that for the whole set of measurable features there are exist subsets of features which adequate to either data analysis task. As a synonym of "informative feature" it is possible to consider term of "adequate feature" or "relevant feature".

4 The Regularization of Signal Recognition: the Principle of Decision Rule Smoothness

In early works [1] the approach to decision rule regularization was proposed by taking into account *a priori* information about the features ordering. To make the learning process prefer decision rules with smooth changing of coefficients of the directional vector of the separable hyperplane we propose to incorporate the additional quadratic penalty on difference of neighboring component to the criterion (1):

$$\sum_{i=1}^{n} a_i^2 + \alpha \sum_{i=2}^{n} (a_i - a_{i-1})^2 + C \sum_{j=1}^{N} \delta_j \to \min(a_1, ..., a_n, b, \delta_1, ..., \delta_N). \quad (7)$$

Here and further coefficient $\alpha \geq 0$ define the ratio of penalty on unsmoothness of ordering coefficients of the sought for optimal separable hyperplane. From the computational point of view both primal and dual tasks remain quadratic. The difference from the classical criterion is in incorporating the additional component $J'(\mathbf{a}) = \sum_{i=2}^{n} (a_i - a_{i-1})^2$. It is clear that such quadratic function can be written as $J'(\mathbf{a}) = \mathbf{a}^T \mathbf{B} \mathbf{a}$, where $\mathbf{B}(n \times n)$ has the following form:

$$\mathbf{B}(n \times n) = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

Therefore, it is more convenient to write the objective function in the problem of finding parameters of the optimal separable hyperplane (7) in the vector form:

$$\mathbf{a}^T (\mathbf{I} + \alpha \mathbf{B}) \mathbf{a} + C \sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \delta_1, ..., \delta_N),$$

under the same restrictions. It is obvious that matrix $\mathbf{B}$ is positive defined. The criterion in dual form also does not undergo changes, but the matrix of quadratic form will be slightly corrected:

$$\begin{cases} \sum_{j=1}^{N} \lambda_j - \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \left(g_j g_k \mathbf{x}_j^T (\mathbf{I} + \alpha \mathbf{B})^{-1} \mathbf{x}_k \right) \lambda_j \lambda_k \to \max(\lambda_1, ..., \lambda_N), \\ \sum_{j=1}^{N} \lambda_j g_j = 0, \qquad 0 \leq 0 \leq \lambda_j \leq C/2, \quad j = 1, ..., N. \end{cases}$$

The directional vector of the optimal separable hyperplane will be calculated by following formula: $\mathbf{a} = (\mathbf{I} + \alpha \mathbf{B})^{-1} \sum_{j: \lambda_j > 0} \lambda_j g_j \mathbf{x}_j$.

5 Selection of Subsets of Ordering Features

The feature selection procedure described in third section does not take into account the specificity of tasks of signal and image analysis. The classical formulation of the pattern recognition problem suppose that objects of recognition are represented by their features and it the order they where recorded does not matter. Roughly speaking, if somebody reorders components of feature vectors then the result of building of decision rule or the result of feature selection will not change. But we also draw attention to following fact: for some specific objects like signals and images peculiarity of their registration, namely, neighborhood of observations (samples/pixels) can be taken into account. Imposing such restrictions is called regularization (sometimes stabilizing) of decision rules of recognition. How to take into account such structural restrictions on a directional vector is shown in Section 4. In this section we will demonstrate how it is possible to join these two techniques: feature selection and assumption that on the set of features there are exist more or less informative groups.

Let modify criterion (4), namely, we will add extra penalty on difference of neighboring weighted coefficients $r_i \geq 0, i = 1,...,n$ under the previous constraints:

$$\sum_{i=1}^{n}\left[\frac{a_i^2 + 1/\mu}{r_i} + \left(\mu + 1 + \frac{1}{\mu}\right)\ln r_i\right] + \alpha\sum_{i=2}^{n} f(r_i, r_{i-1}) + \\ + C\sum_{j=1}^{N}\delta_j \to \min(r_1,...,r_n, a_1,...,a_n, b, \delta_1,...,\delta_N). \quad (\mathbf{8})$$

We propose to solve the problem of minimizing criterion (8) using the method of Gauss-Seidel by separating the variables into two groups: first – $a_i, i = 1,...,n$, b, $\delta_j \geq 0$, $j = 1,...,N$, and second $r_i \geq 0, i = 1,...,n$, and implement step-by-step minimization criteria for one group of variables, with a fixed second one. There is no difficulty to certain that if coefficients $r_i, i = 1,...,n$ are fixed than solution in dual form is coincide with task (5). But finding just informative weights would not be so simple as (6). Therefore, for the search of weight coefficients at the each step of coordinate-wise optimization it is necessary to find the minimum of following criterion (here, for short, we introduce new notions $c_i = a_i^2 + 1/\mu$, $i = 1,...,n$ and $d = \mu + 1 + 1/\mu$, remind that on this substep of iterative procedure values of a_i, $i = 1,...,n$ already found and fixed):

$$\sum_{i=1}^{n}\left[\frac{c_i}{r_i} + d\ln r_i\right] + \alpha\sum_{i=2}^{n} f(r_i, r_{i-1}) \to \min(r_1,...,r_n). \quad (\mathbf{9})$$

In the two next sections we will consider different ways of penalties on differences between weighted coefficients associated with neighboring ordered features of recognition object (for example, signal), namely we will consider penalties in the form of quadratic function and in the form of absolute value function. It is necessary to note, that in this approach the *a priori* information about feature ordering possess

restrictions on weighted coefficients of feature informativeness, but not on components of directional vector of the separable hyperplane as, for example, in [2].

6 Feature Subset Selection with Taking into Account Quadratic Difference between Neighboring Weight Coefficients

In this section we will consider situation where the penalty function is quadratic: $f(r_i, r_{i-1}) = \frac{(r_i - r_{i-1})^2}{r_i r_{i-1}}$, $i = 1,...,n$. In this case the criterion (9) will turn into:

$$\sum_{i=1}^{n}\left[\frac{c_i}{r_i} + d\ln r_i\right] + \alpha\sum_{i=2}^{n}\frac{(r_i - r_{i-1})^2}{r_i r_{i-1}} \to \min(r_1,...,r_n). \quad (\mathbf{10})$$

The search of minimum of (10) reduced to the solution of system of nonlinear equations for the parameters r_i, $i = 1,...,n$:

$$\begin{cases} -\frac{c_1}{r_1} + d + \alpha\left(\frac{r_1}{r_2} - \frac{r_2}{r_1}\right) = 0, \\ -\frac{c_i}{r_i} + d + \alpha\left(-\frac{r_{i-1} + r_{i+1}}{r_i} + \frac{r_i}{r_{i-1}} + \frac{r_i}{r_{i+1}}\right) = 0, \ i = 2,...,n-1, \\ -\frac{c_n}{r_n} + d + \alpha\left(-\frac{r_{n-1}}{r_n} + \frac{r_n}{r_{n-1}}\right) = 0. \end{cases} \quad (\mathbf{11})$$

Each equation in this system includes only 2-3 unknown variables. The method of simple iterations can be used to solve the problem.

7 Feature Subset Selection with Taking into Account Absolute Value of Difference between Neighboring Weight Coefficients

Numerous experiments have shown that taking into account the interrelation between features in the form of a quadratic penalty «dilutes» an informative subarea in the space of the ordered features. To avoid this disadvantage, it was decided to use the absolute value function as the penalty on difference of weight coefficients.

$$\sum_{i=1}^{n}\left[\frac{c_i}{r_i} + d\ln r_i\right] + \alpha\sum_{i=2}^{n}\left|\ln r_i - \ln r_{i-1}\right| \to \min(r_1,...,r_n). \quad (\mathbf{12})$$

The search algorithm for optimum values for the coefficients of the direction vector remains the same, but the minimization of criterion concerning weight factors

$r_i \geq 0, i = 1,...,n$, represents a new problem. Let us substitute variables: $u_i = \ln r_i, i = 1,\ldots n$, then the criterion (12) can be rewritten in the following form:

$$\sum_{i=1}^{n}\left[c_i e^{-u_i} + du_i\right] + \alpha\sum_{i=2}^{n}\left|u_i - u_{i-1}\right| \to \min(u_1,\ldots u_n). \quad (\mathbf{13})$$

Let us denote the functions of one variable in criterion (13) as $\psi_i(u_i) = c_i e^{-u_i} + du_i$, and functions of two variables as $\gamma_i(u_i, u_{i-1}) = \alpha\left|u_i - u_{i-1}\right|$. Then the objective function in criterion (13) takes more general form:

$$J(u_1, u_2, \ldots, u_n) = \sum_{i=1}^{n}\psi_i(u_i) + \sum_{i=2}^{n}\gamma_i(u_{i-1}, u_i) \quad (\mathbf{14})$$

The objective function (14) represents the sum of functions no more than two variables. We will call functions of this structure pair-wise separable. The pair-wise separability of the objective function (14) allows us to take advantages of the minimization procedure based on a principle of Dynamic Programming [11]. The procedure in this case is based on a recurrent decomposition of the initial problem of optimization of a function of n variables into a succession of n elementary problems, each of which consists in optimization of a function of only one variable. The elementary functions of one variable $\tilde{J}_i(u_i)$, to be minimized at each step of minimization of separable function are called here Bellman functions, as well as in the classical dynamic programming procedure.

The procedure of dynamic programming finds a global minimum of pair-wise separable function in two passes, at first in forward direction, and then in the backward direction.

On forward pass $i = 1,...,n$ the Bellman functions are determined in accordance with forward recurrent relation

$$\tilde{J}_i(u_i) = \psi_i(u_i) + \min_{u_{i-1}}\left[\gamma_i(u_{i-1}, u_i) + \tilde{J}_{i-1}(u_{i-1})\right],\ i = 1,...,n\ . \quad (\mathbf{15})$$

Last Bellman function $\tilde{J}_n(u_n)$ directly shows, how the minimum value of criterion, as a whole, depends on value of the variable u_n and therefore, its optimum value can be found as $\hat{u}_n = \arg\min_{u_n}\tilde{J}_n(u_n)$. Other elements of the sought for decision $\hat{u}_i, i = n-1,\ldots,1$, can be found by means of backward recurrent relation

$$\tilde{u}_{i-1}(u_t) = \arg\min_{u_{i-1}}\left[\gamma_i(u_{i-1}, u_i) + \tilde{J}_{i-1}(u_{i-1})\right], \quad (\mathbf{16})$$

which is the inverted form of forward recurrent relation (15).

Application of this relation on the backward move is obvious:

$$\hat{u}_{i-1} = \tilde{u}_{i-1}(u_t),\ i = n-1,\ldots,1 \quad (\mathbf{17})$$

Thus, regardless of the form of functions $\psi_i(u_i)$ and $\gamma_i(u_i, u_{i-1})$ in pair-wise separable objective function, the algorithm of the dynamic programming finds the

point of its global minimum, if, of course, such combination of values of variables exists within the area of their variation, executing the known number of operations, proportional to the number of variables.

In the case of continuous variables, e.g. if $u_i \in \mathrm{R}$, a numerical realization of the dynamic programming procedure, is possible only if there exists a finitely parameterized function family $\tilde{J}(u,\mathbf{q})$ concordant with node functions $\psi_i(u_i)$ and edge functions $\gamma_i(u_{i-1},u_i)$ in the sense that Bellman functions $\tilde{J}_i(u_i)$ belong to this family at each step. In this case, the forward pass of the procedure consists in a recurrent re-evaluating of parameters $\tilde{\mathbf{q}}_i$ that completely represent the Bellman functions $\tilde{J}_i(u) = \tilde{J}(u,\tilde{\mathbf{q}}_i)$. In particular, as is shown in [11], if the node and edge functions are quadratic, the Bellman functions will be quadratic too. The parametric representation is also possible in the case of using absolute value of difference of adjacent variables instead of quadratic node functions [12].

It can be easily proven, in such a case, that if the node functions $\psi_i(u_i)$ and edge functions $\gamma_i(u_{i-1},u_i)$ are convex, all the Bellman functions are also convex. As it is shown in [12], if the function $\gamma_i(u_{i-1},u_i)$ in the objective function (14) has the form $\gamma_i(u_i,u_{i-1}) = \alpha|u_i - u_{i-1}|$, $\alpha > 0$, and functions $\psi_i(u_i)$ are convex and everywhere differentiable in the range of definition, the procedure of dynamic programming can be rewritten in terms of recurrent recalculation of derivatives of the Bellman functions, and the forward recurrent relation (15) takes the form:

$$\tilde{J}'_i(u_i) = \psi'_i(u_i) + \begin{cases} -\alpha, u_i \le \tilde{u}_{i-1}^{-\alpha} \\ \tilde{J}'_{i-1}(u_i), \tilde{u}_{i-1}^{-\alpha} < u_i < \tilde{u}_{i-1}^{\alpha} \\ \alpha, u_i \ge \tilde{u}_{i-1}^{\alpha} \end{cases}, \qquad (\mathbf{18})$$

where $\tilde{u}_{i-1}^{-\alpha}$ and $\tilde{u}_{i-1}^{\alpha}$ can be obtained as the solution of equations

$\tilde{J}'_{i-1}(\tilde{u}_{i-1}^{-\alpha}) = \dfrac{d}{d\tilde{u}_{i-1}^{-\alpha}}[\tilde{J}_{i-1}(\tilde{u}_{i-1}^{-\alpha})] = -\alpha$ and $\tilde{J}'_{i-1}(\tilde{u}_{i-1}^{\alpha}) = \dfrac{d}{d\tilde{u}_{i-1}^{\alpha}}[\tilde{J}_{i-1}(\tilde{u}_{i-1}^{\alpha})] = \alpha$ respectively.

Then, backward recurrent relation (16) gets the simple form:

$$\tilde{u}_{i-1}(u_i) = \begin{cases} \tilde{u}_{i-1}^{-\alpha}, \tilde{u}_i \le \tilde{u}_{i-1}^{-\alpha}, \\ \tilde{u}_i, \tilde{u}_{i-1}^{-\alpha} < \tilde{u}_i < \tilde{u}_{i-1}^{\alpha}, \\ \tilde{u}_{i-1}^{\alpha}, \tilde{u}_i \ge \tilde{u}_{i-1}^{\alpha}. \end{cases} \qquad (\mathbf{19})$$

When parameters c_i and d are nonnegative, the functions $\psi_i(u_i) = c_i e^{-u_i} + du_i$ are convex. The derivative of the first Bellman function $\tilde{J}_1(u_1)$ is equal to the derivative of the function $\psi_1(u_1)$, i.e. $\tilde{J}'_1(u_1) = -c_1 \exp(-u_1) + d$. In accordance with the expression (18), the derivatives of the other Bellman functions are composed from the fragments of functions in the form of $q_k \exp(-u) + p_k$, where k is the number of the fragment. Therefore the boundaries of the fragments, as well as parameters q_k and p_k for each fragment k, constitute parameters of the Bellman function derivatives. The

leftmost boundary of the fragments coincides with $\tilde{u}_{i-1}^{-\alpha}$ and the rightmost boundary will coincides with $\tilde{u}_{i-1}^{\alpha}$.

Thus, for the objective function (13) there exists a parameterized Bellman functions family, concordant with node functions $\psi_i(u_i)$ and edge functions $\gamma_i(u_{i-1}, u_i)$, that makes it possible to use the non-iterative minimization procedure (15-17) on the basis of Dynamic Programming principle, described above.

1. $\tilde{J}_1'(u_1) = -c_1 \exp(-u_1) + d$, $\tilde{u}_1^{-\alpha} = -\ln[(d+\alpha)/c_1]$, $\tilde{u}_1^{\alpha} = -\ln[(d-\alpha)/c_1]$.
2. For $i = 2,\ldots n$,

$$\tilde{J}_i'(u_i) = -c_i \exp(-u_i) + d + \begin{cases} -\alpha, u_i \le \tilde{u}_{i-1}^{-\alpha} \\ \tilde{J}_{i-1}'(u_i), \tilde{u}_{i-1}^{-\alpha} < u_i < \tilde{u}_{i-1}^{\alpha} \\ \alpha, u_i \ge \tilde{u}_{i-1}^{\alpha} \end{cases};$$

$\tilde{u}_i^{\alpha}: \tilde{J}_i'(\tilde{u}_i^{\alpha}) = \alpha$, $\tilde{u}_i^{-\alpha}: \tilde{J}_i'(\tilde{u}_i^{-\alpha}) = -\alpha$.

3. $\tilde{u}_n: \tilde{J}_n'(\tilde{u}_n) = 0$.
4. For $i = n-1,\ldots 1$, $\tilde{u}_{i-1} = \begin{cases} \tilde{u}_{i-1}^{-\alpha}, \tilde{u}_i \le \tilde{u}_{i-1}^{-\alpha}, \\ \tilde{u}_i, \tilde{u}_{i-1}^{-\alpha} < \tilde{u}_i < \tilde{u}_{i-1}^{\alpha}, \\ \tilde{u}_{i-1}^{\alpha}, \tilde{u}_i \ge \tilde{u}_{i-1}^{\alpha}. \end{cases}$

One can easily see that the evaluated value of $\tilde{u}_{i-1}$ is completely defined by the value of adjacent variable u_i in the range of $\tilde{u}_{i-1}^{-\alpha} < u_i < \tilde{u}_{i-1}^{\alpha}$, and is independent of it at the rest of the value area of the variable u_i. It is just the fact that gives such a procedure the ability to preserve abrupt changes of the parameters, and accordingly not "dilute" an informative subarea in the space of the ordered features.

8 Experimental Results

8.1 Experiments on Model Data

For the experimental research of proposed algorithms test data were generated as described below. Two classes of recognition data were distributed near two centers. The centre of the first class is 100 artificial observations with values equal to zero. The centre of second class differs from the first one on interval from 70-th to 80-th samples. Second signal has values of 0.4 instead of 0 (Fig.1).

The examples of weight coefficients values $r_i, i = 1,...,100$ for the SVM, “pure” feature selection, and selection of feature subsets with penalties in form of quadratic function and absolute value function are shown in Fig. 2.

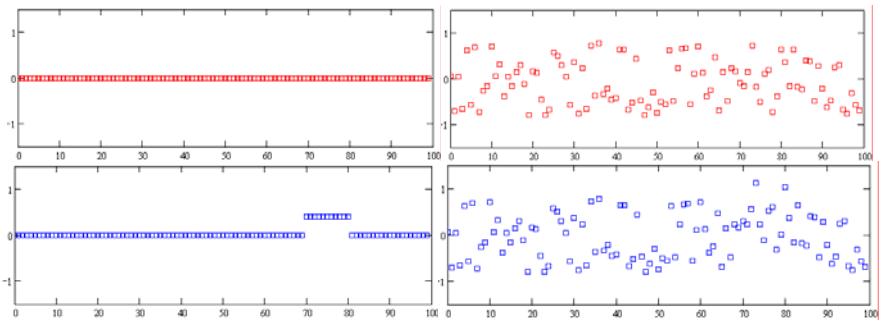

Fig. 1. Centers of first and second classes (left) and examples of recognition objects (right).

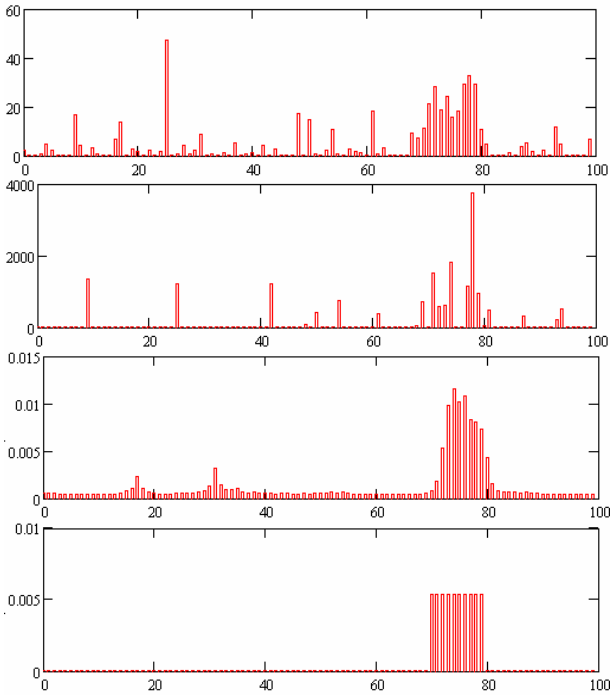

Fig. 2. Examples of weight coefficients values (from the top): SVM (1), SVM with feature selection (4), SVM and feature subset selection taking into account quadratic difference of neighbor weight coefficients (10), SVM and feature subset selection taking into account absolute value difference of neighbor weight coefficients (12).

Experimental results in the form of average error rate on test sets for different sizes of training sets (20-200 objects) are shown in Fig.3. It is clear that adding of the regularization, based on the search of informative subarea in feature space (criterion (10) – dotted line, criterion(12) – dashed line) lead to better predicted properties of decision rule as standard SVM (solid line).

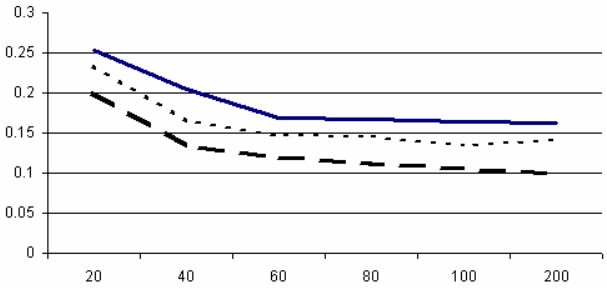

Fig. 3. Experimental results (error rate on test set vs. number of objects in training set).

8.2 Experiments on Real Data

In standard approaches of feature selection there is no take proper account of feature ordering in the tasks of signal or image analysis. The classical statement of pattern recognition theory is guesses that objects are represented by their features and the order of features is unimportant. Roughly speaking, if order of components in feature vector will be changed the results of decision rule making or feature selection will stay stable. Moreover, for example, in tasks of NIPS 2003 Feature Selection Challenge [14] features of datasets (even in "signal" tasks) were randomly reordered. The organizers explain this reordering of features by purity of competition experiment. This fact makes difficulties for finding real-world tasks for experiments. After enduring search the Data on cardiac Single Proton Emission Computed Tomography (SPECT Heart Data Set) [15] were chosen as experimental material. Files of data are available in UCI machine learning repository. Data are groups of 22 features measured on patients in rest and stress. Each of the patients is classified into two categories: normal and abnormal. Data set is divided on two subsets – training (87 objects) and test (80 objects). In [15] two-class task were investigated by CLIP3 algorithm and accuracy on test set of 84 percent were demonstrated. Examples of recognition objects of different classes are shown in Fig.4.

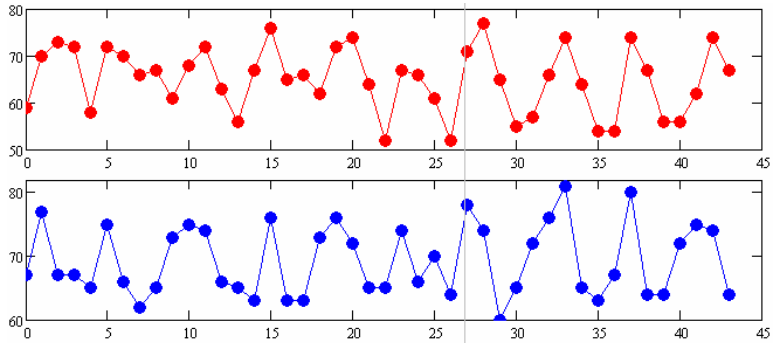

Fig. 4. Example of different recognition objects from SPECT database.

The results of applying proposed methods of selection of subset of features for SPECT data are demonstrated in Table 1.

Table 1. Accuracy for objects of test set for database SPECT

Algorithm	Accuracy, %
SVM	72.19
SVM+features selection	74.3
SVM+subsets of features selection (quadratic function)	77.4
SVM+subsets of features selection (absolute value function)	78.1

It is clear that using technique of subsets of features selection is improving predictive power decision rules; however we can't exceed best result of 84 percent of accuracy on the test set published in [15].

The choice of optimal value for the depth of regularization parameter α remains an open question. We used the procedure of cross validation to find the best value.

9 Conclusion

Article shows a way to combine relevant feature selection and restrictions on such selection, reasonable for the solved task, in one criterion. The summarizing review of our early publications, which actually lead to the proposed idea, has done. The basic idea of the proposed approach is to formalize idea of taking into account the one-dimensional ordering of features, which is typical for the tasks of signals analysis. The criterion is constructed and the scheme of its numerical optimization is offered. It is necessary to provide additional comprehensive analysis of the behavior of the proposed algorithm of selection of subset of relevant features in both modeling and real data experiments. It also seems reasonable to extend the methodology to the case of the two dimensional ordering, that is especially important for image analysis tasks.

Work is supported by grants of Russian Foundation for Basic Research 09-07-00394, 08-01-99003.

References

1. Seredin, O.S., Dvoenko, S.D., Krasotkina, O.V., Mottl, V.V: Machine Learning for Signal Recognition by the Criterion of Decision Rule Smoothness. Pattern Recognition and Image Analysis 11 (1), 87–90 (2001)
2. Seredin, O., Mottl, V.: Regularization in Image Recognition: the Principle of Decision Rule Smoothing. In: Proceedings of the Ninth International Conference Pattern Recognition and Information Processing, Minsk, Belarus, 2007. Vol. II., pp. 151-155 (2007)
3. Mottl, V.V., Seredin, O.S., Krasotkina, O.V., Muchnik, I.B:: Fusing of potential functions in reconstructing dependences from empirical data. Doklady Mathematics 71 (2), 315–319 (2005), From *Doklady Akademii Nauk*, Vol. 401, No. 5, 2005, pp. 607–612.
4. Mottl, V.V., Seredin, O.S., Krasotkina, O.V., Muchnik I.B.: Principles of multi-kernel data mining. In: Perner, P., Imiya, A. (Eds.) Machine Learning and Data Mining in Pattern Recognition. LNAI, vol. 3587, pp. 52 – 61, Springer Verlag (2005)
5. Mottl, V.V., Seredin, O.S., Dvoenko, S.D., Kulikowski, C.A, Muchnik, I.B.: Featureless pattern recognition in an imaginary Hilbert space. In: Proceedings of 16th International Conference Pattern Recognition, ICPR-2002, Quebec City, Canada, August, 2002, vol. II, pp.88-91 (2002)
6. Seredin, O., Mottl, V.: The selection of informative interrelated features in pattern recognition. Tavricheskiy Vestnik Informatiki i Matematiki, №2, 180–185 (2008) (in Russian).
7. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
8. Mottl, V., Tatarchuk, A., Sulimova, V., Krasotkina, O., Seredin, O.: Combining Pattern Recognition Modalities at the Sensor Level Via Kernel Fusion, In: Proceedings of 7th International Workshop Multiple Classifiers Systems, Prague, Czech Republic, 2007, pp. 1–12 (2007)
9. Tatarchuk, A., Mottl, V., Eliseyev, A., Windridge, D.: Selectivity Supervision in Combining Pattern-Recognition Modalities by Feature- and Kernel-Selective Support Vector Machines. IN: Proceedings of the 19th International Conference on Pattern Recognition, December 7-11, 2008, Tampa, Florida, USA (2008)
10. Seredin, O., Kopylov, A., Mottl, V., Pryimak, A.: Selection of subsets of interrelated features in pattern recognition problem. In: Proceedings of 9th International Conference "Pattern Recognition and Image Analysis: New Information Technologies", Nizhni Novgorod, 2008, Vol. 2, pp. 151-154 (2008)
11. Mottl, V., Kopylov, A., Blinov, A., Kostin, A.: Optimization techniques on pixel neighborhood graphs for image processing. Graph-Based Representations in Pattern Recognition. Computing, Supplement 12, pp. 135-145, Springer-Verlag Wien (1998)
12. Kopylov, A.V.: Parametric dynamic programming procedures for edge preserving in signal and image smoothing. In: Proceedings of the 7th International Conference on Pattern Recognition and Image Analysis, St.Petersburg October 18-23, 2004. Volume I, pp. 281-284 (2004)
13. Kopylov, A.V.: Dynamic programming for edge-preserving smoothing in signal and image analysis and pattern recognition with interrelated features. In: Proceedings of 9th International Conference "Pattern Recognition and Image Analysis: New Information Technologies", Nizhni Novgorod, 2008, Vol. 1, pp. 325-328 (2008)

14. Guyon, I., Li, J., Mader, T., Pletscher, P.A., Schneider, Uhr, M.: Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. Pattern Recognition Letters 28, 1438-1444 (2007)
15. Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M., Goodenday, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23 (2), 149-169 (2001)