6

Relaxed Linear Separability (RLS) Approach to Feature (Gene) Subset Selection

Leon Bobrowski[1,2] and Tomasz Łukaszuk[1]
[1]*Faculty of Computer Science, Białystok University of Technology,*
[2]*Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw*
Poland

1. Introduction

Feature selection is one of active research area in pattern recognition or data mining methods (Duda et al., 2001). The importance of feature selection methods becomes apparent in the context of rapidly growing amount of data collected in contemporary databases (Liu & Motoda, 2008).

Feature subset selection procedures are aimed at neglecting as large as possible number of such features (measurements) which are irrelevant or redundant for a given problem. The feature subset resulting from feature selection procedure should allow to build a model on the base of available learning data sets that generalizes better to new (unseen) data. For the purpose of designing classification or prediction models, the feature subset selection procedures are expected to produce higher classification or prediction accuracy.

Feature selection problem is particularly important and challenging in the case when the number of objects represented in a given database is low in comparison to the number of features which have been used to characterise these objects. Such situation appears typically in exploration of genomic data sets where the number of features can be thousands of times greater than the number of objects.

Here we are considering the *relaxed linear separability* (*RLS*) method of feature subset selection (Bobrowski & Łukaszuk, 2009). Such approach to feature selection problem refers to the concept of linear separability of the learning sets (Bobrowski, 2008). The term "relaxation" means here deterioration of the linear separability due to the gradual neglect of selected features. The considered approach to feature selection is based on repetitive minimization of the convex and piecewise-linear (*CPL*) criterion functions. These *CPL* criterion functions, which have origins in the theory of neural networks, include the cost of various features (Bobrowski, 2005). Increasing the cost of individual features makes these features falling out of the feature subspace. Quality the reduced feature subspaces is assessed by the accuracy of the CPL optimal classifiers built in this subspace.

The article contains a new theoretical and experimental results related to the RLS method of feature subset selection. The experimental results have been achieved through the analysis, inter alia, two sets of genetic data.

2. Linear separability of two learning sets

Suppose that m objects O_j described in the database are represented by feature vectors $\mathbf{x}_j[n] = [x_{j1},...,x_{jn}]^T$ ($j = 1,\ldots,m$). The feature vectors $\mathbf{x}_j[n]$ can be treated as points in the n-dimensional feature space $F[n]$ ($\mathbf{x}_j[n] \in F[n]$). The component x_{ji} of the vector $\mathbf{x}_j[n]$ is the numerical value of the i-th feature x_i of the object O_j. For example, in the case of clinical database, components x_{ji} can be the numerical results of the i-th diagnostic examinations of a given patient O_j.

Consider two learning sets G^+ and G^- built from n-dimensional feature vectors $\mathbf{x}_j[n]$. The *positive set* G^+ contains m^+ feature vectors $\mathbf{x}_j[n]$ and the *negative set* G^- contains m^- vectors $\mathbf{x}_j[n]$:

$$G^+ = \{\mathbf{x}_j[n]: j \in J^+\} \quad \textbf{\textit{and}} \quad G^- = \{\mathbf{x}_j[n]: j \in J^-\} \tag{1}$$

where J^+ and J^- are disjoined sets ($J^+ \cap J^- = \varnothing$) of indices j.

The positive set G^+ usually contains vectors $\mathbf{x}_j[n]$ of only one category. For example, the set G^+ may contain feature vectors $\mathbf{x}_j[n]$ representing patients with cancer and set G^- may represent patients without cancer.

Definition 1: The sets G^+ and G^- (1) are linearly separable, if and only if there exists such a weight vector $\mathbf{w}[n] = [w_1,...,w_n]^T$ ($\mathbf{w}[n] \in R^n$) and threshold θ ($\theta \in R$), that all the below inequalities are fulfilled:

$$\begin{aligned}(\exists\, \mathbf{w}[n], \theta\,)\ (\forall \mathbf{x}_j[n] \in G^+)\ \mathbf{w}[n]^T\mathbf{x}_j[n] > \theta \\ \textbf{\textit{and}}\ (\forall \mathbf{x}_j[n] \in G^-)\ \mathbf{w}[n]^T\mathbf{x}_j[n] < \theta\end{aligned} \tag{2}$$

The parameters $\mathbf{w}[n]$ and θ define the separating hyperplane $H(\mathbf{w}[n],\theta)$ in the feature space $F[n]$ ($\mathbf{x}[n] \in F[n]$):

$$H(\mathbf{w}[n],\theta) = \{\mathbf{x}[n]: \mathbf{w}[n]^T\mathbf{x}[n] = \theta\} \tag{3}$$

If the relations (2) are fulfilled, then all the elements $\mathbf{x}_j[n]$ of the set G^+ are located on the positive side of the hyperplane $H(\mathbf{w}[n],\theta)$ (3) and all the elements of the set G^- are located on the negative side of this hyperplane.

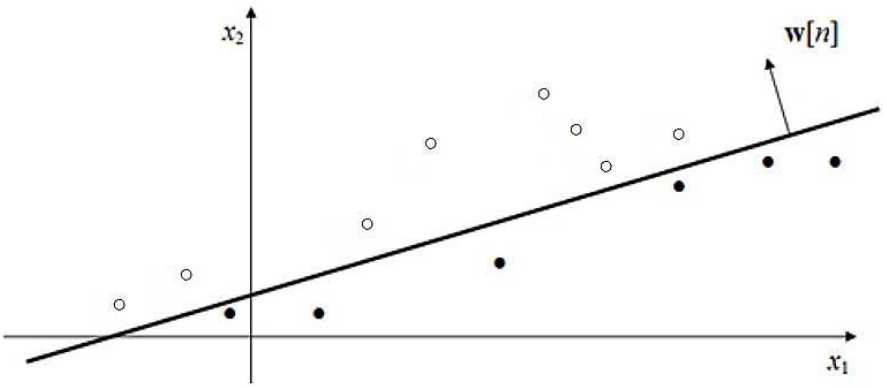


Fig. 1. An example of linearly separable sets G^+ (denoted by ○) and G^- (denoted by •) in the two-dimensional feature space F[2], where $m^+ = 8$ and $m^- = 6$

Lemma 1: Such sets G^+ and G^- (1) which are linearly separable (2) in the feature space $F[n]$, are also linearly separable in any greater feature space $F'[n']$, where $F[n] \subset F'[n']$.
The proof of the *Lemma* 1 is self-evident. The *Lemma* 1 shows, inter alia, that for any constant c the sets $G^+ = \{\mathbf{x}_j[n]: x_{ji'} > c\}$ and $G^- = \{\mathbf{x}_j[n]: x_{ji'} < c\}$ are linearly separable in each feature space $F[n]$.
Lemma 2: The sets G^+ and G^- (1) constructed of linearly independent feature vectors $\mathbf{x}_j[n]$ are always linearly separable (2) in the feature space $F[n]$.
The *Lemma* 2 can be proved by using arguments related to the construction of bases in the feature space $F[n]$ (Bobrowski, 2005). A base in the feature space $F[n]$ can be created by any n feature vectors $\mathbf{x}_j[n]$ which are linearly independent. Such n vectors $\mathbf{x}_j[n]$ can be separated by the hyperplane $H(\mathbf{w}[n],\theta)$ (3) for any subsets G^+ and G^- (1).
It can be seen that the linear separability (2) can be formulated equivalently to (2) as (Bobrowski, 2005):

$$\begin{array}{l}(\exists\, \mathbf{v}[n+1])\ (\forall\, \mathbf{y}_j[n+1] \in G^+)\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \geq 1 \\ \quad \textbf{\textit{and}}\ \ (\forall\, \mathbf{y}_j[n+1] \in G^-)\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \leq -1\end{array} \tag{4}$$

where $\mathbf{y}_j[n+1]$ are the *augmented feature vec*tors, and $\mathbf{v}[n+1]$ is the *augmented weight vector* (Duda et al., 2001):

$$(\forall j \in \{1,\ldots,m\})\ \mathbf{y}_j[n+1] = [\mathbf{x}_j[n]^T, 1]^T \textit{ and } \mathbf{v}[n+1] = [\mathbf{w}[n]^T, -\theta]^T \tag{5}$$

The inequalities (4) are used in the definition of the convex and piecewise-linear (*CPL*) penalty functions $\varphi_j^+(\mathbf{v}[n+1])$ and $\varphi_j^-(\mathbf{v}[n+1])$.

3. Convex and piecewise linear (*CPL*) criterion functions

Let us define the convex and piecewise-linear penalty functions $\varphi_j^+(\mathbf{v}[n+1])$ and $\varphi_j^-(\mathbf{v}[n+1])$ using the augmented feature vectors $\mathbf{y}_j[n+1]$ (5), and the weight vector $\mathbf{v}[n+1]$ (Bobrowski, 2005):

$$(\forall \mathbf{y}_j[n+1] \in G^+)\ \ \varphi_j^+(\mathbf{v}[n+1]) = \begin{array}{ll} 1 - \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] & \textbf{\textit{if}}\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] < 1 \\ 0 & \textbf{\textit{if}}\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \geq 1 \end{array}$$

and

$$(\forall \mathbf{y}_j[n+1] \in G^-)\ \ \varphi_j^-(\mathbf{v}[n+1]) = \begin{array}{ll} 1 + \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] & \textbf{\textit{if}}\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] > -1 \\ 0 & \textbf{\textit{if}}\quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \leq -1 \end{array}$$

The penalty function $\varphi_j^+(\mathbf{v}[n+1])$ is equal to zero if and only if the vector $\mathbf{y}_j[n+1]$ ($\mathbf{y}_j[n+1] \in G^+$) is situated on the positive side of the hyperplane $H(\mathbf{v}[n+1])$ (3) and is not too near to it (Fig. 2). Similarly, $\varphi_j^-(\mathbf{v}[n+1])$ is equal to zero if the vector $\mathbf{y}_j[n+1]$ ($\mathbf{y}_j[n+1] \in G^-$) is situated on the negative side of the hyperplane $H(\mathbf{v}[n+1])$ and is not too near to it (Fig. 3).
The perceptron criterion function $\Phi(\mathbf{v}[n+1])$ is defined on the sets G^+ and G^- (1) as the weighted sum of the penalty functions $\varphi_j^+(\mathbf{v}[n+1])$ and $\varphi_j^-(\mathbf{v}[n+1])$ (Bobrowski, 2005):

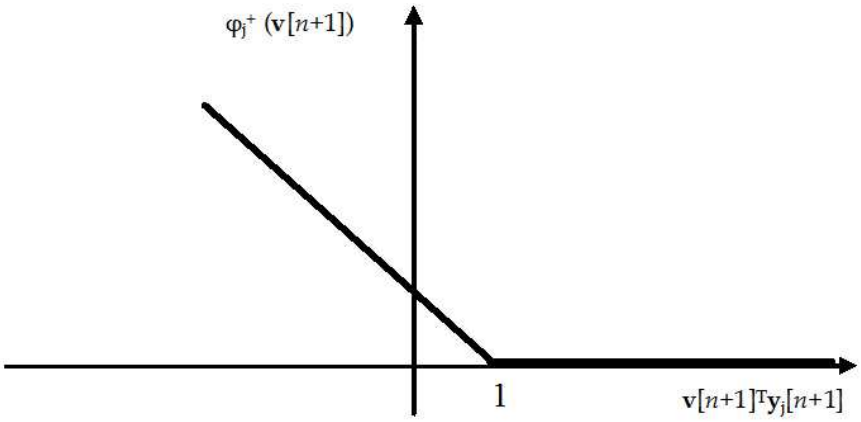


Fig. 2. The positive penalty function $\varphi_j^+(\mathbf{v}[n+1])$ (6).

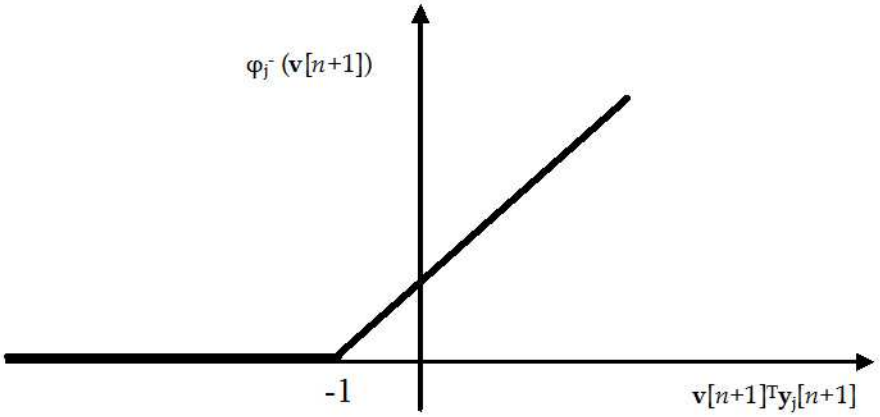


Fig. 3. The negative penalty function $\varphi_j^-(\mathbf{v}[n+1])$ (7).

$$\Phi(\mathbf{v}[n+1]) = \sum_{j \in J+} \alpha_j \varphi_j^+(\mathbf{v}[n+1]) + \sum_{j \in J-} \alpha_j \varphi_j^-(\mathbf{v}[n+1]) \tag{8}$$

where nonnegative parameters α_j determine *prices* of particular feature vectors $\mathbf{x}_j[n]$.
We are interested in the finding minimum $\Phi(\mathbf{v}_k^*[n+1])$ of the criterion function $\Phi(\mathbf{v}[n+1])$:

$$(\forall \mathbf{v}[n+1])\ \Phi(\mathbf{v}[n+1]) \geq \Phi(\mathbf{v}_k^*[n+1]) = \Phi^* \tag{9}$$

It has been proved that the minimal value Φ^* is equal to zero ($\Phi^* = 0$) if and only if the sets G^+ and G^- (1) are linearly separable (4) (Bobrowski, 2005).

$$(\Phi^* = 0) \Leftrightarrow (G^+ \text{ and } G^- \textit{ are linearly separable } (4)) \tag{10}$$

A modified *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ which includes additional penalty functions $\phi_i(\mathbf{v}[n+1])$ and the *costs* γ_i ($\gamma_i > 0$) related to particular features x_i has been introduced in order of the feature selection (Bobrowski, 2005):

$$(\forall i \in \{1,...,n\})\ \ \phi_i(\mathbf{v}[n+1]) = |w_i| = \begin{array}{ll} -\mathbf{e}_i[n+1]^T\mathbf{v}[n+1] & \textit{if}\quad \mathbf{e}_i[n+1]^T\mathbf{v}[n+1] < 0 \\ \mathbf{e}_i[n+1]^T\mathbf{v}[n+1] & \textit{if}\quad \mathbf{e}_i[n+1]^T\mathbf{v}[n+1] \geq 0 \end{array}$$

and

$$\Psi_\lambda(\mathbf{v}[n+1]) = \Phi(\mathbf{v}[n+1]) + \lambda \sum_{i \in I} \gamma_i \phi_i(\mathbf{v}[n+1]) \tag{12}$$

where λ ($\lambda \geq 0$) is the *cost level*, and $I = \{1,\ldots\ldots,n\}$.

Let us relate the hyperplane $h_j^+[n+1]$ in the parameter space R^{n+1} to each augmented feature vector $\mathbf{y}_j[n+1]$ (5) from the set G^+ (1), and the hyperplane $h_j^-[n+1]$ to each element $\mathbf{y}_j[n+1]$ (5) of the set G^-.

$$(\forall j \in J^+)\ h_j^+[n+1] = \{\mathbf{v}[n+1]: \mathbf{y}_j[n+1]^T\mathbf{v}[n+1] = 1\} \tag{13}$$

and

$$(\forall j \in J^-)\ h_j^-[n+1] = \{\mathbf{v}[n+1]: \mathbf{y}_j[n+1]^T\mathbf{v}[n+1] = -1\}$$

The first n unit vectors $\mathbf{e}_i[n+1] = [0,...,0,1,0,...,0]^T$ ($i = 1,...,n$) without the vector $\mathbf{e}_{n+1}[n+1] = [0,\ldots,0,1]^T$ are used in defining hyperplanes $h_i^0[n+1]$ in the augmented parameter space R^{n+1} (5):

$$(\forall i \in \{1,\ldots,n\})\ \ h_i^0[n+1] = \{\mathbf{v}[n+1]: \mathbf{e}_i[n+1]^T\mathbf{v} = 0\} = \{\mathbf{v}[n+1]: v_i = 0\} \tag{14}$$

The hyperplanes $h_j^+[n+1]$, $h_j^-[n+1]$ and $h_i^0[n+1]$ divide the parameter space R^{n+1} (5) in the disjoined regions $R_l[n+1]$. Each region $R_l[n+1]$ is a convex polyhedron in the parameter space with number of vertices $\mathbf{v}_k[n+1]$. The *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) is linear inside each region $R_l[n+1]$. It has been shown based on the theory of linear programming that the minimum of the *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) can be found in one of vertices $\mathbf{v}_k[n+1]$ of some region $R_l[n+1]$ (Bobrowski, 2005). Each vertex $\mathbf{v}_k[n+1]$ in the parameter space R^{n+1} is the intersection point of at lest $(n + 1)$ hyperplanes $h_j^+[n+1]$, $h_j^-[n+1]$ or $h_i^0[n+1]$. The below equations are fulfilled in each vertex $\mathbf{v}_k[n+1]$:

$$\begin{array}{l} (\forall j \in J_k^+)\ \mathbf{y}_j[n+1]^T\mathbf{v}_k[n+1] = 1, \textit{ and} \\ (\forall j \in J_k^-)\ \mathbf{y}_j[n+1]^T\mathbf{v}_k[n+1] = -1, \textit{ and} \\ (\forall i \in I_k^0)\ \mathbf{e}_i[n+1]^T\mathbf{v}_k[n+1] = 0 \end{array} \tag{15}$$

where $J_k^+$and J_k^- are the sets of indices j such hyperplanes $h_j^+[n+1]$, $h_j^-[n+1]$ (13) that pass through the vertex $\mathbf{v}_k[n+1]$, I_k^0 is the set of indices i such hyperplanes $h_i^0[n+1]$ (14) that pass through the vertex $\mathbf{v}_k[n+1]$.

The above equations can be given in the matrix form:

$$\mathbf{B}_k[n+1]\ \mathbf{v}_k[n+1] = \boldsymbol{\delta}_k'[n+1] \tag{16}$$

where $\mathbf{B}_k[n+1]$ is a non-singular matrix (*basis*) with the rows constituted by the linearly independent vectors $\mathbf{y}_j[n+1]$ ($j \in J_k^+ \cup J_k^-$) or the unit vectors $\mathbf{e}_i[n+1]$ ($i \in I_k^0$), and $\boldsymbol{\delta}_k'[n+1]$ is the *margin vector* with components equal to 1, -1 or 0 according to (15).

Remark 1: The number n_1 of the independent vectors $\mathbf{y}_j[n+1]$ in the matrix $\mathbf{B}_k[n+1]$ (16) can be not greater than the *rank r* of the data set $G^+ \cup G^-$ (1). So, the number n_0 of the unit vectors $\mathbf{e}_i[n+1]$ ($i \in I_k^0$) (15) in the basis $\mathbf{B}_k[n+1]$ (16) is not less than $n - r$ ($n_0 \geq n - r$).
The vertex $\mathbf{v}_k[n+1]$ can be computed by using the basis $\mathbf{B}_k[n+1]$ and the margin vector $\boldsymbol{\delta}_k'[n+1]$ (16):

$$\mathbf{v}_k[n+1] = \mathbf{B}_k[n+1]^{-1} \boldsymbol{\delta}_k'[n+1] \tag{17}$$

The criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12), similarly to the function $\Phi(\mathbf{v}[n+1])$ (8) is convex and piecewise-linear (*CPL*). The minimum of this function is located in one of the vertices $\mathbf{v}_k[n+1]$ (17):

$$(\exists \mathbf{v}_k\hat{}[n+1])\ (\forall \mathbf{v}[n+1])\ \Psi_\lambda(\mathbf{v}[n+1]) \geq \Psi_\lambda(\mathbf{v}_k\hat{}[n+1]) = \Psi_\lambda\hat{} \tag{18}$$

The basis exchange algorithms allow to find efficiently the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ constituting the minimum of the *CPL* function $\Psi_\lambda(\mathbf{v}[n+1])$ (12), even in the case of large, multidimentional data sets G^+ and G^- (1) (Bobrowski, 1991).
Remark 2: Such components w_{ki} of the vertex $\mathbf{v}_k[n+1] = [\mathbf{w}_k[n]^T, -\theta_k]^T = [w_{k1},...,w_{kn}, -\theta_k]^T$ (5) which are related to the unit vectors $\mathbf{e}_i[n+1]$ ($i \in I_k^0$) in the basis $\mathbf{B}_k[n+1]$ (16) are equal to zero ($w_{ki} = 0$) (15).
The n_0 features x_i ($i \in I_k^0$) (15) with the weights w_i equal to zero in the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) can be reduced without changing the separating hyperplane $H(\mathbf{w}_k\hat{}[n+1], \theta_k\hat{})$ (4). The following rule of *feature reduction* has been proposed on this base:

$$(\forall i \in I_k^0)\ \mathbf{e}_i[n+1]^T \mathbf{v}_k\hat{}[n+1] = 0 \Rightarrow w_i = 0 \Rightarrow \text{the feature } x_i \text{ is reduced} \tag{19}$$

Remark 3: A sufficiently large increase of the *cost level* λ ($\lambda \geq 0$) in the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) results in an increase of the number n_0 of unit vectors $\mathbf{e}_i[n+1]$ in the basis $\mathbf{B}_k\hat{}[n+1]$ (16) linked to the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) (Bobrowski, 2005).
An arbitrary number n_0 of features x_i can be omitted and the feature space $F[n]$ can be reduced to the subspace $F_k\hat{}[n - n_0]$ by using of adequate value λ_k of the parameter λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). For example, the value $\lambda = 0$ means that the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) constitutes the minimum of the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) defined in the full feature space $F[n]$. On the other hand, sufficiently large value of the parameter λ results in the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) equal to zero ($\mathbf{v}_k\hat{}[n+1] = \mathbf{0}$). Such solution is not constructive, because it means that all the features x_i have been reduced (19) and the separating hyperplane $H(\mathbf{w}[n], \theta)$ (3) cannot be defined.
For a given parameter value $\lambda = \lambda_k$ (12) the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) is determined unambiguously as the minimum (18) of the convex and piecewise linear function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). This vertex is characterized by the subset of such $n - n_0$ features x_i which are not related to the unit vectors $\mathbf{e}_i[n+1]$ ($i \notin I_k^0$) in the basis $\mathbf{B}_k\hat{}[n+1]$ (16) related to the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18). The feature subspace $F_k\hat{}[n_k] = F_k\hat{}[n - n_0]$ can be also determined by such $n - n_0$ features x_i. Quality of the feature subspace $F_k\hat{}[n_k]$ can be determined on the basis of the quality of the optimal linear classifier designed in this subspace of dimensionality n_k. The optimal feature subspace $F_k^*[n_k]$ can be identified as one that enables create the best linear classifier. The *RLS* method of feature subset selection is based on this scheme (Bobrowski, 2008; Bobrowski & Łukaszuk, 2009).
Comparing our approach with the approach based on the least-squares criterion, we can conclude that the discriminant function based on the least-squares criterion can be linked to

the Euclidean distance L_2, whereas our method based on the convex and piece-wise linear criterion function (*CPL*) can be linked to the L_1 norm distance function.

4. Characteristics of the optimal vertices in the case of linear separabilty

Let us consider the case of "*long vectors*" in the exploratory data analysis. In this case, the dimensionality n of the feature vectors $\mathbf{x}_j[n]$ is much greater than the number m ($n >> m$) of these vectors ($j = 1,\ldots, m$). We may expect in this case that the vectors $\mathbf{x}_j[n]$ are linearly independent (Duda et al., 2001). In accordance with the *Lemma* 2, the arbitrary sets G^+ and G^- (1) of linearly independent vectors $\mathbf{x}_j[n]$ are linearly separable (6). The minimal value Φ^* (9) of the criterion function $\Phi(\mathbf{v}[n+1])$ (8) defined on linearly separable sets G^+ and G^- (1) is always equal to zero ($\Phi^* = 0$) (Bobrowski,2005). The minimum $\Phi(\mathbf{v}_k^*[n+1])$ (9) of the function $\Phi(\mathbf{v}[n+1])$ (8) can be located in the optimal vertex $\mathbf{v}_k^*[n+1]$ (9), where the below equations hold (15):

$$\begin{gathered}(\forall j \in J_k^+)\ \mathbf{v}_k^*[n']^T\mathbf{y}_j'[n'] = 1 \\ \textit{and}\ (\forall j \in J_k^-)\ \mathbf{v}_k^*[n']^T\mathbf{y}_j'[n'] = -1\end{gathered} \tag{20}$$

where $n' = n - n_0$ is the dimensionality of the reduced feature vectors $\mathbf{y}_j'[n']$ obtained from $\mathbf{y}_j[n+1]$ (5) after neglecting n_0 features x_i related to the set I_k^0 (15) and $\mathbf{v}_k^*[n']$ is the reduced vertex obtained from $\mathbf{v}_k^*[n+1]$ (9) by neglecting n_0 components w_i equal to zero ($w_i = 0$).
The vectors $\mathbf{y}_j'[n']$ belong to the reduced feature subspace $F_k[n']$ ($\mathbf{y}_j'[n'] \in F_k[n']$). We can remark that if the learning sets $G^+[n']$ and $G^-[n']$ constituted from the vectors $\mathbf{y}_j'[n']$ are linearly separable (4) in a given feature subspace $F_k[n']$, there may be more than one optimal vertex $\mathbf{v}_k^*[n']$ creating the minimum (9) of the function $\Phi_k(\mathbf{v}[n'])$ (8) ($\Phi_k(\mathbf{v}_k^*[n']) = 0$). In this case, each optimal vertex $\mathbf{v}_k^*[n']$ linearly separates (4) the sets $G^+[n']$ and $G^-[n']$ (Bobrowski, 2005):

$$\begin{gathered}(\forall \mathbf{y}_j'[n'] \in G^+[n'])\ \mathbf{v}_k^*[n']^T\mathbf{y}_j'[n'] \geq 1 \\ \textit{and}\ (\forall \mathbf{y}_j'[n'] \in G^-[n'])\ \mathbf{v}_k^*[n']^T\ \mathbf{y}_j'[n'] \leq -1\end{gathered} \tag{21}$$

Moreover, in the case of "*long vectors*" there may exist many such feature subspaces $F_k[n']$ of a given feature space $F[n]$ ($F_k[n'] \subset F[n]$) which can assure the linear separability (21). Therefore, a question arises which of the vertices $\mathbf{v}_k^*[n']$ constituting the minimum (9) of the perceptron function $\Phi(\mathbf{v}[n+1])$ (8) is the best one.
The answer for a such question can be given on the basis of minimization of the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). In contrary to the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) has only one optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (16). The vertex $\mathbf{v}_k\hat{}[n+1]$ (16) which constitutes minimum (18) of the function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) is unambiguously determined and can be treated as the optimal one.
It can be proved that the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) with a sufficiently small cost level λ ($\lambda \geq 0$), has the minimal value (18) in the same vertex $\mathbf{v}_k^*[n+1]$ (9) as the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) (Bobrowski, 2005):

$$(\exists\, \lambda_{max})\ (\forall \lambda \in (0, \lambda_{max}))\ (\forall \mathbf{v}[n+1])\ \Psi_\lambda(\mathbf{v}[n+1]) \geq \Psi_\lambda(\mathbf{v}_k^*[n+1]) \tag{22}$$

In other words, the replacement of the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) by the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) does not necessarily mean changing the position of the minimum.

The modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) can be expressed in the following manner for such points $\mathbf{v}[n+1]$ which separate linearly (4) the sets G^+ and G^- (1):

$$\Psi_\lambda'(\mathbf{v}[n+1]) = \lambda \sum_{i \in I} \gamma_i \phi_i(\mathbf{v}[n+1]) = \lambda \sum_{i \in I} \gamma_i |v_{ki}| \tag{23}$$

Therefore, the minimization of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) can be replaced by the minimization of the function $\Psi_\lambda'(\mathbf{v}[n+1])$ (23) under the condition that the point $\mathbf{v}[n+1]$ linearly separates (4) the sets G^+ and G^- (1).

Remark 5: If the sets G^+ and G^- (1) are linearly separable, then the vertex $\mathbf{v}_k^*[n+1]$ constituting the minimum of the function $\Psi_\lambda'(\mathbf{v}[n+1])$ (23) with equal feature costs γ_i has the lowest L_1 norm $||\mathbf{v}_k^*[n+1]||_{L1} = \Sigma_i |v_{ki}|$ among all such vectors $\mathbf{v}[n+1]$ which linearly separate (4) these sets.

The *Remark* 5 points out a possible similarity between the *CPL* solution $\mathbf{v}_k^*[n+1]$ (22) and the optimal vector $\mathbf{v}^*[n+1]$ obtained in the *Support Vector Machines* (*SVM*) approach (Vapnik, 1998). But the use of the *CPL* function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) also allows obtain other types of solutions $\mathbf{v}_k^*[n+1]$ (22) by another specification of feature costs γ_i and the cost level λ parameters. The modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) gives possibility to introduce different feature costs γ_i ($\gamma_i > 0$) related to particular features x_i. As a result, the outcome of feature subset selection process can be influenced by the feature costs γ_i (12).

5. Relaxed linear separability (*RLS*) approach to feature selection

The initial *feature space* $F[n]$ ($\mathbf{x}_j[n] \in F[n]$) is composed of the all n features x_i from a given set $\{x_1, \ldots x_n\}$. Feature reduction rule (19) results in appearance of the feature subspaces $F_k[n_k]$ ($F_k[n_k] \subset F[n]$ and $n_k < n$).

Successive increase of the value of the cost level λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) allows to reduce (19) additional features x_i and, as a result, allows generate the descended sequence of feature subspaces $F_k[n_k]$:

$$F[n] \supset F_1[n_1] \supset F_2[n_2] \supset \ldots \supset F_{k'}[n_{k'}], \text{ where } n_k > n_{k+1} \tag{24}$$

The sequence (24) of the feature subspaces $F_k[n_k]$ is generated in a deterministic manner on the basis data sets G^+ and G^- (1) in accordance with the relaxed linear separability (*RLS*) method (Bobrowski & Łukaszuk, 2009). Each step $F_k[n_k] \rightarrow F_{k+1}[n_{k+1}]$ is realized by an adequate increase $\lambda_k \rightarrow \lambda_{k+1} = \lambda_k + \Delta_k$ (where $\Delta_k > 0$) of the cost level λ in the criterion function $\Psi_\lambda(\mathbf{w}[n],\theta)$ (12).

One of the problems in applying the *RLS* method is to assess the quality characteristics of successive subspaces $F_k[n_k]$ (24). In this approach, a quality of a given subspace $F_k[n_k]$ is evaluated on the basis of the optimal linear classifier designed in this subspace. The better optimal linear classifier means the better feature subspace $F_k[n_k]$.

The feature subspace $F_k[n_k]$ can be obtained from the initial feature space $F[n]$ by reducing the $n - n_k$ features x_i. Such reduction can be based on the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) with

the related basis $\mathbf{B}_k\hat{}[n+1]$ (16). The optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) appoints the minimum of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) with the adequate value λ_k of the cost level λ.

Definition 2: The *reduced feature vectors* $\mathbf{y}_j'[n_k]$ ($\mathbf{y}_j'[n_k] \in F_k[n_k]$) are obtained from the feature vectors $\mathbf{y}_j[n+1] = [\mathbf{x}_j[n]^T,1]^T$ (5) after neglecting $n - n_k$ features x_i related to the set I_k^0 (15) of the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18). The *reduced vertex* (parameter vector) $\mathbf{v}\hat{}[n_k]=[\mathbf{w}\hat{}[n_{k-1}]^T,-\theta\hat{}]^T$ (5) and $\mathbf{v}_k\hat{}[n_k]$ is obtained from the optimal vertex $\mathbf{v}_k\hat{}[n+1]$ (18) by neglecting of these $n - n_k$ components w_i, which are equal to zero ($w_i = 0$).

The reduced parameter vector $\mathbf{v}[n_k] = [\mathbf{w}[n_{k-1}]^T,-\theta]^T$ (5) defines the linear classifier $LC(\mathbf{v}[n_k])$ in the feature subspace $F_k[n_k]$. The linear classifier $LC(\mathbf{v}[n_k])$ can be characterized by the following decision rule:

$$\begin{aligned} &\textbf{if } \mathbf{v}[n_k]^T \mathbf{y}'[n_k] \geq 0, \textbf{ then } \mathbf{y}'[n_k] \textit{ is allocated to the category } \omega^+ \\ &\textbf{if } \mathbf{v}[n_k]^T \mathbf{y}'[n_k] < 0, \textbf{ then } \mathbf{y}'[n_k] \textit{ is allocated to the category } \omega^- \end{aligned} \tag{25}$$

where $\mathbf{y}'[n_k] \in F_k[n_k]$, and the category (*class*) ω^+ is represented by elements $\mathbf{x}_j[n]$ of the learning set G^+ (1) and the category ω^- is represented by elements of the set G^-.

Definition 3: The *CPL optimal* linear classifier $LC(\mathbf{v}^*[n_k])$ is defined in the feature subspace $F_k[n_k]$ by a reduced parameter vector $\mathbf{v}^*[n_k]$ that constitutes the minimum $\Phi^* = \Phi_k(\mathbf{v}^*[n_k])$ (9) of the perceptron criterion function $\Phi_k(\mathbf{v}[n_k])$ (8).

The perceptron criterion function $\Phi_k(\mathbf{v}[n_k])$ is defined (8) on reduced feature vectors $\mathbf{y}_j'[n_k]$ ($\mathbf{y}_j'[n_k] \in F_k[n_k]$) that belong to the reduced learning set $G^+[n_k]$ or $G^-[n_k]$ (1).

$$G^+[n_k] = \{\mathbf{y}_j'[n_k]: j \in J^+\} \textbf{ and } G^-[n_k] = \{\mathbf{y}_j'[n_k]: j \in J^-\} \tag{26}$$

Remark 6: The minimal value Φ_k^* of the criterion function $\Phi_k(\mathbf{v}[n_k])$ (8) on reduced feature vectors $\mathbf{y}_j'[n_k]$ is equal to zero ($\Phi_k^* = 0$) if and only if the sets $G^+[n_k]$ and $G^-[n_k]$ are linearly separable (4) in the feature subspace $F_k[n_k]$ (similarly as (10)) (Bobrowski, 2005).

It has been proved that, if the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) are linearly separable (4), then the decision rule (25) based on the optimal vector $\mathbf{v}_k^*[n_k]$ (9) allocates correctly all elements $\mathbf{y}_j'[n_k]$ of these learning sets (Bobrowski, 2005). It means that (21):

$$\begin{aligned} &(\forall \mathbf{y}_j'[n_k] \in G^+[n_k])\ \mathbf{v}^*[n_k]^T \mathbf{y}_j'[n_k] > 0, \textbf{ and} \\ &(\forall \mathbf{y}_j'[n_k] \in G^-[n_k])\ \mathbf{v}^*[n_k]^T \mathbf{y}_j'[n_k] < 0 \end{aligned} \tag{27}$$

If the sets $G^+[n_k]$ and $G^-[n_k]$ (26) are not linearly separable (4), then not all but only a majority of the vectors $\mathbf{y}_j'[n_k]$ fulfil the above inequalities.

According to the considerations of the previous paragraph, if the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) are linearly separable (4), then there is more than one vertex $\mathbf{v}_i^*[n_k]$ forming a minimum of the function $\Phi_k(\mathbf{v}[n_k])$ (8). To avoid such ambiguity, the criterion function $\Phi_k(\mathbf{v}[n_k])$ (8) can be replaced by the modified criterion function $\Psi_{k\lambda}(\mathbf{v}[n_k])$ (12) with the small value (22) of the parameter λ.

6. Evaluation of linear classifiers

The quality of the linear classifier $LC(\mathbf{v}^*[n_k])$ (25) can be evaluated by using the error estimator (*apparent error rate*) $e_a(\mathbf{v}^*[n_k])$ as the fraction of wrongly classified elements $\mathbf{y}_j'[n_k]$ of the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26):

$$e_a(\mathbf{v}^*[n_k]) = m_e(\mathbf{v}^*[n_k]) \;/\; m \tag{28}$$

where m is the number of all elements $\mathbf{y}_j'[n_k]$ of the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) $\mathbf{x}_j[n]$, and $m_e(\mathbf{v}^*[n_k])$ is the number of elements $\mathbf{y}_j'[n_k]$ wrongly allocated by the rule (25).

The parameters $\mathbf{v}^*[n_k]$ of the linear classifier $LC(\mathbf{v}^*[n_k])$ (25) are estimated from the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) through minimization of the perceptron criterion function $\Phi_k(\mathbf{v}[n_k])$ (8) determined on elements $\mathbf{y}_j'[n_k]$ of these sets. It is known that if the same data $\mathbf{y}_j'[n_k]$ is used for classifier designing and for classifier evaluation, then the evaluation results are too optimistic (*biased*). The error rate (28) evaluated on the elements $\mathbf{y}_j'[n_k]$ of the learning sets is called the *apparent error* (*AE*). For example, if the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) are linearly separable (4), then the relation (27) holds and, as a result, the *apparent error* (28) evaluated on elements $\mathbf{y}_j'[n_k]$ is equal to zero ($e_a(\mathbf{v}^*[n_k]) = 0$). But it is observed in practice that the error rate of the classifier (25) evaluated on new vectors $\mathbf{y}'[n_k]$ is usually greater than zero.

For the purpose of the classifier's bias reducing, the cross validation procedures are applied (Lachenbruch, 1975). The term *p-fold cross validation* means that the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) have been divided into p parts G_i, where $i = 1,\ldots,p$ (for example $p = 10$). The vectors $\mathbf{y}_j'[n_k]$ contained in $p - 1$ parts G_i are used for definition of the criterion function $\Phi_k(\mathbf{v}[n_k])$ (8) and computing of the parameters $\mathbf{v}^*[n_k]$. The remaining vectors $\mathbf{y}_j'[n_k]$ are used as the *test set* (one part $G_{i'}$) for computing (evaluation) the error rate $e(\mathbf{v}^*[n_k])$ (28). Such evaluation is repeated p times, and each time different part $G_{i'}$ is used as the test set. The cross valid1ation procedure allows to use different vectors $\mathbf{y}_j'[n_k]$ (1) for the classifier (25) designing and evaluation (28) and as a result, to reduce the bias of the error rate estimation (28). The error rate (28) estimated during the *cross validation* procedure will be called the *cross-validation error* (*CVE*).

The *CVE error rate* $e_{CVE}(\mathbf{v}^*[n_k])$ (28) of the linear classifier (25) is used in the relaxed linear separability (*RLS*) method as a basic criterion in evaluation of particular feature subspaces $F_k[n_k]$ in the sequence (24) (Bobrowski & Łukaszuk, 2009). Feature subspace $F_k[n_k']$ that is linked to the linear classifier $LC(\mathbf{v}^*[n_k])$ (25) with the lowest *CVE error rate* $e_{CVE}(\mathbf{v}^*[n_k])$ can be considered as the optimal one in accordance with the *RLS* method of feature selection.

7. Toy example

The data set used in the experiment was generated by the authors. In the two-dimensional space seven points were selected. Four of them were assigned to the positive set G^+, three to the negative set G^-. The allocation of points to the sets G^+ and G^- were made in a way that the linear separability of sets was preserved. After that each point was extended to 10 dimensions. The values the remaining coordinates were drawn from the distribution N(0,1). Table 1 contains the complete data set. Features x_2 and x_7 constitute the coordinates of points in the initial two-dimensional space.

Previously described the *RLS* method was applied to the data set presented in Table 1. Table 2 shows a sequence of feature subsets studied by the method and values of the *apparent error* (28) and the *cross-validation error* obtained in particular subsets of features. The best subset of features designated by the method is a subset $F_k[2] = \{x_7, x_2\}$. It is characterized by the lowest value of the *cross-validation error*.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	subset
$\mathbf{x}_1[10]$	0,04	-0,20	0,66	0,83	0,12	-0,06	2,70	-0,37	0,04	-0,43	G^+
$\mathbf{x}_2[10]$	-1,47	0,70	1,16	-0,54	-0,15	-0,47	1,80	0,24	0,12	0,15	G^+
$\mathbf{x}_3[10]$	0,34	1,10	0,27	0,22	2,45	1,19	1,30	0,45	-1,06	-1,25	G^+
$\mathbf{x}_4[10]$	-1,44	2,60	1,23	-1,86	-0,31	1,26	-0,30	0,34	0,19	0,14	G^+
$\mathbf{x}_4[10]$	-0,48	-0,80	-0,55	-0,77	-0,13	0,41	1,10	-0,13	-0,83	-0,97	G^-
$\mathbf{x}_6[10]$	0,54	0,20	-0,53	0,90	-0,25	0,54	0,30	-0,34	-0,60	0,70	G^-
$\mathbf{x}_7[10]$	-0,06	1,20	1,65	-1,77	0,34	1,41	-0,80	-0,65	0,98	-0,27	G^-

Table 1. Feature vectors $\mathbf{x}_j[10]$ constituting the sets G^+ and G^-

Subset of features	AE	CVE
$F_k[5] = \{x_7, x_2, x_1, x_8, x_3\}$	0	0,28571
$F_k[4] = \{x_7, x_2, x_8, x_3\}$	0	0,14286
$F_k[3] = \{x_7, x_2, x_3\}$	0	0,14286
$F_k[2] = \{x_7, x_2\}$	0	0
$F_k[1] = \{x_7\}$	0,2619	0,28571

Table 2. Subsets of features evaluated by the RLS method, *apparent error rate* (AE) and *cross-validation error rate* (CVE) obtained in particular subsets of features

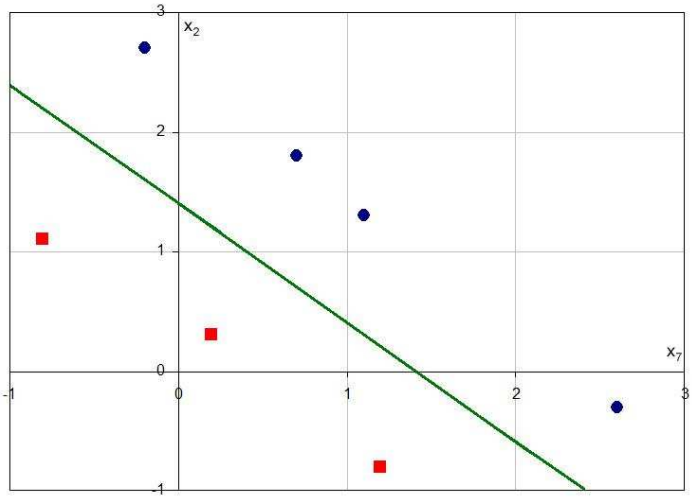


Fig. 4. Points in the feature space selected by the *RLS* method, hyperplane separated points falling within the sets G^+ (denoted by circles) and G^- (denoted by squares)

The *RLS* method in addition to the designation of the best subset of features has also determined the hyperplane separating objects from the sets G^+ and G^-.

$$H(\mathbf{w}[2],\theta) = \{\mathbf{x}[2]: 1{,}0204\, x_7 + 1{,}0884\, x_2 = 1{,}5238\} \tag{29}$$

8. Experiment on synthetic data

The data set used in the experiment contained 1000 objects, each described by 100 features. Data were drawn from a multivariate normal distribution. The values of each feature had a mean equal to 0 and standard deviation equal to 1. All the features were independent of each other (diagonal covariance matrix). The objects were divided into two disjoined subsets G^+ and G^- (1) in accordance with the values of the following linear combination:

$$3x_4+4x_{10}-7x_{17}+2x_{28}-6x_{36}+3x_{41}+3x_{58}-8x_{63}+x_{75}-x_{92}+5 \quad (30)$$

Objects corresponded to the value of expression (30) greater than 0 were assigned to subset G^+. Objects corresponded to the value of expression (30) less than 0 were assigned to subset G^-. The result was two linearly separable subsets G^+ and G^- (1) containing 630 and 370 objects.

The *RLS* method of feature selection was applied in analysis of the so-prepared synthetic data. The expected result was the preference by the method the subset of features used in the expression (30).

Figure 4 shows the *apparent error* (*AE*) and *cross-validation error* (*CVE*) values in the various tested features subspaces generated by the RLS method. Each subspace larger than 10 features ships with all 10 features used in the expression (30). Subspace of size 10 consists only of the features used in the expression (30).

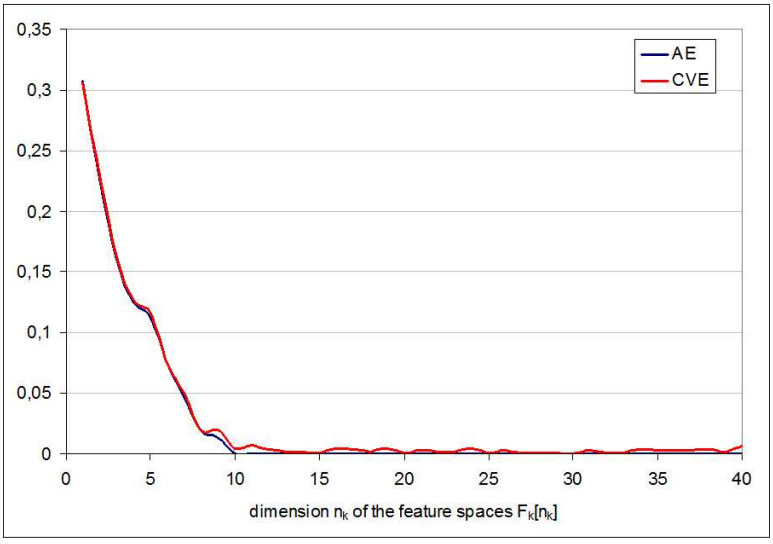


Fig. 5. The *apparent error* (AE) and the *cross-validation error* (CVE) in different feature subspaces $F_k[n_k]$ of the synthetic data set

9. Experiment on the *Leukemia* and the *Breast cancer* data sets

The *Leukemia* (Golub et al., 1999) data set contains expression levels of 7129 genes taken over 72 samples. Labels of objects indicate which of two variants of leukemia is present in the sample: acute myeloid (AML, 25 samples), or acute lymphoblastic leukemias (ALL, 47 samples).
The *Breast cancer* (van't Veer et al., 2002) data set describes the patients tested for the presence of breast cancer. The data contains 97 patient samples, 46 of which are from patients who had developed distance metastases within 5 years (labelled as "relapse"), the rest 51 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years (labelled as "non-relapse"). The number of genes is 24481.

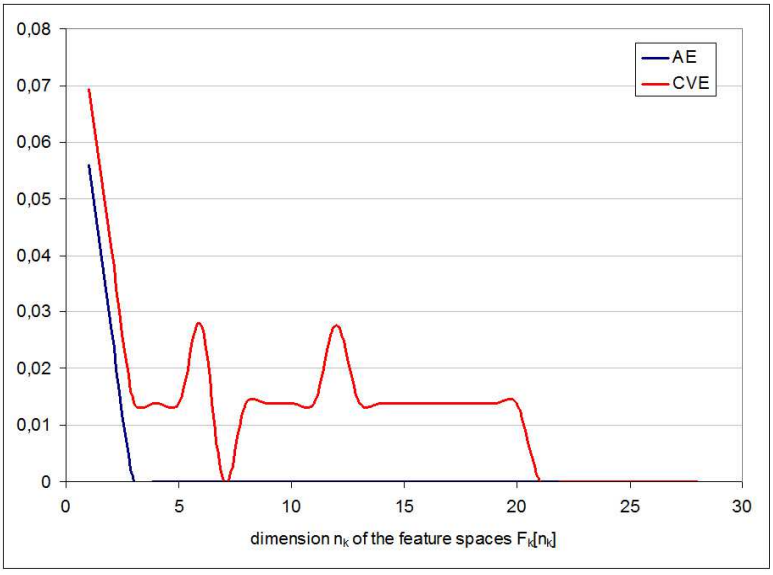


Fig. 6. The apparent error (AE) and the cross-validation error (CVE) in different feature subspaces $F_k[n_k]$ of the *Leukemia* data set

feature name	$F_k[7]$ weights w_i	$F_k[3]$ weights w_i
attribute4951	-0,99614	-1,71845
attribute1882	-0,73666	-11,6251
attribute3847	-0,55316	-
attribute6169	-0,47317	-
attribute4973	0,41573	-
attribute6539	-0,25898	-
attribute1779	-0,1519	-1,69028
threshold θ	-0,55316	2,53742

Table 3. Features x_i constituting the optimal subspace $F_k[7]$ characterised by the lowest *cross-validation error* (CVE) and features x_i constituting the lowest subspace $F_k[3]$ with *apparent error* (AE) equal to 0 of the *Leukemia* data set

Original data sets come with training and test samples that were drawn from different conditions. Here we combine them together for the purpose of cross validation. Data have also been standardized before experiment.

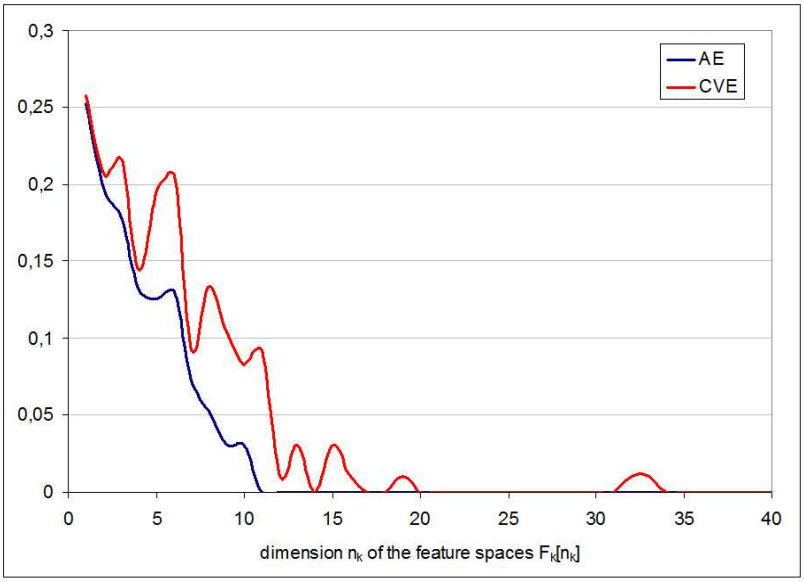


Fig. 7. The apparent error (AE) and the cross-validation error (CVE) in different feature subspaces $F_k[n_k]$ of the *Breast cancer* data set

feature name	$F_k[14]$ weights w_i	$F_k[11]$ weights w_i
Contig32002_RC	0,81467	1,89334
NM_000127	0,76305	2,67913
Contig412_RC	-0,71647	-
D86979	0,65172	-
Contig38438_RC	-0,63018	-1,58491
NM_016153	0,62345	1,81026
NM_015434	0,58631	1,43095
NM_013360	-0,58122	-1,38681
NM_002200	0,47752	1,94796
Contig44278_RC	-0,42246	-0,906564
NM_019886	0,4143	2,75393
AF055033	0,37546	1,24607
AL080059	0,31843	1,50648
NM_000909	-0,27537	-
threshold θ	0,1117	-0,247492

Table 4. Features x_i constituting the optimal subspace $F_k[14]$ characterised by the lowest *cross-validation error* (CVE) and features x_i constituting the lowest subspace $F_k[11]$ with *apparent error* (AE) equal to 0 of the *Breast cancer* data set

Figures 6 and 7 show the *apparent error* (*AE*) and *cross-validation error* (*CVE*) obtained in different feature subspaces generated by the *RLS* method. Full separability of data subsets is preserved in feature subsets much smaller than the initial very large sets of genes.

10. Conclusion

The problem of feature selection is usually resolved in practice through the evaluation of the usefulness (the validity) of individual *features* (*attributes*, *factors*) (Liu & Motoda, 2008). In this approach, resulting feature subsets are composed of such features which have the strongest individual influence on the analysed outcome. Such approach is related to the assumption about the independence of the factors. However, in a complex system, such as the living organism, these factors are often related. An advantage of the relaxed linear separability (*RLS*) method is that one may identify directly and efficiently a subset of features that influences the outcome and assesses the *combined* effect of these features as prognostic factors.

In accordance with the *RLS* method, the feature selection process involves two basic actions. The first of these actions is to generate the descending sequence (24) of feature subspaces $F_k[n_k]$. The second of the these actions is to evaluate the quality of the individual feature subspaces $F_k[n_k]$ in the sequence (24).

Generation of descending sequence (24) of feature subspaces $F_k[n_k]$ is done in the deterministic manner by multiple minimization of the criterion function the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) combined with gradual increasing of the parameter λ value. The criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) depends on the three nonnegative parameters: α_j - *prices* of the particular feature vectors $\mathbf{x}_j[n]$ (1), γ_i - feature costs, and λ - the cost level. The composition of the consecutive feature subspaces $F_k[n_k]$ (24) depends on the choice of these parameters. For example, the costly features x_i should have a sufficiently large values of the parameter γ_i. A high value of the parameter γ_i increases the chance for elimination of a given feature x_i.

Evaluation of the quality of the individual feature subspaces $F_k[n_k]$ is based in the *RLS* method on the cross-validation of the *CPL* optimal (*Definition* 3) linear classifier (25) designed in this subspace. The optimal linear classifier (25) is designed in the feature subspace $F_k[n_k]$ through the multiple minimization of the perceptron criterion function $\Phi_k(\mathbf{v}[n_k])$ defined (8) on the reduced feature vectors $\mathbf{y}_j'[n_k]$ ($\mathbf{y}_j'[n_k] \in F_k[n_k]$) or the modified criterion function $\Psi_{k\lambda}(\mathbf{v}[n_k])$ (12) with a small value (22) of the cost level λ.

This article also contains a description of the experiments with feature selection based on the R*LS* method. Experiments of the first group were carried out on synthetic data. The multivariate synthetic data were generated randomly and deterministically divided into two learning sets according of predetermined key. The given key was in the form of linear combination of the unknown number of selected features. The aim of the experiment was to find an unknown key, based on available sets of multidimensional data. The experiment confirmed this possibility.

Experiments of the second group were carried out on the genetic data sets *Leucemia* (Golub et al., 1999) and *Brest cancer* (van't Veer et al., 2002). These experiments have shown, inter alia, that the *RLS* method enables finding interesting and not too large subsets of features, even if the number of features at the beginning is a huge. For example, in the case of the *Brest cancer* set, the feature space was reduced from the dimensionality $n = 24481$ till $n_k = 11$ while the linear separability (27) of the learning sets $G^+[n_k]$ and $G^-[n_k]$ (26) were preserved.

The results of calculations described in this paper were obtained by using its own implementation of the basis exchange algorithms (http://irys.wi.pb.edu.pl/dmp).

Calculations in so high dimensional feature space $F[n]$ as n = 24481 were made possible by achieving a high efficiency of these algorithms.

11. Acknowledgment

This work was supported by the by the NCBiR project N R13 0014 04, and partially financed by the project S/WI/2/2011 from the Białystok University of Technology, and by the project 16/St/2011 from the Institute of Biocybernetics and Biomedical Engineering PAS.

12. References

Bobrowski, L. (1991). Design of piecewise linear classifiers from formal neurons by some basis exchange technique, In: *Pattern Recognition*, 24(9), pp. 863-870

Bobrowski, L. & Łukaszuk, T. (2004). Selection of the linearly separable feature subsets, In: *Artificial Intelligence and Soft Computing - ICAISC 2004*, eds. L. Rutkowski et al., Springer Verlag, pp. 544-549

Bobrowski, L. (2005). *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych (Data mining based on convex and piecewise linear (CPL) criterion functions)* (in Polish), Białystok University of Technology

Bobrowski, L. (2008). Feature subsets selection based on linear separbilty, In: *Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice*, ed. by H. Bacelar-Nicolau, L. Bobrowski, J. Doroszewski, C. Kulikowski, N. Victor, June 2008, Warsaw

Bobrowski L. & Łukaszuk T. (2009). Feature selection based on relaxed linear separability, In: *Biocybernetical and Biomedical Engineering*, vol.29, nr 2, pp. 43-58

Duda, O. R.; Hart, P. E. & Stork D. G. (2001). *Pattern Classification*, J. Wiley, New York

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*, Academic Press

Golub, T. R., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Sciences*, 286, pp. 531–537

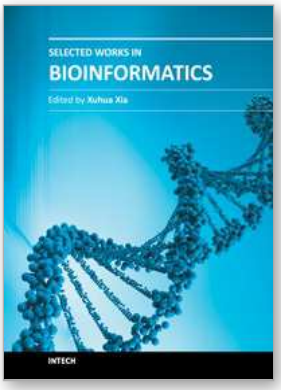
Guyon, I.; Weston, J.; Barnhill, S. & Vapnik, V. N. (2002). Gene Selection for Cancer Classification using Support Vector Machines, In: *Machine Learning*, 46, pp. 389-422

Lachenbruch, P.A. (1975). *Discriminant Analysis*, Hafner Press, New York.

Liu, H. & Motoda, H. (Eds.) (2008). *Computational Methods of Feature Selection*, Chapmann & Hall/CRC, New York

Vapnik, V. N. (1998). *Statistical Learning Theory*, J. Wiley, New York

van't Veer, L. J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415(6871), pp. 530–536



Selected Works in Bioinformatics

Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
Publisher InTech
Published online 19, October, 2011
Published in print edition October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Leon Bobrowski and Tomasz Łukaszuk (2011). Relaxed Linear Separability (RLS) Approach to Feature (Gene) Subset Selection, Selected Works in Bioinformatics, Dr. Xuhua Xia (Ed.), ISBN: 978-953-307-281-4, InTech, Available from: http://www.intechopen.com/books/selected-works-in-bioinformatics/relaxed-linear-separability-rls-approach-to-feature-gene-subset-selection

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821