Biocybernetics and Biomedical Engineering
2009, Volume 29, Number 2, pp. 43–59

Feature Selection Based on Relaxed Linear Separability

LEON BOBROWSKI[1, 2,*] , TOMASZ ŁUKASZUK[1]

[1] *Faculty of Computer Science, Technical University Białystok*
[2] *Institute of Biocybernetics and Biomedical Engineering, Warsaw, Poland*

Feature selection problem appears where large number of features constraint effective data analysis and processing. Identification of the most important feature subsets is a crucial challenge in many important applications. For example, a basic question in bioinformatics which is identification of genes functionalities, can be formulated and answered as a problem of this kind. Identification of the most important feature subsets through minimisation of convex and piecewise-linear (CPL) criterion function is described and analysed in the paper. This approach is combined with relaxation of the linear separability assumption.

K e y w o r d s: feature selection, relaxed linear separability, CPL criterion function

1. Introduction

Feature selection is one of the fundamental problems in pattern recognition [1]. Feature selection methods are used for removing irrelevant or redundant features. The importance of feature selection methods becomes apparent in the context of rapidly growing amount of collected data in contemporary databases [2].

It is assumed in the paper that objects collected in a given database are represented in a standard form as feature vectors of the same dimensionality and type, and are used for the purpose of decision support [3]. Components of each feature vector are numerical results of particular examinations of a given object. It is also assumed that objects stored in a database have been divided in accordance with expert opinion into disjoined categories (the supervised case). Basing on expert's decisions, the family of feature vectors has been divided into the disjoined learning sets. Each learning set contains the referential feature vectors describing objects of the same category. For example, particular learning sets may contain feature

* Correspondence to: Leon Bobrowski, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland, e-mail: leon.bobrowski@ibib.waw.pl
Received 05 January 2009; Accepted 09 February 2009

vectors describing patients with the same disease and can be used in the system of diagnosis support.

Decision support systems work in accordance with their decision rules. Such decision rules can be designed on the basis of learning sets by means of various methods. Evaluation and comparison of different decision rules is an important part of the designing process. In particular, statistical evaluation of different decision rules is demanded before their application in practice. A low number of objects in comparison to a large number of features is a serious obstacle in statistical evaluation of decision rules. Such problem appears usually in exploration of genomic data where the number of features can be greater thousand times than the number of objects [4]. Feature selection procedures are applied to data sets in order to decrease the number of features used during the decision stage.

Here we are considering such approach to the feature selection problem which refers to the concept of linear separability of the learning sets. The relaxation of linear separability combined with the feature selection is discussed in the paper. The term "relaxation" means here deterioration of the linear separability as a result of successive omitting of selected features. The considered approach to the feature selection is based on minimisation of the convex and piecewise-linear (*CPL*) criterion functions. The perceptron criterion function originated from the theory of neural network belongs to the considered *CPL* family [5].

2. Linear Separability of Two Learning Sets

Let us assume that m objects O_j contained in a given database has been represented as feature vectors $\mathbf{x}_j[n] = [x_{j1},...,x_{jn}]^T$ or as points in the n-dimensional feature space $F[n]$ ($j = 1,\ldots,m$). 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, the components x_{ji} can be the numerical results of diagnostic examinations of a given patient O_j.

Let us consider two learning sets G^+ and G^- of 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 $\mathbf{x}_j[n]$:

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

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

In practice, the positive set G^+ contains vectors $\mathbf{x}_j[n]$ of only one category. For example, the set G^+ may contain the feature vectors $\mathbf{x}_j[n]$ representing patients with cancer and the 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 $\mathbf{w}[n]$ ($\mathbf{w}[n] \in R^n$) and threshold θ ($\theta \in R$), that all the below inequalities are fulfilled:

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

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 situated on the positive side of the hyperplane $H(\mathbf{w}[n],\theta)$ (3) and all the elements of the set G^- are situated on the negative side of this hyperplane.

Remark 1: If m feature vectors $\mathbf{x}_j[n]$ are linearly independent, then the arbitrary sets G^+ and G^- (1) of these vectors are linearly separable [6].

Remark 2: If the sets G^+ and G^- (1) are linearly separable (2) in the feature space $F[n]$, then these sets are also linearly separable in any greater feature space $F'[n']$, where $F[n] \subset F'[n']$.

In accordance with the *Remark* 2, 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 $F[n]$.

It can be seen that linear separability (2) can be formulated equivalently to (2) as [7]:

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

where $\mathbf{y}_j[n+1]$ are the *augmented feature vec*tors, and $\mathbf{v}[n+1]$ is the *augmented weight vector*:

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

The inequalities (4) will be directly 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 introduce the convex and piecewise-linear penalty functions $\varphi_j^+(\mathbf{v}[n+1])$ and $\varphi_j^-(\mathbf{v}[n+1])$ [7]

$(\forall \mathbf{y}_j[n+1] \in G^+)$

$$\varphi_j^+(\mathbf{v}[n+1]) = \begin{cases} 1-\mathbf{v}[n+1]^T\mathbf{y}_j[n+1] & \textit{if} \quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] < 1 \\ 0 & \textit{if} \quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \geq 1 \end{cases} \tag{6}$$

and

$$(\forall \mathbf{y}_j[n+1] \in G^-)$$

$$\varphi_j^-(\mathbf{v}[n+1]) = \begin{cases} 1+\mathbf{v}[n+1]^T\mathbf{y}_j[n+1] & \textit{if} \quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] > -1 \\ 0 & \textit{if} \quad \mathbf{v}[n+1]^T\mathbf{y}_j[n+1] \le -1 \end{cases} \tag{7}$$

The 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. 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.

The perceptron criterion function $\Phi(\mathbf{v}[n+1])$ can be defined on the sets G^+ and G^- (1) as [6]:

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

where nonnegative parameters α_j determine *prices* of the particular feature vectors $\mathbf{x}_j[n]$.

We are interested in finding the minimum $\Phi(\mathbf{v}_k^*[n+1])$ of the criterion function $\Phi(\mathbf{v}[n+1])$:

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

It has been proved that the value Φ^* is equal to zero ($\Phi^* = 0$) if and only if the sets G^+ and G^- (1) are linearly separable (4) [6].

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

A modified *CPL* criterion function $\Phi_\lambda(\mathbf{v}[n+1])$ 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 [6]:

$$(\forall i \in \{1,\ldots,n\}) \quad \phi_i(\mathbf{v}[n+1]) = |w_i| = |\, \mathbf{e}_i[n+1]^T\mathbf{v}[n+1] \,| \tag{11}$$

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

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^-.

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

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 the hyperplanes $h_i^0[n+1]$ in the augmented parameter space R^{n+1} (5):

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

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 proved that the minimum of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) can be found in one of vertices $\mathbf{v}_k[n+1]$ of some region $R_l[n+1]$. Each vertex $\mathbf{v}_k[n+1]$ in the parameter space R^{n+1} is the intersection point of at least $(n+1)$ hyperplanes $h_j^{+}[n+1]$, $h_j^{-}[n+1]$ or $h_i^0[n+1]$. The below equations should be fulfilled in the vertex $\mathbf{v}_k[n+1]$:

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

The above equations can be given in the matrix form:

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

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 $\delta'[n+1]$ is the *margin vector* with components equal to 1, –1 or 0 according to (15).

Remark 3: The number n_1 of the independent vectors $\mathbf{y}_j[n+1]$ in the matrix $\mathbf{B}_k[n+1]$ (16) cannot be greater than the *rank r* of the data set $G^{+} \cup G^{-}$ (1). Therefore, the number n_0 of the unit vectors $\mathbf{e}_i[n+1]$ ($i \in I_k^{0}$) (15) in the matrix $\mathbf{B}_k[n+1]$ is not less than $n - r$ ($n_0 \geq n - r$).

The vertex $\mathbf{v}_k[n+1]$ (16) can be computed as follows:

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

The criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12), similarly to the function $\Phi(\mathbf{v}[n+1])$ (9) is convex and piecewise-linear (*CPL*). The minimum of this function is situated in one of the vertices $\mathbf{v}_k[n+1]$ (16):

$$(\forall \mathbf{v}[n+1]) \quad \Psi_\lambda(\mathbf{v}[n+1]) \geq \Psi_\lambda(\mathbf{v}_k^\wedge[n+1]) = \Psi_\lambda^\wedge. \tag{18}$$

The basis exchange algorithms allow to find efficiently the parameters (*vertex*) $\mathbf{v}_k^\wedge[n+1]$ constituting the minimum of the *CPL* function, even in the case of large data sets G^+ and G^- (1) [8].

The components w_{ki} of the vertex $\mathbf{v}_k[n+1]$ 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 vertex $\mathbf{v}_k^\wedge[n+1]$ (18) can be reduced without changing the separating hyperplane $H(\mathbf{w}_k^\wedge[n+1],\theta_k^\wedge)$ (4):

$$\begin{gathered}(\forall i \in I_k^0) \\ \mathbf{e}_i[n+1]^T\mathbf{v}_k^\wedge[n+1] = 0 \Rightarrow w_i = 0 \Rightarrow \text{the feature } x_i \text{ can be reduced.}\end{gathered} \tag{19}$$

As a result, the vertex $\mathbf{v}_k^\wedge[n+1]$ (18) can be fully characterized by the subset of $n-n_0$ features x_i ($i \notin I_k^0$). The vertex $\mathbf{v}_k^\wedge[n+1]$ (15) is characterized by the optimal 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 $\mathbf{B}_k^\wedge[n+1]$ (16) related to the optimal vertex $\mathbf{v}_k^\wedge[n+1]$ (18). As a result, the optimal feature subset $F_k^\wedge[n-n_0]$ can be identified in this approach by minimization (18) of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). One of procedures of feature subset selection can be based on this scheme.

Remark 4: 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 the increase of the number n_0 of unit vectors $\mathbf{e}_i[n+1]$ in the optimal base $\mathbf{B}_k^\wedge[n+1]$ linked to the vertex $\mathbf{v}_k^\wedge[n+1]$ (18) [7].

Therefore, the dimensionality of the optimized feature subset $F_k^\wedge[n-n_0]$ can be reduced arbitrarily with the increase 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 the vertex $\mathbf{v}_k^\wedge[n+1]$ (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 $\mathbf{v}_k^\wedge[n+1]$ (18) equal to zero ($\mathbf{v}_k^\wedge[n+1] = \mathbf{0}$). Such solution is not constructive, because it means that all the features x_i have been reduced (18) and the separating hyperplane $H(\mathbf{w}[n],\theta)$ (3) cannot be defined.

4. Selection of Optimal Feature Subset Based on Linear Separability

Let us consider the case of "*long vectors*", where the dimensionality n of the feature vectors $\mathbf{x}_j[n]$ is much greater than the number m ($n >> m$) of these vectors ($j = 1,..,m$). We can expect in such case that the vectors $\mathbf{x}_j[n]$ are linearly independent [7]. In accordance with the *Remark* 1, the arbitrary data sets G^+ and G^- of linearly independent

vectors $\mathbf{x}_j[n]$ are linearly separable (6). The minimal value Ψ^* (9) of the criterion function $\Psi(\mathbf{v}[n+1])$ (8) defined on linearly separable sets G^+ and G^- (4) is equal to zero ($\Psi^*=0$). The minimum of the function $\Psi(\mathbf{v}[n+1])$ (8) can be situated in the optimal vertex $\mathbf{v}_k^*[n+1]$ (9), where the below equations hold (15):

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

where $n' = n - n_0$ is the dimensionality of the reduce feature vectors $\mathbf{y}_j'[n']$ obtained from $\mathbf{y}_j[n+1]$ 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).

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 sets G^+ and G^- are linearly separable (4) in a given feature subspace $F_k[n']$ there can be more than one vertex $\mathbf{v}_k^*[n']$ constituting the minimum (9) of the function $\Phi(\mathbf{v})$ (8), which separates of these sets:

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

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 $\mathbf{v}_k^*[n']$ (20) 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 $\Psi_\lambda(\mathbf{v}[n+1])$ (13). Such vertex $\mathbf{v}_k^\wedge[n+1]$ (16) which constitutes minimum (18) of the function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) can be treated as the optimal one.

It can be proved that if the sets G^+ and G^- (1) are not linearly separable (21), then the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) 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) [5]:

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

The value of the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) in such points $\mathbf{v}[n+1]$ which separate linearly (21) the sets G^+ and G^- (1) can be expressed in the below manner (12):

$$\Psi_\lambda'(\mathbf{v}[n+1]) = \lambda \sum_{i \in I} \gamma_I \phi_i(\mathbf{v}[n+1]) = \lambda \sum_{i \in I} \gamma_i \left| \mathrm{v}_{ki} \right|. \tag{23}$$

Therefore, the minimization of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (13) can be replaced by the minimization of the function $\Psi_\lambda'(\mathbf{v}[n+1])$ (23) under the constraint that the point $\mathbf{v}[n+1]$ linearly separates 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 (22) of the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) 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 (21) 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 [9]. But *CPL* approach also allows to obtain different types of solution $\mathbf{v}_k^*[n+1]$ (22) by another specification of feature costs γ_i and the cost level λ parameters (12).

The feature selection procedure can be based on the minimization of the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) or of the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). In this approach, the minimal values $\Phi(\mathbf{v}_k^*[n+1])$ (9) or $\Psi_\lambda(\mathbf{v}_k^*[n+1])$ (22) of the criterion functions are used in the evaluation process of different feature subspaces $F_i[k]$ ($F_i[k] \subset F[n]$). The modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) gives additional possibility to introduce 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). The feature subset selection process considered in this paragraph assumed linear separability of the learning sets $G^+$and G^- (21). Below are presented considerations of feature selection with some relaxation of the linear separability assumption.

5. Feature Selection with the Linear Separability Relaxation

Different feature subsets $\{x_{i(1)},\ldots,x_{i(k)}\}$ or different feature subspaces $F_i[k]$ ($F_i[k] \subset F[n]$) can by evaluated by the minimal values of the *CPL* criterion function $\Phi(\mathbf{v}[n+1])$ (8) or the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). The minimal values $\Phi(\mathbf{v}_1^*[n+1])$ (9) or $\Psi_\lambda(\mathbf{v}_1\hat{}[n+1])$ (18) can be used as a measure of linear separability of the learning sets $G^+$and G^- (21) in the [7]. In order to compare the different feature subspaces $F_i[k]$ in this way, the criterion functions should be defined separately for each subspace.

Let the symbol $\Phi_i(\mathbf{v}[k+1])$ means the perceptron criterion function (8) defined on the augmented vectors $\mathbf{y}_j[k+1]$ (5), where k-dimensional feature vectors $\mathbf{x}_j[k]$ belong to the feature subspace $F_i[k]$ ($\mathbf{x}_j[k] \in F_i[k]$). It means that the penalty functions $\varphi_j^+(\mathbf{v}[k+1])$ (6) and $\varphi_j^-(\mathbf{v}[k+1])$ (7) are also defined on the augmented vectors $\mathbf{y}_j[k+1]$ (5). In this case, the minimal values $\Phi(\mathbf{v}_l^*[k+1])$ (9) or $\Psi_\lambda(\mathbf{v}_l\hat{}[k+1])$ (18) can be used as the measure of linear separability of the learning sets $G_i^+[k]$ and $G_i^-[k]$ (21) in the feature subspace $F_i[k]$ [7].

It can be proved that the criterion function $\Phi(\mathbf{v}[k+1])$ (8) has the *monotonocity property* [7]:

$$F_{i'}[k'] \subset F_i[k] \Rightarrow (\Phi_{i'}^* \geq \Phi_i^*) \tag{24}$$

where Φ_i^* is the minimal value (9) of the criterion function $\Phi_i(\mathbf{v}[k+1])$ (8) defined in the feature subspace $F_i[k]$, and $\Phi_{i'}^*$ is the minimal value of this criterion function defined in the feature subspace $F_{i'}[k']$.
Similar *monotonocity property* occurs for the modified criterion function $\Psi_\lambda(\mathbf{v}[k+1])$ (12):

$$F_{i'}[k'] \subset F_i[k] \Rightarrow (\Psi_{i'}^* \geq \Psi_i^*) \tag{25}$$

where Ψ_i^* is the minimal value (18) of the criterion function $\Psi_\lambda(\mathbf{v}[k+1])$ (12) defined in the feature subspace $F_i[k]$.

Let us call the minimal value Φ_i^* (24) of the criterion function $\Phi(\mathbf{v}[k+1])$ (8) the *measures of linear inseparability* of the learning sets $G_i^+[k]$ and $G_i^-[k]$ (21) in the feature subspace $F_i[k]$ [7]. In accordance with the above relations, neglecting of some features x_i from the subspace $F[n]$ cannot decrease the measures of linear inseparability Φ_k^* (24). Neglecting of sufficiently large number of features x_i results in increasing of the values Φ_k^* (24). In other words, the condition of linear separability can be relaxed in this way. The feature selection problem can be formulated on the basis of the measure Φ_i^* (24) in the below manner [7]:

Feature selection problem: Neglect maximal number of features x_i from the subspace $F[n]$ under the condition that an increase of the *measure of linear inseparability* Φ^* (24) is smaller than the given a priori margin γ_0 ($\gamma_0 \geq 0$):

$$\Phi_{i'}^* - \Phi_i^* \leq \gamma_0 \tag{26}$$

where Φ_i^* is the minimal value (18) of the criterion function $\Phi(\mathbf{v}[n+1])$ (8) defined in the feature space $F[n]$, and $\Phi_{i'}^*$ is the minimal value (9) of the criterion function $\Phi(\mathbf{v}[k+1])$ (8) defined in optimal feature subspace $F_i^*[k']$, composed of minimal number k' of features x_i.

The condition (26) reflects the concept of the relax linear separability because allows to worsening of the linear separability (21) in a some limit γ_0. The sets $G_i^+[n]$ and $G_i^-[n]$ which can be linearly separable (21) in the initial feature space $F[n]$ are not linearly separable in the reduced feature subspace $F_i^*[k']$. The solution of the feature selection problem (26) allows to identify the optimal feature subspace $F_i^*[k']$, and the optimal feature subset $S_i^*[k'] = \{x_{i(1)},\ldots,x_{i(k')}\}$. The solution of this problem can be reached through minimization of the criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) combined with an increasing of the *cost level* λ ($\lambda \geq 0$). In accordance with the *Remark* 4, it is possible to increase arbitrarily the number n_0 of the reduced features x_i by an increase of the parameter λ. Moreover, the less important features x_i can be reduced in this way.

6. Feature Selection in the Context of Linear Classification

The optimal vertex $\mathbf{v}_k^*[n+1] = [-\theta_k^*, \mathbf{w}_k^*[n]^T]^T$ (5) which constitutes the minimum (9) of the perceptron criterion function $\Phi(\mathbf{v}[n+1])$ (8) can be used also in definition of the linear classifier with the below decision rule concerning the allocation of the feature vector $\mathbf{x}[n]$ to one of the category ω_1 or ω_0:

$$\begin{aligned} &\textbf{\textit{if}} \quad \mathbf{w}_k^*[n]^T\mathbf{x}[n] \geq \theta_k^*, \quad \textbf{\textit{then}}\ \mathbf{x}[n] \textit{ is allocated to the category } \omega_1 \\ &\textbf{\textit{if}} \quad \mathbf{w}_k^*[n]^T\mathbf{x}[n] < \theta_k^*, \quad \textbf{\textit{then}}\ \mathbf{x}[n] \textit{ is allocated to the category } \omega_0 \end{aligned} \tag{27}$$

where the category (*class*) ω_1 is represented by elements $\mathbf{x}_j[n]$ of the learning set G^+ and the category ω_0 is represented by elements of the set G^-.

It has been proved that, if the sets G^+ and G^- (1) are linearly separable (4), then the above rule allocates correctly all elements $\mathbf{x}_j[n]$ of these learning sets [1]. It means that (21):

$$\begin{aligned} (\forall \mathbf{x}_j[n] \in G^+) \quad \mathbf{w}_k^*[n]^T\mathbf{x}_j[n] > \theta_k^*, \quad \textbf{\textit{and}} \\ (\forall\ \mathbf{x}_j[n] \in G^-) \quad \mathbf{w}_k^*[n]^T\mathbf{x}_j[n] < \theta_k^* \end{aligned} \tag{28}$$

If the sets G^+ and G^- (1) are not linearly separable (4), then not all but only a majority of the vectors $\mathbf{x}_j[n]$ fulfil the above inequalities.

The quality of the linear classifier (27) can be evaluated by using the error estimator (*error rate*) $e(\mathbf{w}_k^*[n],\theta_k^*)$ as the fraction of wrongly classified elements $\mathbf{x}_j[n]$ of the sets G^+ and G^- (1):

$$e(\mathbf{w}_k^*[n],\theta_k^*) = m_e\,(\mathbf{w}_k^*[n],\theta_k^*) \,/\, m \tag{29}$$

where m is the number of all elements $\mathbf{x}_j[n]$ of the sets G^+ and G^- (1), and $m_e(\mathbf{w}_k^*[n],\theta_k^*)$ is the number of elements $\mathbf{x}_j[n]$ wrongly allocated by the rule (27).

The parameters $\mathbf{w}_k^*[n]$ and θ_k^* of the linear classifier (27) are estimated from data sets G^+ and G^- (1) through minimization of the perceptron criterion function $\Psi(\mathbf{v}[n+1])$ (8) determined on elements $\mathbf{x}_j[n]$ of these sets. It is known that if the same data $\mathbf{x}_j[n]$ is used for classifier designing and classifier evaluation, then the evaluation results are too optimistic (*biased*). The error rate (29) evaluated on the elements $\mathbf{x}_j[n]$ of the learning sets is called the *apparent error*. For example, if the sets G^+ and G^- (1) are linearly separable (4), then the relation (28) holds and, as a result, the *apparent error* (29) evaluated on elements $\mathbf{x}_j[n]$ (1) is equal to zero ($e(\mathbf{w}_k^*[n],\theta_k^*) = 0$). But it is observed in practice that the error rate of the classifier (27) evaluated on new vectors $\mathbf{x}[n]$ is often greater than zero.

For the purpose of the classifier's bias reducing, the cross validation procedures are applied [3]. The term *p-fold cross validation* means that data sets G^+ and G^- (1) have been divided into p parts G_i, where $i = 1,\ldots, p$ (for example $p = 10$). The vec-

tors $\mathbf{x}_j[n]$ contained in $p-1$ parts G_i are used for definition of the criterion function $\Psi(\mathbf{v}[n+1])$ (8) and computing of the parameters $\mathbf{w}_k^*[n]$ and θ_k^*. The remaining vectors $\mathbf{x}_j[n]$ are used as the *test set* (one part $G_{i'}$) for computing (evaluation) the error rate $e(\mathbf{w}_k^*[n],\theta_k^*)$. Such evaluation is repeated p times, and each time different part $G_{i'}$ is used as the test set. The cross validation procedure allows to use different vectors $\mathbf{x}_j[n]$ (1) for the classifier (27) designing and evaluation (29) and as a result, to reduce the bias of the error rate estimation (29). The error rate (29) estimated during the *cross validation* procedure will be called the *cross-validation error*.

Another type of the classifier (27) evaluation is based on the so-called *confusion matrix* $\boldsymbol{T}(\mathbf{w}_k^*[n],\theta_k^*)$:

$$\boldsymbol{T}(\mathbf{w}_k^*[n],\theta_k^*) = \begin{bmatrix} m_{11} & m_{10} \\ m_{01} & m_{00} \end{bmatrix} \tag{30}$$

where m_{11} is the number of elements $\mathbf{x}_j[n]$ of the set G^+ (1) correctly allocated (27) in the category ω_1, and m_{10} is the number of elements in this set wrongly allocated in the category ω_0. Similarly, m_{00} is the number of elements $\mathbf{x}_j[n]$ of the set G^- (1) correctly allocated (27) in the category ω_0, and m_{01} is the number of elements in this set wrongly allocated in the category ω_1.

The confusion matrix $\boldsymbol{T}(\mathbf{w}_k^*[n],\theta_k^*)$ allows to estimate different types of errors related to the classifier (27). The cross validation procedures can be also used for estimating such types of errors.

Both the error rate $e_i(\mathbf{w}_k^*[k],\theta_k^*)$ (29) as well as *the confusion matrix* $\boldsymbol{T}_i(\mathbf{w}_k^*[n],\theta_k^*)$ (30) can be estimated in different feature subspaces $F_i[k]$ ($F_i[k] \subset F[n]$). Different feature subspaces $F_i[k]$ can be evaluated and compared on this basis. In general, the best feature subspaces $F_i[k]$ should be characterised by the lowest error rate $e_i(\mathbf{w}_k^*[k],\theta_k^*)$ (29) and a near diagonal confusion matrix $\boldsymbol{T}_i(\mathbf{w}_k^*[n],\theta_k^*)$ (30). Feature selection procedures can be organised in accordance with this demand.

7. Examples of Experimental Results

Arrhythmia data set and *Colon data set* was chosen for experimentation with described earlier feature selection procedures.

Arrhythmia data set was taken from the *UCI Machine Learning Repository* (http://www.ics.uci.edu/~mlearn/MLRepository.html) [10]. This data set describes patients with presence or absence of cardiac arrhythmia. The explored data set contained 420 patient's records described by 258 features. The features reflected the patient physical form (e.g. age, sex, weight, height) as well as the parameters of EEG signals. The records have been divided into two approximately equal groups (1): G^+ – *cardiac arrhythmia* and G^- – *normal EEG*.

Colon data set has been reported by Alon et al. [11]. This data set contains descriptions of 22 normal and 40 colon cancer samples. Each sample contains 2000

gene expression values. The gene expression values have been log transformed, and then normalized. This data set is available at http://www.iitk.ac.in/kangal/bioinfo.shtml and its description at http://microarray.princeton.edu/oncology.

During the first stage of the experiment the dimensionality of the feature vectors $\mathbf{x}_j[n]$ has been reduced while preserving the linear separability (21) of the data sets $G^+[n]$ and $G^-[n]$ (1). The minimization (18) of the modified criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12) combined with feature reduction (19) has been used for this purpose. In a result, the dimensionality of the *Arrhythmia data set* has been reduced from $n = 258$ to $n' = 163$. The dimensionality of the *Colon data set* has been reduced from $n = 2000$ to $n' = 39$. The most important feature subsets assuring linear separability have been also identified this way.

During the second stage of the experiment the dimensionality n' of the feature vectors $\mathbf{x}_j[n']$ has been reduced further while relaxing in some limits of the linear separability (21) of the data sets $G^+[n']$ and $G^-[n']$. Neglecting of the additional features x_i at this stage resulted in worsening of the linear separability. The minimization (9) of the perceptron criterion function $\Phi(\mathbf{v}[n'+1])$ (8) combined with the optimized feature selection (26) has been used at this stage.

The feature subspaces $F_k[n']$ obtained during the second stage have been additionally evaluated by the estimated error rate $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29). The *p-fold cross validation* procedure with $p = 5$ has been applied for this purpose. This procedure has been applied for selected dimensionalities n_i ($n_i < n'$). The data sets $G^+[n']$ and $G^-[n']$ has been divided randomly into p folds. For each dimensionality n_i, the data sets $G^+[n']$ and $G^-[n']$ has been divided into p folds K times (K=100). The repeated

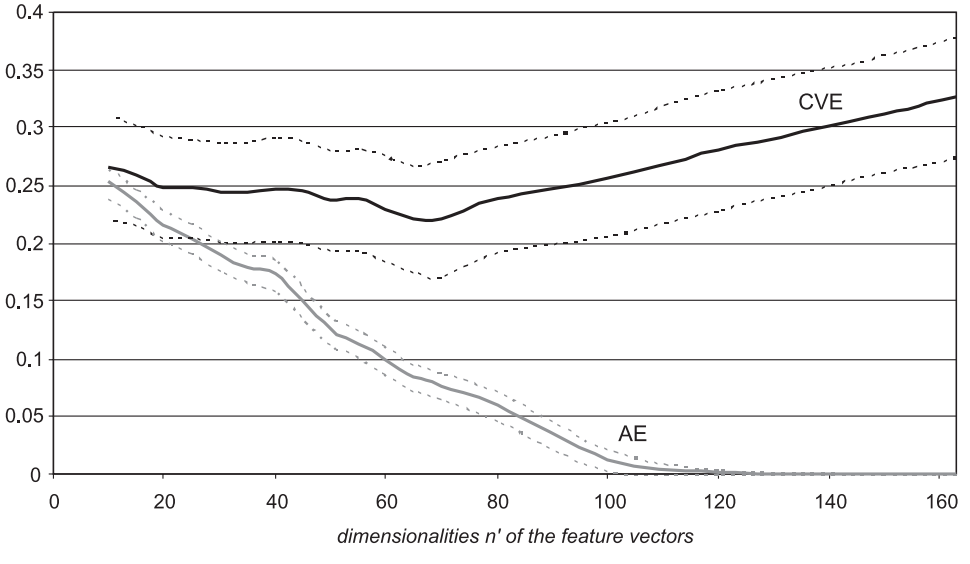

Fig. 1. Mean values and standard deviations of the cross validation error (CVE) $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29) and the apparent error (AE) estimated for various dimensionalities n'

divisions of the data sets $G^{+}[n']$ and $G^{-}[n']$ allowed for the evaluation both the mean value as well as the standard deviation (variance) of the estimated error rate $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29).

The results of these computations are summarised in the below tables and figures.

Table 1. Mean values and standard deviations of the cross validation error (CVE) $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29) and the apparent error (AE) for various dimensionalities n' (Arrhythmia data set)

n'	AE (mean)	AE (std dev)	CVE (mean)	CVE (std dev)
163	0	0	0.326567	0.0524481
150	1.79E-05	0.000229848	0.311954	0.0507423
120	0.00162306	0.0021981	0.280443	0.0518441
100	0.0124394	0.00911507	0.255526	0.0491135
80	0.0595116	0.0132822	0.237787	0.046031
70	0.0767527	0.0117279	0.221127	0.0507594
65	0.0834023	0.0116217	0.221069	0.0462739
60	0.0990721	0.0121999	0.230105	0.0452354
55	0.112449	0.0118556	0.237612	0.044522
50	0.125232	0.0123466	0.23675	0.0432401
45	0.149776	0.0136048	0.245164	0.0437401
40	0.172631	0.0131828	0.246641	0.0449471
35	0.178375	0.0129352	0.244646	0.0433411
30	0.189034	0.0130021	0.243791	0.0430754
25	0.203502	0.0124633	0.247908	0.0425431
20	0.215769	0.0133489	0.248619	0.0438193
15	0.23499	0.0126528	0.25896	0.0434154
10	0.252184	0.0133701	0.266102	0.0451637

Table 2. The *confusion matrix* $\boldsymbol{T}(\mathbf{w}_k^*[n'],\theta_k^*)$ (30) evaluated by the cross validation for $n' = 163$ and $n' = 65$ (Arrhythmia data set)

$n' = 163$	$G_1[n']$	$G_0[n']$
ω_1	168.71	68.29
ω_0	68.92	114.08

$n' = 65$	$G_1[n']$	$G_0[n']$
ω_1	190.58	46.42
ω_0	46.47	135.53

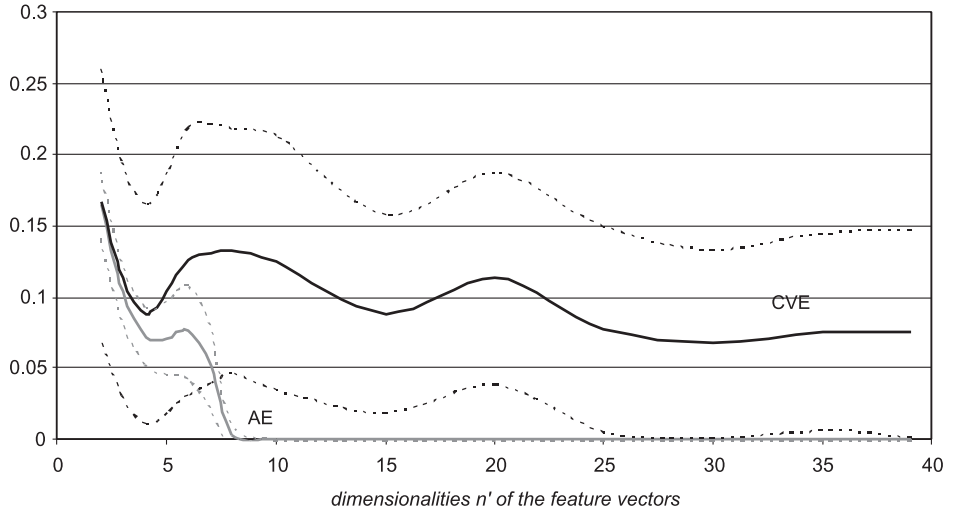

Fig. 2. Mean values and standard deviations of the cross validation error $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29) and the apparent error for various dimensionalities n' (Colon data set)

Table 3. Mean values and standard deviations of the cross validation error $e(\mathbf{w}_k^*[n'],\theta_k^*)$ (29) and the apparent error for various dimensionalities n' (Colon data set)

n'	AE (mean)	AE (std dev)	CVE (mean)	CVE (std dev)
39	0	0	0.0753366	0.072651
35	0	0	0.0760804	0.0689891
30	0	0	0.0677334	0.0663773
25	0	0	0.0777835	0.0719197
20	0	0	0.113534	0.0746782
15	0	0	0.0887336	0.0696158
10	0	0	0.125156	0.0891572
8	0.00329345	0.00776852	0.132832	0.0855875
7	0.0509652	0.0296641	0.130792	0.0908926
6	0.0763155	0.0325885	0.12595	0.0943788
5	0.0712929	0.0256584	0.103741	0.0839435
4	0.072621	0.0203354	0.0884397	0.0775479
3	0.104343	0.0203249	0.113357	0.0821719
2	0.164566	0.0232248	0.167032	0.0946489

Table 4. The confusion matrix $\boldsymbol{T}(\mathbf{w}_k^*[n'],\theta_k^*)$ (30) evaluated by the cross validation for $n' = 39$, $n' = 30$, $n' = 4$ (Colon data set)

$n' = 39$	$G_1[n']$	$G_0[n']$
ω_1	35.35	4.65
ω_0	0.040	21.96

$n' = 30$	$G_1[n']$	$G_0[n']$
ω_1	35.845	4.155
ω_0	0.05	21.95

$n' = 4$	$G_1[n']$	$G_0[n']$
ω_1	35.52	4.345
ω_0	1.004	20.96

8. Concluding Remarks

The process of feature selection has been divided in the paper into two stages. The first stage is relevant for a situation when the learning sets $G^+[n]$ and $G^-[n]$ (1) are linearly separable (2) in the initial feature space $F[n]$. Such situation often occurs in the case of *long vectors*, when the dimensionality n is much greater than the number m of feature vectors $\mathbf{x}_j[n]$ [2]. During the first stage number of features x_i are neglected in such a manner that the linear separability (2) of the learning sets $G^+[n']$ and $G^-[n']$ (1) is preserved in the reduced feature subspace $F_i[n']$ ($F_i[n'] \subset F[n]$).

The feature selection procedure during the first stage can be carried out efficiently through the minimization of the modified *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+1])$ (12). This criterion function depends on the three nonnegative parameters: α_j – *prices* of feature vectors, γ_i – feature costs, and λ – cost level. Properties of the resulting feature subspace $F_i[n']$ depend on the choice of values of these parameters. For example, the costly features x_i should have a sufficiently large values of the parameter γ_i. As a result of the parameter γ_i increasing a chance for the feature x_i neglecting also increases.

The presented second stage of feature selection is based on relaxation of the linear separability (2) of the learning sets $G_1^-[n']$ and $G_0^+[n']$ (1). Neglecting successive features x_i during this stage deteriorates linear separability (2). As a result, the minimal value $\Phi(\mathbf{v}_k^*[n'+1])$ (9) of the perceptron criterion function $\Phi(\mathbf{v}[n'+1])$ (8) increases successively ($\Phi(\mathbf{v}_k^*[n'+1]) > 0$). Departure from linear separability can be controlled by the margin γ_0 in the condition (26). At this stage, the feature selection procedure is aimed at reducing the maximal number of features x_i while the increase of the value $\Phi(\mathbf{v}_k^*[n'+1])$ (9) should be no greater than the margin γ_0. It was assumed here that the remaining features x_i constitute the feature subset $\{x_{i(1)},\ldots,x_{i(n')}\}$ with the greatest discriminative power. Such assumption has been supported by experimental results obtained on two data sets.

Acknowledgment

This work was supported by the KBN grant 3T11F01130, and partially financed by the grantS/WI/2/2008 from the Białystok University of Technology, and by the grant 16/St/2009 from the Institute of Biocybernetics and Biomedical Engineering PAS.

References

1. Duda O. R., Hart P. E., Stork D. G.: Pattern Classification, J. Wiley, New York 2001.
2. Liu H., Motoda H. (Eds.): Computational Methods of Feature Selection, Chapmann & Hall/ CRC, New York 2008.
3. Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press 1972.
4. Guyon I., Weston J., Barnhill S., Vapnik V. N.: Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, 2002, 46, 389–422.
5. Bobrowski L., Łukaszuk T.: Selection of the linearly separable feature subsets, in: Artificial Intelligence and Soft Computing – ICAISC 2004, L. Rutkowski et al. (Eds.), Springer Lecture Notes in Artificial Intelligence 3070, Springer Verlag 2004, 544–549.
6. Bobrowski L.: Feature subsets selection based on linear separabilty, in: VII-th ICB Seminar: Statistics and Clinical Practice, H. Bacelar- Nicolau, L. Bobrowski, J. Doroszewski, C. Kulikowski, N. Victor /Eds/, June 2008, Warsaw 17–23.
7. Bobrowski L.: Data mining based on convex and piecewise linear (CPL) criterion functions (in Polish), Technical University Białystok 2005.
8. Bobrowski L: Design of piecewise linear classifiers from formal neurons by some basis exchange technique. Pattern Recognition, 1991, 24, 9, 863–870.
9. Vapnik V.N.: Statistical Learning Theory, J. Wiley, New York 1998.
10. Asuncion A., Newman D.J.: UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, School of Information and Computer Science, 2007.
11. Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA, 1999, 96, Issue 12, June 8, 6745–6750.