FEATURE SELECTION FOR PROGNOSTIC MODELS BY LINEAR SEPARATION OF SURVIVAL GENETIC DATA SETS

Leon Bobrowski[1,2], Tomasz Łukaszuk[1]

[1] Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

[2] Institute of Biocybernetics and Biomedical Engineering PAS, Warsaw, Poland

Abstract: Designing regression models based on high dimensional (e.g. genetic) data sets through exploring linear separability problem is considered in the paper. The linear regression model designing has been reformulated here as the linear separability problem. Exploring the linear separability problem has been based on minimization of the convex and piecewise linear (CPL) criterion functions. The minimization of the CPL criterion functions was used not only for estimating the prognostic model parameters, but also for most effective selecting feature subsets (model selection) in accordance with the relaxed linear separability (RLS) method. This approach to designing prognostic models has been used in experiments both with synthetic multivariate data, and with genetic data sets containing censored values of dependent variable. The quality of the prognostic models resulting from the linear separability postulate has been evaluated by using the measure of the model discrepancy and the estimated classification error rate. In order to reduce the bias of the evaluation, the value of the model discrepancy and the classification error have been computed in different feature subspaces, in accordance with the cross-validation procedure. A series of new experiments described in this paper shows that the designing of regression models can be based on the linear separability principle. More specifically, the high-dimensional genetic sets with censored dependent variable can be used in designing procedure. The proposed measure of prognostic model discrepancy can be effectively used in the search for the optimal feature subspace and for selecting the linear regression model.

Keywords: data mining, interval regression, model selection, relaxed linear separability

1. Introduction

Multivariate regression analysis includes many techniques aimed at modelling the linear relationship between dependent variable and independent variables. In this case,

Advances in Computer Science Research, vol. 14, pp. 31-54, 2018. DOI: 10.24427/acsr-2018-vol14-0003

the value of a dependent variable is predicted to be the linear combination of some independent variables. The linear regression function is based on a finite number of unknown parameters that are estimated from the learning data set. The least squares method of the parameters estimation is commonly used in the regression analysis [11].

It has been recently demonstrated that the task of linear regression model designing can be formulated as a linear separability problem [3,5]. The linear separability problem has been investigated for many years in the context of the theory of neural networks and pattern recognition [1,9]. We use the convex and piecewise linear (*CPL*) criterion functions in our approach to the linear separability problem [2]. The basis exchange algorithms, which are similar to the linear programming, allow to efficiently find the minimal value of the *CPL* criterion function [6]. The parameters that create the minimum of an adequate *CPL* criterion function can be also used in the definition of the optimal regression models.

The *perceptron criterion function* belongs to the family of the *CPL* criterion functions [2]. The perceptron criterion function was modified by adding a regularization component for the purpose of the feature subset selection in accordance with the relaxed linear separability (*RLS*) method [4]. This regularization component used in the *RLS* method has a similar structure to those used in the *Lasso regression* [14]. The relaxed linear separability (*RLS*) method of feature subset selection is based on minimization of the modified perceptron criterion function. This method allows for a successive reduction of unnecessary features while preserving the linear separability of the learning sets.

Prognostic models in the area of survival analysis are designed on the basis of the so-called *censored* data sets. The Cox model plays a fundamental role in the survival analysis [12]. The modified perceptron criterion function can be also used for designing prognostic models (selection) on the basis of censored data sets. The possibility of the regression (prognostic) models selection from high-dimensional genetic data set with censored dependent variable is considered in the paper. Particular attention is paid to evaluating the quality of prognostic models obtained in this way.

The novelties introduced in the paper include: a) introduction of a new prognostic model quality measure, i.e. *discrepancy* coefficient, b) the series of new experiments proving the correctness of the concept adopted.

2. Methods

2.1 Different types of regression learning sets

Multivariate regression models are based on linear (affine) transformations of n-dimensional feature vectors $\mathbf{x}[n]$ taken from a given feature space $F[n]$ ($\mathbf{x}[n] \in F[n]$) on points t on the line ($t \in R^1$):

$$t(\mathbf{x}[n]) = \mathbf{w}[n]^T \mathbf{x}[n] + w_0 \tag{1}$$

where $\mathbf{w}[n] = [w_1, ..., w_n]^T \in R^n$ is the parameters' (*weight*) vector and w_0 is the threshold (*intercept coefficient*) ($w_0 \in R^1$).

Properties of the model (1) depend on the choice of the parameters $\mathbf{w}[n]$ and w_0. The weights w_i and the threshold w_0 are estimated from regression learning sets. In the case of classical regression analysis the learning sets are structured as follows:

$$\begin{aligned} C_0 = \{\mathbf{x}_j[n]; t_j\} = \{x_{j1}, ..., x_{jn}; t_j\}, \\ where \;\; j = 1, ..., m_0 \end{aligned} \tag{2}$$

Each object O_j in the set C_0 is characterized by values x_{ji} of n *independent variables* (*features*) X_i, and by the observed value t_j ($t_j \in R^1$) of the *dependent* (*target*) *variable* T. Components x_{ji} of the j-th feature vector $\mathbf{x}_j[n]$ could be treated as numerical results of n standardized examinations of the given object O_j ($x_{ji} \in \{0,1\}$ or $x_{ji} \in R^1$). Each feature vector $\mathbf{x}_j[n]$ can be also treated as a point in the n-dimensional feature space $F[n]$ [9].

In case of *classical regression*, the parameters $\mathbf{w}[n]$ and w_0 are estimated on the basis of set C_0 (2), in accordance with the method of *least squares* in such a way that the sum of the squared differences $(t_j - \hat{t_j})^2$ between the observed target variable y_j and the modelled variable $\hat{t_j} = \mathbf{w}[n]^T \mathbf{x}_j[n] + w_0$ (1) is minimal [11].

In case of *interval regression* learning sets C_I, additional knowledge about values t_j of the dependent variable T of particular objects O_j is represented by the intervals $[t_j^-, t_j^+]$ ($t_j^- < t_j^+$) instead of exact values t_j (1) [8,10]:

$$C_I = \{(\mathbf{x}_j[n], [t_j^-, t_j^+]), where \; j \in J_I\} \tag{3}$$

where J_I is the set of indices j of m_R objects O_j (feature vectors $\mathbf{x}_j[n]$), t_j^- is the lower bound ($t_j^- \in R^1$) and t_j^+ is the upper bound ($t_j^+ \in R^1$) of unknown value t_j ($t_j^- < t_j < t_j^+$) of the target variable T, which accompanies the j-th feature vector $\mathbf{x}_j[n]$.

Let us introduce the *right censored* set C_R and the *left censored* set C_L:

$$C_R = \{\mathbf{x}_j[n], [t_j^-, +\infty)\}, where\ j \in J_R \tag{4}$$

and

$$C_L = \{\mathbf{x}_j[n], (-\infty, t_j^+]\}, where\ j \in J_L \tag{5}$$

The set J_R contains the indices j of m_R objects O_j (feature vectors $\mathbf{x}_j[n]$) which are characterized by *right censored* values of the *dependent* variable T [12]. Similarly, the set J_L contains the indices j of such objects O_j, that are characterized by *left censored* values of the *dependent* variable T. It is assumed, that the sets C_R and C_L are disjoined ($C_R \cap C_L = \emptyset$). The censored sets C_R (4) or C_L (5) can be treated as a special type of the interval regression set C_I (3) in which either $t_j^+ = +\infty$ or $t_j^- = -\infty$.

The classical learning set C_0 (2) can be transformed into the interval learning set C_I (3) through introducing artificial boundary values $t_j^- = t_j - \varepsilon$ and $t_j^+ = t_j + \varepsilon$, where ε is a small positive parameter (*margin*) ($\varepsilon > 0$):

$$C_I' = \{\mathbf{x}_j[n], [t_j - \varepsilon, t_j + \varepsilon]\}, where\ j = 1, ..., m_0 \tag{6}$$

The following linear inequalities can be expected in case of prognostic model (1) designing on the basis of the interval learning set C_I (3):

$$(\forall j \in J_I)\ \ t_j^- < \mathbf{w}[n]^T \mathbf{x}_j[n] + w_0 < t_j^+ \tag{7}$$

or equivalently

$$\begin{aligned}(\forall j \in J_I)\ \ &\mathbf{w}[n]^T \mathbf{x}_j[n] + w_0 - t_j^- > 0 \\ and\ &\mathbf{w}[n]^T \mathbf{x}_j[n] + w_0 - t_j^+ < 0\end{aligned} \tag{8}$$

The feature vectors $\mathbf{x}_j[n]$ belonging to the censored sets C_R (4) or C_L (5) can be linked in a similar way to the below linear inequalities:

$$(\forall j \in J_R)\ \ \mathbf{w}[n]^T \mathbf{x}_j[n] + w_0 - t_j^- > 0 \tag{9}$$

and

$$(\forall j \in J_L)\ \ \mathbf{w}[n]^T \mathbf{x}_j[n] + w_0 - t_j^+ < 0 \tag{10}$$

We can note that censoring of some feature vector $\mathbf{x}_j[n]$ ($t_j^+ = +\infty$ or $t_j^- = -\infty$) results in removing one inequality from the set of inequalities (8).

2.2 Linear separability of the positive set $Z^{+}[n+2]$ and the negative set $Z^{-}[n+2]$

The interval learning set C_I (3) can be represented as the below sum of the disjoined subsets C_I', C_R (4), and C_L (5):

$$C_I = C_I' \cup C_R \cup C_L \tag{11}$$

where the subset C_I' contains such intervals $[t_j^-, t_j^+]$ that are not censored (the constraints t_j^- and t_j^+ are finite):

$$C_I' = \{(\mathbf{x}_j[n], [t_j^-, t_j^+]) : -\infty < t_j^- < t_j^+ < +\infty\} \tag{12}$$

The subsets C_R (4), C_L (5) and C_I' (12) are used in defining the *augmented feature vectors* $\mathbf{z}_j^+[n+2]$ and $\mathbf{z}_j^-[n+2]$ based on the linear inequalities (8), (9), and (10):

$$(\forall \mathbf{x}_j[n] \in C_I' \cup C_R) \quad \mathbf{z}_j^+[n+2] = [\mathbf{x}_j[n]^T, 1, -t_j^-]^T \tag{13}$$

and

$$(\forall \mathbf{x}_j[n] \in C_I' \cup C_L) \quad \mathbf{z}_j^-[n+2] = [\mathbf{x}_j[n]^T, 1, -t_j^+]^T \tag{14}$$

Let us introduce the *positive set* $\mathbf{Z}^+[n+2]$ and the *negative set* $\mathbf{Z}^-[n+2]$ which are composed of $(n+2)$-dimensional vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14):

$$\begin{aligned} &\mathbf{Z}^+[n+2] = \{\mathbf{z}_j^+[n+2]\} \text{ } \textit{and} \\ &\mathbf{Z}^-[n+2] = \{\mathbf{z}_j^-[n+2]\} \end{aligned} \tag{15}$$

Definition 1. *The positive set $\mathbf{Z}^+[n+2]$ and the negative set $\mathbf{Z}^-[n+2]$ (15) are linearly separable, if and only if there exists a parameter vector $\mathbf{v}'[n+2]$ ($\mathbf{v}'[n+2] \in R^{n+2}$), for which all the below inequalities are fulfilled [3,5]:*

$$\begin{aligned} (\exists \mathbf{v}'[n+2]) \; &(\forall \mathbf{z}_j^+[n+2] \in \mathbf{Z}^+[n+2]) \\ &\mathbf{v}'[n+2]^T \mathbf{z}_j^+[n+2] \geq 1 \\ \textit{and} \quad &(\forall \mathbf{z}_j^-[n+2] \in \mathbf{Z}^-[n+2]) \\ &\mathbf{v}'[n+2]^T \mathbf{z}_j^-[n+2] \leq 1 \end{aligned} \tag{16}$$

The parameter vector $\mathbf{v}'[n+2]$ defines the below hyperplane $H(\mathbf{v}'[n+2])$ in the feature space $F[n+2]$ ($\mathbf{z}[n+2] \in F[n+2]$):

$$H(\mathbf{v}'[n+2]) = \{\mathbf{z}[n+2] : \mathbf{v}'[n+2]^T \mathbf{z}[n+2] = 0\} \tag{17}$$

If all the inequalities (16) are fulfilled, then the hyperplane $H(\mathbf{v}'[n+2])$ (17) separates the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15). This means that each augmented feature vector $\mathbf{z}_j^+[n+2]$ (13) from the set $\mathbf{Z}^+[n+2]$ is situated on the *positive side* of the hyperplane $H(\mathbf{v}'[n+2])$ (17) ($\mathbf{v}'[n+2]^T\mathbf{z}_j^+[n+2] > 0$) and each augmented feature vector $\mathbf{z}_j^-[n+2]$ (14) from the set $\mathbf{Z}^-[n+2]$ is situated on the *negative side* of this hyperplane ($\mathbf{v}'[n+2]^T\mathbf{z}_j^+[n+2] < 0$).

The desirable inequalities (8), (9), (10) can be represented as the linear separability problem (16), if the parameter vector $\mathbf{v}[n+2]$ has the below structure [3]:

$$\mathbf{v}[n+2] = [v_1, ..., v_{n+2}]^T = [\mathbf{w}[n]^T, w_0, \beta]^T \tag{18}$$

where β is the *interval parameter* ($\beta \in R^1$).

The parameter vector $\mathbf{v}[n+2]$ (18) allows to define the below prognostic model:

$$y(\mathbf{x}[n]) = (\mathbf{w}[n]/\beta)^T\mathbf{x}[n] + w_0/\beta \tag{19}$$

The following *Lemma* can be proved [3]:

Lemma 1. *All the desirable inequalities (8), (9), (10) are fulfilled by the prognostic model $y(\mathbf{x}[n])$ (19) defined by the parameters vector $\mathbf{v}'[n+2] = [\mathbf{w}'[n]^T, w_0', \beta']$ (18) if and only if the hyperplane $H(\mathbf{v}'[n+2])$ (17) fully separates (16) the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15).*

If the number n of features X_i is larger than the number m of the vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14) in the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15), then such sets are usually linearly separable [9]. The most interesting are cases when linear separability (16) occurs in the opposite circumstances, when the number m of feature vectors is large in comparison to the number n of features.

The concept of *linear separability* has been used for many years in the theory of neural networks and in pattern recognition methods. The linear separability has been used in the proof of the convergence of the error-correction algorithm - classical learning algorithm of neural networks [9]. The optimal linear classifiers in pattern recognition can be designed through exploration of the linear separability of the learning sets [2].

2.3 Convex and piecewise linear (CPL) criterion function defined on the positive set $\mathbf{Z}^+[n+2]$ and the negative set $\mathbf{Z}^-[n+2]$

The positive set $\mathbf{Z}^+[n+2]$ and the negative set $\mathbf{Z}^-[n+2]$ (15) are composed of the $(n+2)$ - dimensional vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14), adequately. The sets

$\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) are linearly separable if and only if the inequalities (16) are fulfilled.

The below convex and piecewise-linear (*CPL*) penalty functions $\varphi_j^+(\mathbf{v}[n+2])$ and $\varphi_j^-(\mathbf{v}[n+2])$ are introduced for solving inequalities (16):

$$(\forall \mathbf{z}_j^+[n+2])$$
$$\varphi_j^+(\mathbf{v}[n+2]) = \begin{array}{l} 1-\mathbf{v}[n+2]^T\mathbf{z}_j^+[n+2] \\ \quad \mathit{if}\ \mathbf{v}[n+2]^T\mathbf{z}_j^+[n+2] < 1 \\ 0 \\ \quad \mathit{if}\ \mathbf{v}[n+2]^T\mathbf{z}_j^+[n+2] \geq 1 \end{array} \tag{20}$$

$$(\forall \mathbf{z}_j^-[n+2])$$
$$\varphi_j^-(\mathbf{v}[n+2]) = \begin{array}{l} 1+\mathbf{v}[n+2]^T\mathbf{z}_j^-[n+2] \\ \quad \mathit{if}\ \mathbf{v}[n+2]^T\mathbf{z}_j^-[n+2] > -1 \\ 0 \\ \quad \mathit{if}\ \mathbf{v}[n+2]^T\mathbf{z}_j^-[n+2] \leq -1 \end{array} \tag{21}$$

The *perceptron criterion function* $\Phi(\mathbf{v}[n+2])$ is defined as the weighted sum of the penalty functions $\varphi_j^+(\mathbf{v}[n+2])$ (20) and $\varphi_j^-(\mathbf{v}[n+2])$ (21) [2]:

$$\Phi(\mathbf{v}[n+2]) = \sum_j \alpha_j^+\varphi_j^+(\mathbf{v}[n+2]) + \sum_j \alpha_j^-\varphi_j^-(\mathbf{v}[n+2]) \tag{22}$$

where non-negative parameters α_j^+ ($\alpha_j^+ > 0$) determine the *importance* of particular vectors $\mathbf{z}_j^+[n+2]$ (13) and parameters α_j^- ($\alpha_j^+ > 0$) determine the *importance* of particular vectors $\mathbf{z}_j^-[n+2]$ (14). Standard values of the parameters α_j^+ and α_j^- can be provided as follows [2]:

$$\begin{array}{ll} (\forall \mathbf{z}_j^+[n+2]) & \alpha_j^+ = 1/(2m^+)\ \mathit{and} \\ (\forall \mathbf{z}_j^-[n+2]) & \alpha_j^- = 1/(2m^-) \end{array} \tag{23}$$

where m^+ is the number of the vectors $\mathbf{z}_j^+[n+2]$ (13) and m^- is the number of the vectors $\mathbf{z}_j^-[n+2]$ (14).

The optimal vector $\mathbf{v}^*[n+2]$ constitutes the global minimum of the *CPL* criterion function $\Phi(\mathbf{v}[n+2])$ (22):

$$(\forall \mathbf{v}[n+2])\ \Phi(\mathbf{v}[n+2]) \geq \Phi(\mathbf{v}^*[n+2]) = \Phi^* \geq 0 \tag{24}$$

where $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$, and $\mathbf{w}^*[n] = [w_1^*, ..., w_n^*]^T$.

The basis exchange algorithms which are similar to linear programming, allow to find the minimal value Φ^* (24) of the function $\Phi(\mathbf{v}[n+2])$ (22) and the optimal parameters $\mathbf{v}^*[n+2]$ efficiently, even in case of large, multidimensional data sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) [3,5].

The following remarks describe useful properties of the minimal value $\Phi^* = \Phi(\mathbf{v}^*[n+2])$ (24) of the perceptron criterion function $\Phi(\mathbf{v}[n+2])$ (22) [2]:

Remark 1. detection of the linear separability The minimal value $\Phi^* = \Phi(\mathbf{v}^*[n+2])$ (24) of the criterion function $\Phi(\mathbf{v}[n+2])$ (22) with the standard values (23) of the parameters α_j^+ and α_j^- is contained in the interval $< 0, 1 >$

$$0 \le \Phi^* \le 1 \tag{25}$$

where $\Phi^* = 0$ if and only if the positive set $\mathbf{Z}^+[n+2]$ and the negative set $\mathbf{Z}^-[n+2]$ (15) are linearly separable (16).

Remark 2. the positive monotonicity property The omission of an arbitrary pair $(\mathbf{x}_j[n], [y_j^-, y_j^+])$ from the learning set C_l (3) can not increase the value of Φ^* (24) (the value of Φ^* usually *decreases*).

Remark 3. the negative monotonicity property The omission of any of the component x_{ji} (feature X_i) in all the m feature vectors $\mathbf{x}_j[n] = [x_{j1}, ..., x_{jn}]^T$ (3) can not reduce the value of Φ^* (24) (the value of Φ^* usually *increases*).

Remark 4. the invariancy property The minimal value Φ^* (24) of the criterion function $\Phi(\mathbf{v}[n+2])$ (22) does not depend on linear (affine), nonsingular transformations of feature vectors $\mathbf{x}_j[n]$ (3):

$$\begin{gathered} if\ (\forall j \in \{1, ..., m\})\ \mathbf{x}_j'[n] = \mathbf{A}\mathbf{x}_j[n] + \mathbf{b}[n], \\ where\ \mathbf{A}^{-1}\ exists,\ then\ \Phi_{x'}^* = \Phi_x^* \end{gathered} \tag{26}$$

where $\mathbf{b}[n]$ is a constant vector ($\mathbf{b}[n] \in R^n$), and $\Phi_{x'}^*$ is the minimal value (24) of the perceptron criterion function $\Phi_{x'}(\mathbf{v}[n+2])$ (22) defined on elements of the learning set $C_l' = \{(\mathbf{x}_j'[n], [y_j^-, y_j^+]),\ where\ j \in J_l\}$ (3).

The optimal parameters $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$ (24) that constitute the minimal value $\Phi^* = \Phi(\mathbf{v}^*[n+2])$ (24) of the criterion function $\Phi(\mathbf{v}[n+2])$ (22) are used in the definition of the below *CPL* prognostic model [5]:

$$t^*(\mathbf{x}[n]) = (\mathbf{w}^*[n]^T \mathbf{x}[n] + w_0^*)/\beta^* \tag{27}$$

Lemma 2. *If the minimal value* $\Phi^* = \Phi(\mathbf{v}^*[n+2])$ *(24) of the criterion function* $\Phi(\mathbf{v}[n+2])$ *(22) is equal to zero (*$\Phi^* = 0$*) in the extreme point* $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$ *with* $\beta^* > 0$*, then the optimal prognostic model (27) fulfills all the inequalities (7):*

$$(\forall j \in J_I)\; t_j^- < (\mathbf{w}^*[n]/\beta^*)^T \mathbf{x}_j[n] + w_0^*/\beta^* < t_j^+ \tag{28}$$

The above conditions can be proved directly from the linear separability inequalities (16). If the minimal value Φ^* (24) is greater than zero ($\Phi^* > 0$) in the extreme point $\mathbf{v}^*[n+2]$, then the optimal prognostic model (27) satisfies the majority but not all the inequalities (28) [5].

2.4 The modified criterion function $\Psi(\mathbf{v}[n+2])$

The perceptron criterion function $\Phi(\mathbf{v}[n+2])$ (22) has been modified in order to allow selecting features task though including *feature penalty functions* $\phi_i(\mathbf{v}[n+2])$ and the *costs* γ_i ($\gamma_i \geq 0$) related to particular features X_i [4]. The feature penalty functions $\phi_i(\mathbf{v}[n+2])$ are defined in the below manner:

$$\begin{aligned}&(\forall i \in \{1, ..., n\})\\ &\quad \phi_i(\mathbf{v}[n+2]) = |\mathbf{e}_i[n+2]^T \mathbf{v}[n+2]| = |w_i|\end{aligned} \tag{29}$$

The modified criterion function $\Psi(\mathbf{v}[n+2])$ is the sum of the basic function in the form of perceptron criterion function $\Phi(\mathbf{v}[n+2])$ (20) and regularization components with the penalty functions $\phi_i(\mathbf{v}[n+2])$ [4]:

$$\begin{aligned}&\Psi_\lambda(\mathbf{v}[n+2]) =\\ &\quad = \Phi(\mathbf{v}[n+2]) + \lambda \textstyle\sum_{i \in \{1,...,n\}} \gamma_i \phi_i(\mathbf{v}[n+2]) =\\ &\quad = \Phi(\mathbf{v}[n+2]) + \lambda \textstyle\sum_{i \in \{1,...,n\}} \gamma_i |w_i|\end{aligned} \tag{30}$$

where λ ($\lambda \geq 0$) is the *cost level*, and γ_i are feature costs ($\gamma_i \geq 0$).

The standard assumption about the *feature costs* γ_i is that these costs are equal to one:

$$(\forall i \in \{1, ..., n\})\; \gamma_i = 1 \tag{31}$$

The modified criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) is used in the *relaxed linear separability* (*RLS*) method of feature subset selection [2]. The regularization component $\lambda \sum \gamma_i |w_i|$ used in the modified criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) is similar to these used in the *Lasso* method [14]. The *Lasso* method was developed as a part of regression analysis for the model selection [14]. The main difference between the *Lasso* and the *RLS* methods is the type of the basic criterion function. This difference affects the computational techniques used to minimize the modified criterion functions. The perceptron criterion function $\Phi(\mathbf{v}[n+2])$ (22) plays fundamental role in case of the *RLS* method. Both the basic criterion $\Phi(\mathbf{v}[n+2])$ (22), as well as the penalty functions $\phi_i(\mathbf{v}[n+2])$ (29) are convex and piecewise-linear (*CPL*). As a result, the modified criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) is also convex and piecewise-linear (*CPL*). The basis exchange algorithms allow to find efficiently the optimal vector of parameters (*vertex*) $\mathbf{v}_\lambda^*[n+2]$ constituting the minimum of the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) with the cost level λ:

$$\begin{aligned}&(\exists \mathbf{v}_\lambda^*[n+2])\ (\forall \mathbf{v}[n+2])\\ &\quad \Psi_\lambda(\mathbf{v}[n+2]) \geq \Psi_\lambda(\mathbf{v}_\lambda^*[n+2]) = \Psi_\lambda^*\end{aligned} \tag{32}$$

where $\mathbf{v}_\lambda^*[n+2] = [\mathbf{w}_\lambda^*[n]^T, w_{\lambda 0}^*, \beta_\lambda^*]^T = [w_{\lambda 1}^*, ..., w_{\lambda n}^*, w_{\lambda 0}^*, \beta_\lambda^*]^T$ (24). In the *RLS* method the optimal parameters $w_{\lambda i}^*$ are used in the feature reduction rule below:

$$(w_{\lambda i}^* = 0)\ => \ (the\ feature\ X_i\ is\ reduced) \tag{33}$$

We can remark that the features X_i which have the weights $w_{\lambda i}^*$ equal to zero ($w_{\lambda i}^* = 0$) in the optimal vertex $\mathbf{v}_\lambda^*[n+2]$ (32) can be reduced (33) without changing the optimal prognostic model $y^*(\mathbf{x}[n])$ (27) which is defined by the parameters $\mathbf{v}_\lambda^*[n+2]$ (32).

It can be proved that the vector $\mathbf{v}_\lambda^*[n+2]$ which constitutes the minimum (32) of the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) can be located on the optimal vertex $\mathbf{v}_k^*[n+2]$ ($\mathbf{v}_\lambda^*[n+2] = \mathbf{v}_k^*[n+2]$) linked to some basis $\mathbf{B}_k^*[n+2]$ in the $(n+2)$ - dimensional feature space $F[n+2]$ [2]:

$$\mathbf{B}_k^*[n+2]\mathbf{v}_k^*[n+2] = \delta_k^*[n+2] \tag{34}$$

where $\mathbf{B}_k^*[n+2]$ is the non-singular matrix (*basis*) with rows consisting of $(n+2)$ linearly independent vectors $\mathbf{z}_j^+[n+2]$ $(j \in J_k^+)$, $\mathbf{z}_j^-[n+2]$ $(j \in J_k^-)$ or by unit vectors $\mathbf{e}_i[n+2]$ $(i \in I_k^0)$, and $\gamma_k^*[n+1]$ is the *margin vector* with components equal to 1, -1 or 0, adequately to the below equations fulfilled in the vertex $\mathbf{v}_k^*[n+2]$:

$$\begin{aligned}&(\forall j \in J_k^+)\ \mathbf{z}_j^+[n+2]^T\mathbf{v}_k^*[n+2] = 1,\ and\\ &(\forall j \in J_k^-)\ \mathbf{z}_j^-[n+2]^T\mathbf{v}_k^*[n+2] = -1,\ and\\ &(\forall i \in I_k^0)\ \mathbf{e}_i[n+2]^T\mathbf{v}_k^*[n+2] = 0\end{aligned} \tag{35}$$

where J_k^+, J_k^- and I_k^0, are the sets of indices of the basis vectors $\mathbf{z}_j^+[n+2]$, $\mathbf{z}_j^-[n+2]$, and $\mathbf{e}_i[n+2]$, adequately.

Remark 5. The features X_i which are linked to the unit vectors $\mathbf{e}_i[n+2]$ $(i \in I_k^0)$ in the optimal basis $\mathbf{B}_k^*[n+2]$ (34) can be reduced (33) in the related vertex $\mathbf{v}_k^*[n+2] = [\mathbf{w}_k^*[n]^T, w_{k0}^*, \beta_k^*]^T = [w_{k1}^*, ..., w_{kn}^*, w_{k0}^*, \beta_k^*]^T$.

The Remark 5 can be justified by the below implication (33) [4]:

$$\begin{aligned}&(\forall i \in I_k^0)\\&(\mathbf{e}_i[n+2]^T\mathbf{v}_k^*[n+2] = 0) => (w_{ki}^* = 0) =>\\&=> (the\ feature\ X_i\ is\ reduced)\end{aligned} \tag{36}$$

The minimal value $\Psi_\lambda(\mathbf{v}_\lambda^*[n+2])$ (32) of the *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) in the vertex $\mathbf{v}_k^*[n+2]$ represents an equilibrium between the "force" of linear separability (16) and the "force" of features costs determined by the parameters λ and γ_i. We can remark that an increase of the value of the parameter λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) causes an increase in the number of the unit vectors $\mathbf{e}_i[n+2]$ in the basis $\mathbf{B}_k^*[n+2]$ (34) linked to the optimal vertex $\mathbf{v}_k^*[n+2]$ (32). As a consequence, an increase of the *cost level* λ value in the minimized function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) results in an increased number of the reduced features X_i (36). Furthermore, the dimensionality of the feature $F[n]$ can be reduced arbitrarily in accordance with the rule (36) by a sufficient increase of the parameter λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30). Such method of feature selection is called *relaxed linear separability* (*RLS*) [4]. A successive increase of the *cost level* λ in the minimized function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) allows to reduce the less important (redundant) features X_i and to generate descending sequence of feature subspaces $F_k[n_k]$ ($F_k[n_k] \supset F_{k+1}[n_{k+1}]$, where $n_k > n_{k+1}$) [4]:

$$\begin{gathered}F[n] \rightarrow F_1[n_1] \rightarrow ... \rightarrow F_k[n_k],\\ where\ 0 \leq \lambda_0 < \lambda_1 < ... < \lambda_k\end{gathered} \tag{37}$$

Each feature subspace $F_k[n_k]$ in the above sequence has been linked to a certain value λ_k of the cost level λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30). The sequence (37) of the feature subspaces $F_k[n_k]$ is generated in a deterministic manner based on the positive set $\mathbf{Z}^+[n+2]$ and the negative set $\mathbf{Z}^-[n+2]$ (15) in accordance with the *relaxed linear separability* (*RLS*) method [4]. Each step $F_k[n_k] \rightarrow F_{k+1}[n_{k+1}]$ has been realized by a minimal increase $\lambda_k \rightarrow \lambda_{k+1} = \lambda_k + \Delta_k$ (where $\Delta_k > 0$) of the cost level λ in the criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30).

A high value λ_k of the cost level λ in criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) can cause replacement of all vectors $\mathbf{z}_j^+[n+2]$ (13) or $\mathbf{z}_j^-[n+2]$ (14) in the basis $\mathbf{B}_k^*[n+2]$ (34) linked to the optimal vertex $\mathbf{v}_k^*[n+2]$ (32) by unit vectors $\mathbf{e}_i[n+2]$ and the solution $\mathbf{v}_k^*[n] = 0$ could appear. Such solution is not a constructive one, because it means that all features X_i have been eliminated (33). A compromise solution is needed, which would allow to preserve the most important feature subset. Such postulate can be realized through an adequate stop criterion in the process of the feature space $F[n]$ reduction (37). The stop criterion could be based on evaluating the quality of particular feature subspaces $F_k[n_k]$ in the sequence (37).

In accordance with the *relaxed linear separability* (*RLS*) approach to feature subset selection, the quality of particular subspaces $F_k[n_k]$ (37) is evaluated on the basis of the optimal linear classifier designed in this subspace [4]. A better optimal linear classifier means a better feature subspace $F_k[n_k]$. In the context of this paper, a better feature subspace $F_k[n_k]$ (37) should guarantee a better prognostic model (27).

2.5 Evaluation of the *CPL* regression models based on non censored data

Let us consider the learning set C_I (3) composed of the disjoined subset of not censored data C_0 (2) and the censored subsets C_R (4), and C_L (5):

$$C_I = C_0 \cup C_R \cup C_L \tag{38}$$

The *right censored* subset C_R (4) as well as the *left censored* subset C_L (5) could be empty. The nonempty subset C_0 (2) of feature vector $\mathbf{x}_j[n]$ with not censored values t_j can be transformed into the subset C_I' (6) with the intervals $[t_j - \varepsilon, t_j + \varepsilon]$ in order to define the augmented vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14). Each element $\mathbf{x}_j[n]$ of the not censored data set C_0 (2) generates two augmented vectors $\mathbf{z}_j^+[n+2]$ and $\mathbf{z}_j^-[n+2]$:

$$\begin{aligned} &(\forall \mathbf{x}_j[n] \in C_0) \\ &\mathbf{z}_j^+[n+2] = [\mathbf{x}_j[n]^T, 1, -t_j + \varepsilon]^T \\ &\mathbf{z}_j^-[n+2] = [\mathbf{x}_j[n]^T, 1, -t_j - \varepsilon]^T \end{aligned} \tag{39}$$

Each element $\mathbf{x}_j[n]$ of the censored subsets C_R (4) or C_L (5) generates only one augmented vector $\mathbf{z}_j^+[n+2]$ (13) or $\mathbf{z}_j^-[n+2]$ (14):

$$(\forall \mathbf{x}_j[n] \in C_R)\ \mathbf{z}_j^+[n+2] = [\mathbf{x}_j[n]^T, 1, -tj - \varepsilon]^T \tag{40}$$

and

$$(\forall \mathbf{x}_j[n] \in C_L)\ \mathbf{z}_j^-[n+2] = [\mathbf{x}_j[n]^T, 1, -tj + \varepsilon]^T \tag{41}$$

The convex and piecewise-linear (*CPL*) criterion functions $\Phi(\mathbf{v}[n+2])$ (22) and $\Psi_\lambda(\mathbf{v}[n+2])$ (30) were defined on the elements $\mathbf{z}_j^+[n+2]$ (39) or (40) of the set $\mathbf{Z}^+[n+2]$ (15) and on the elements $\mathbf{z}_j^-[n+2]$ (39) or (41) of the set $\mathbf{Z}^-[n+2]$ (15). The parameters $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$ constituting the minimal value (24) of the criterion functions $\Phi(\mathbf{v}[n+2])$ (22) were used for defining the prognostic model $t^*(\mathbf{x}[n])$ (27). The *CPL* prognostic model (27) allows to compute the predicted values t_j^* for the feature vectors $\mathbf{x}_j[n]$ belonging to the not censored data set C_0 (2):

$$(\forall \mathbf{x}_j[n] \in C_0)\ t_j^* = \mathbf{a}[n]^T \mathbf{x}_j[n] + b \tag{42}$$

where

$$\mathbf{a}[n] = \mathbf{w}^*[n]/\beta^* \ \ and \ \ b = w_0^*/\beta^* \tag{43}$$

We can compare the predicted values t_j^* with the observed values t_j from the not censored data set C_0 (2). In the case of the *classical regression*, the quality of the prognostic model (42) is evaluated on the basis of the sum of the squared differences $(t_j - t_j^*)^2$ between the observed variable t_j and the modelled t_j^* value (42).

We are using the absolute differences $|t_j - t_j^*|$ instead of the squared differences $(t_j - t_j^*)^2$ in the prognostic model (42) evaluation because the absolute differences $|t_j - t_j^*|$, in the same way as the criterion functions $\Phi(\mathbf{v}[n+2])$ (22) and $\Psi_\lambda(\mathbf{v}[n+2])$ (30) can be linked to the L_1 norm [2]. On the other hand, the squared differences $(t_j - t_j^*)^2$ are linked to the L_2 (*Euclidean*) norm [11].

The *discrepancy* coefficient Q_a of the prognostic model (42) is determined as the mean value of the absolute differences $|t_j - t_j^*|$:

$$Q_a = \sum_j |t_j - t_j^*|/m_0 \tag{44}$$

where the summation is over the all m_0 elements of the not censored data set C_0 (2).

The *CPL* prognostic model (42) appears as a result of attempted linear separation (16) of the positive set $\mathbf{Z}^+[n+2]$ from the negative set $\mathbf{Z}^-[n+2]$ (15). This linear separation (16) also allows to define the below linear classifier of the augmented vectors $\mathbf{z}_j[n+2]$ (39), (40), (41) [3]:

$$\begin{array}{l} if\ \mathbf{v}^*[n+2]^T \mathbf{z}_j[n+2] \geq 0, \\ \ \ then\ \mathbf{z}_j[n+2]\ is\ located\ in\ the\ set\ \mathbf{Z}^+[n+2] \\ if\ \mathbf{v}^*[n+2]^T \mathbf{z}_j[n+2] < 0, \\ \ \ then\ \mathbf{z}_j[n+2]\ is\ located\ in\ the\ set\ \mathbf{Z}^-[n+2] \end{array} \tag{45}$$

The quality of the linear classifiers (45) can be evaluated in the usual way by using the error estimator (*apparent error rate*) $e_a(\mathbf{v}^*[n+2])$ as the fraction of wrongly classified elements $\mathbf{z}_j[n+2]$ (39), (40), (41) of the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15):

$$e_a(\mathbf{v}^*[n+2]) = m_a(\mathbf{v}^*[n+2])/m \tag{46}$$

where m is the number of all elements $\mathbf{z}_j^+[n+2]$ (39), (40) of the set $\mathbf{Z}^+[n+2]$ (15) and the elements $\mathbf{z}_j^-[n+2]$ (39), (41) of the set $\mathbf{Z}^-[n+2]$ (15) and $m_a(\mathbf{v}^*[n+2])$ is the number of these elements which are wrongly allocated by the rule (45).

The optimal parameters $\mathbf{v}^*[n+2]$ in the classification rule (45) are obtained through the minimization of the criterion function $\Phi(\mathbf{v}[n+2])$ (22) which is defined on m elements $\mathbf{z}_j^+[n+2]$ and $\mathbf{z}_j^-[n+2]$ of the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15), adequately.

It is known that if the same vectors $\mathbf{z}_j[n+2]$ are used for classifier (45) designing and classifier evaluation (46), then the evaluation results are too optimistic (*biased*). In order to reduce the bias of the apparent error rate estimator $e_a(\mathbf{v}^*[n+2])$ (46) is usually replaced by the cross-validation error rate $e_{CVE}(\mathbf{v}^*[n+2])$ [2].

It was assumed in the earlier approach to the *CPL* prognostic model (27) evaluation that a "good" prognostic model should have the lowest error rate $e_{CVE}(\mathbf{v}^*[n+2])$ estimated through the cross-validation procedure [2]. The *discrepancy* coefficient Q_a (44) can also be used in the *CPL* prognostic model (42) evaluation. The *discrepancy* coefficient Q_a (44), similarly to the error rate $e_a(\mathbf{v}^*[n+2])$ can be a biased evaluation of the quality of the prognostic model $t^*(\mathbf{x}[n])$ (27). The cross-validation (*leave-one-out*) procedure can be used for the bias reduction of the *discrepancy* coefficient Q_a (44). The *leave-one-out* evaluation Q_{CVE} of the discrepancy coefficient can be defined for this purpose:

$$Q_{CVE} = \sum_i Q'(i)/m_0 \tag{47}$$

where the above summation is over all of the m_0 temporarily removed elements of the non-censored data set C_0 (2) and

$$Q'(i) = \sum_{j \in J(i)} |t_j - t_j^*(i)|/(m_0 - 1) \tag{48}$$

where the symbol $J(i)$ stands for the set of indices j of all $m_0 - 1$ elements $\mathbf{x}_j[n]$ of the set C_0 (2) apart from the i-th element $\mathbf{x}_i[n]$, which is temporarily removed.

The parameters $\mathbf{v}_i^*[n+2]$ of the prognostic model $t_j^*(i)$ (42) in the above expression (48) were determined through the minimization of the criterion function $\Phi_i(\mathbf{v}[n+2])$ (22) which was defined on all the $m_0 - 1$ elements of the set C_0 (2), except for the i-th element $\mathbf{x}_i[n]$.

In line with the modification proposed in this paper of the relaxed linear separability (*RLS*) method, the quality of particular subspaces $F_k[n_k]$ in the descending sequence (37) is evaluated based on the cross-validation values Q_{CVE} (47) of the discrepancy coefficient Q_a (44). It has been assumed here that a better feature subspace

$F_k[n_k]$ (37) allows to design *CPL* prognostic models $t^*(\mathbf{x}[n_k])$ (42) characterized by a lower cross-validation value Q_{CVE} (47) of the discrepancy coefficient.

The minimal value of the *discrepancy* coefficient Q_{CVE} (47) is used in this paper as the stop criterion for the process of feature space $F[n]$ reduction described by the descending sequence (37).

3. Experimental results and discussions

3.1 A toy model identification

The toy data set used in the experiment was generated by the authors. Seven points x_j $(j = 1,...,7)$ were arbitrarily selected on the line $(x_j \in R^1)$. The values t_j of dependent variable Y were generated for each point x_j in accordance with the below model (1):

$$(\forall j \in \{1,...,7\})\; t_j = 1 - x_j + \zeta_j \tag{49}$$

where the numbers ζ_j $(\zeta_j \in R^1)$ were generated in accordance with the normal probability distribution $(\zeta_j \sim N(0,\sigma))$ with the expected value zero and the variance σ^2 equal to three different values (0.3, 0.5, 0.7). The generated data sets are given in Table 1.

Table 1. Three toy data sets.

		Data 1	Data 2	Data 3
x_j	$t_j = 1 - x_j$	$t_j = 1 - x_j + \zeta_j$	$t_j = 1 - x_j + \zeta_j$	$t_j = 1 - x_j + \zeta_j$
		$(\sigma^2 = 0.3)$	$(\sigma^2 = 0.5)$	$(\sigma^2 = 0.7)$
-5	6	6.113	6.370	5.671
-4	5	5.237	3.627	5.018
-2	3	1.605	4.209	1.995
0	1	1.036	0.335	1.062
1	0	-0.456	-0.459	-0.419
3	-2	-2.544	-2.319	-0.514
5	-4	-4.302	-3.232	-3.741

In this case, the classical learning set C_0 (2) and the interval learning set C_I (3) have the below form:

$$C_1' = \{x_j, t_j,\; where\; j = 1,...,7\} \tag{50}$$

$$C_2' = \{x_j, [t_j - \varepsilon, t_j + \varepsilon],\; where\; j = 1,...,7\} \tag{51}$$

where $\varepsilon = 0.5$. The value of the parameter ε specifying the length of the interval $[t_j - \varepsilon, t_j + \varepsilon]$ was set to 0.5 in all experiments described in this subsection.

The prognostic (regression) model designed on the basis of the learning sets C'_1 (50) or C'_2 (51) has the following form, depending on parameters w_1 ($w_1 \in R^1$) and w_0 ($w_0 \in R^1$):

$$t(x) = w_1 x + w_0 \tag{52}$$

The parameters w_1 and w_0 were estimated based on the learning set C'_1 (50) by using the classical method of least squares [13]. The parameters w_1 and w_0 of the model (52) were also estimated based on the interval learning set C'_2 (51) through minimization (24) of the *CPL* criterion function $\Phi(\mathbf{v}[n+2])$ (22). The results of these experiments are shown in the Table 2 and the Figure 1.

Table 2. Parameters w_1 and w_0 of the model (52) estimated from the toy data sets.

	classical regression		interval regression	
	w_1	w_0	w_1	w_0
Data 1	-1.038	0.659	-1.060	0.801
Data 2	-0.956	0.946	-0.960	1.035
Data 3	-0.887	1.043	-0.941	0.965

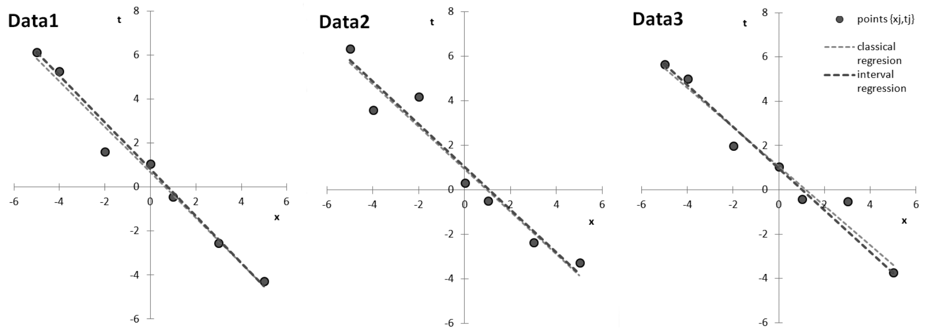

Fig. 1. The model $y = 1 - x$ identification by a classical regression and the model $t = 1 - x$ identification by an interval regression.

From Figure 1, it can be seen that the interval regression allowed to estimate the parameters w_1 and w_0 which are similar to the model of classical regression from this toy dataset. We can also note that he lowest identification quality of the model $t = 1 - x$ (45) has been obtained in the case of *Data 3* set which is characterized by the highest level of noise.

The *right censored* set C'_R (4) and the *left censored* set C'_L (5) have been also generated randomly from the toy data sets collected in the Table 1. The *indicator of censoring* δ_j ($\delta_j = 1$ or $\delta_j = 0$) has been used for this purpose. The value $\delta_j = 1$ implied that the interval $[t_j - \varepsilon, t_j + \varepsilon]$ (51) was censored to the form (4) or (5). In other words, if the value $\delta_j = 1$ appeared, the interval $[t_j - \varepsilon, t_j + \varepsilon]$ was replaced with equal probability $p = 0.5$ by $[-\infty, t_j + \varepsilon]$ or by $[t_j - \varepsilon, +\infty]$. The value $\delta_j = 0$ implied that the interval $[t_j - \varepsilon, t_j + \varepsilon]$ (51) was not changed. The censoring process was controlled by the parameter p_c called *probability of censoring* ($0 \leq p_c \leq 1$). The censoring ($\delta_j = 1$) was drawn for each interval $[t_j - \varepsilon, t_j + \varepsilon]$ (51) with the probability $p = p_c$. The *CPL* prognostic models (27) obtained for several values of the parameter p_c (*probability of censoring*) were sketched in Figure 2.

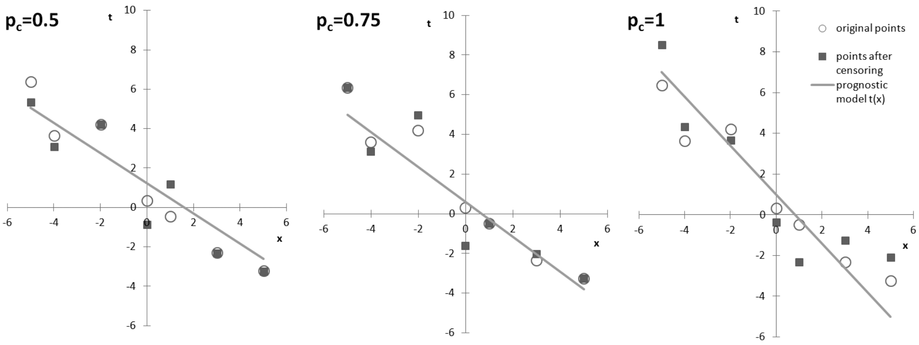

Fig. 2. The *CPL* prognostic models $t(x)$ (52) estimated from the toy data sets (Table 1) for several values of the censoring probability p_c.

We can remark that even learning sets with the full censoring ($p_c = 1$) allow to obtain a reasonably good prognostic model $t(x)$ (52).

3.2 Prognostic model selection on synthetic data

The synthetic data set contained $m = 100$ objects (feature vectors) $\mathbf{x}_j[n]$ ($j = 1, ..., 100$). Each object $\mathbf{x}_j[n]$ was represented by $n = 100$ features X_i ($i = 1, ..., 100$). The value x_{ji} of each feature X_i of particular object $\mathbf{x}_j[n]$ were drawn from a uniform distribution on the unit interval $[0, 1]$ ($X_i \in [0, 1]$). The value of the dependent variable t_j was computed as the bellow linear combination (*linear key*) with coefficients α_{ji}

of the selected components x_{ji} of the feature vector $\mathbf{x}_j[n]$:

$$\begin{aligned}&(\forall j \in \{1,...,m\})\\ &t_j = 3x_{j5} + 4x_{j10} + 7x_{j16} + 2x_{j37} + 6x_{j45} +\\ &\quad +3x_{j50} + 3x_{j67} + 8x_{j72} + x_{j84} + x_{j91} + 10\end{aligned} \tag{53}$$

The set of 10 features X_i and their coefficients α_{ji} in the above linear key were defined arbitrarily before the experiment. The linear key (53) was used for generating the classical regression learning set $C_0 = \{(\mathbf{x}_j[n]; y_j)\}$ (2).

Some of the dependent values y_j in this set C_0 were censored. The below scheme of censoring was adopted for the synthetic data set C_0. The censoring process was controlled, as in the case of the toy data set, by the parameter p_c ($0 \leq p_c \leq 1$). The censoring ($\delta_j = 1$) was drawn for each element $\mathbf{x}_j[n]$ of the learning set C_0 (2) with the probability p_c. As a result, m_c elements $\mathbf{x}_j[n]$ were selected to be censored. If the value $\delta_j = 1$ was drawn, the dependent value t_j was replaced by the censored value t_j^c ($0 \leq t_j^c \leq t_j$). The *right censored* value t_j^c was randomly generated in accordance with the triangle probability distribution determined on the interval $[0, t_j]$.

To allow the use of the *CPL* functions $\Phi(\mathbf{v}[n+2])$ (22) and $\Psi_\lambda(\mathbf{v}[n+2])$ (30) the non-censored elements $(\mathbf{x}_j[n]; t_j)$ of the classical learning set C_0 (2) were transformed in the interval learning set C_2' (51):

$$C_2' = \{\mathbf{x}_j[n], [t_j - \varepsilon, t_j + \varepsilon]\} \tag{54}$$

where $\varepsilon = 0.1$.

The *CPL* prognostic model (27) was defined in the experiment with the synthetic data set by the parameters $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$ (22) constituting the minimal value (24) of the criterion functions $\Phi(\mathbf{v}[n+2])$ (22), where $n = 100$. The *leave-one-out* evaluation of the *discrepancy* coefficients Q_a (44) and $e_a(\mathbf{v}^*[n+2])$ (46) were used in the experiment for the purpose of the bias reduction. In accordance with the leave-one-out procedure, the criterion functions $\Phi(\mathbf{v}[n+2])$ (22) was defined for each time by using $m - m_c - 1$ non censored elements $\mathbf{x}_j[n]$, because one non-censored element $\mathbf{x}_j[n]$ was used only for evaluating the resulting model (42). As a result, the criterion functions $\Phi(\mathbf{v}[n+2])$ (22) were defined on $2(m - m_c - 1) + m_c$ augmented vectors $\mathbf{z}_j^+[n+2]$ and $\mathbf{z}_j^-[n+2]$ (39), (40), (41).

The results of the experiments on the synthetic data set are shown in Figure 3. Two types of curves (the *learning curve* and the *testing curve*) are used to evaluate the *CPL* prognostic model (27) in different feature subspaces $F_k[n_k]$ (37), generated in accordance with the *RLS* method [4]. The same prognostic model (27) has been evaluated in a two manners (two curves) by using the same *discrepancy* coefficient

Q_a (44). The *learning curve* shows the averaged values of the coefficient Q_a (44) computed on all $m - m_c$ non censored elements $(\mathbf{x}_j[n];t_j)$ of the data set C_0 (2). The *testing curve* shows the averaged values of the coefficient Q_{CVE} (47) computed on m temporarily removed elements $(\mathbf{x}_{j'}[n];t_{j'})$ of the data set C_0 (2). In the case of the *learning curve*, the averaging is taking place from m calculations of the coefficient Q_a (44). In the case of the *testing curve*, the discrepancy coefficient Q_{CVE} (47) is computed in different feature subspaces $F_k[n_k]$ (37).

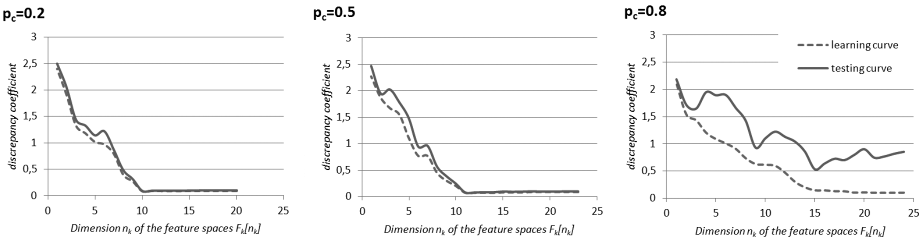

Fig. 3. The discrepancy evaluation Q_a (44) (learning curve) and Q_{CVE} (47) (testing curve) of the model (27) in different feature subspaces $F_k[n_k]$ (37) on the base of synthetic data with a few probabilities of censoring p_c.

We can note that the minimal value of the discrepancy coefficient Q_{CVE} (47) on the Figure 3 is located in the feature subspace $F_k[n_k]$ (37) of dimensionality n_k approximately equal to 10 ($n_k \approx 10$), as it was assumed in the linear key (53). Both the features X_i and their coefficients α_{ji} constituting the linear key (53) were approximately reproduced as a result of the *CPL* prognostic model designing. The linear key (53) was most accurately reproduced from the synthetic data set with the lowest probability of censoring $p_c = 0.2$.

The results of these experiments on the synthetic data set show the usefulness of a criterion based on the discrepancy coefficient Q_{CVE} (47) fore discovering the linear key (decision rule) separating two linear sets in a reasonably good manner.

3.3 Prognostic model selection based on the *Adrenocortical carcinoma* data set

The *Adrenocortical carcinoma* [7] data set consists of patient samples suffering from this type of cancer. The set contains 79 objects, each described by 20533 features (age, gender and gene expression values). Each object has a specified time value measured from start of observation until death (on average 915 days) or censoring (on average 1765 days). 51 patients (64.5%) were still alive at the final follow-up visit (censoring observations).

On the basis of 79 objects from the *Adrenocortical carcinoma* data set, 107 elements $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14) were created. The *RLS* method was applied to the newly formed data sets (15).

The main objective of this experiment was to examine the possibility of the *CPL* prognostic models $t^*(\mathbf{x}[n_k])$ (27) evaluation in different feature subspaces $F_k[n_k]$ (37) by using the discrepancy coefficient Q_{CVE} (47). More specifically, the possibility of using the minimal value of the discrepancy coefficient Q_{CVE} (47) as the stop criterion for the descending sequence of subspaces $F_k[n_k]$ (37) was examined. The optimal feature subspace $F_k^*[n_k]$ defined by this stop criterion was the last stage of the feature reduction procedure (35).

We can observe in Figure 4 that the minimal value of the discrepancy coefficient Q_{CVE} (47) has been reached in the feature subspace $F_k^*[n_k]$ (37) with dimensionality n_k of about 26 ($n_k \approx 26$).

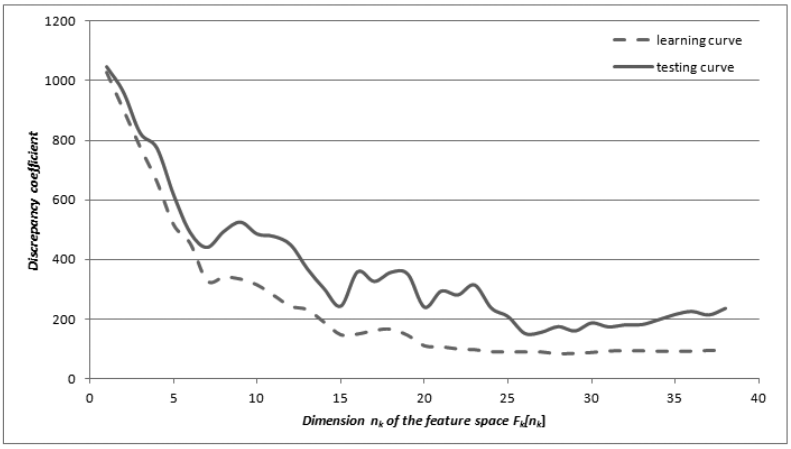

Fig. 4. The discrepancy evaluations Q_a (44) (learning curve) and Q_{CVE} (47) (testing curve) of the model (27) in different feature subspaces $F_k[n_k]$ (37) of the *Adrenocortical carcinoma* data set.

The prognostic model $t^*(\mathbf{x}[n_k])$ (27) was also evaluated, by using the cross validation error $e_{CVE}(\mathbf{v}^*[n+2])$ [2] of the linear classifier (45) of the augmented feature vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14). The optimal feature subspace $F_k^*[n_k]$ indicated in the sequence (37) by the minimal value of the discrepancy coefficient Q_{CVE} (47) is similar to the optimal feature subspace identified in cross validation error

$e_{CVE}(\mathbf{v}^*[n+2])$ (see Figure 5). This result might have some practical meaning, because it is an additional confirmation that our methodology is correct. The next time, the optimal feature subspace $F_k^*[n_k]$ was identified through an attempted linear separation of the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) of the augmented feature vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14). The augmented feature vectors $\mathbf{z}_j^+[n+2]$ (13) and $\mathbf{z}_j^-[n+2]$ (14) can represent both the non-censored (39), as well as censored cases (40), (41). As a result, the criterion based on the attempted linear separation of the sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) is less clear than the criterion based on the minimal discrepancy Q_{CVE} (47) intuitively assessed only on the non-censored cases (37).

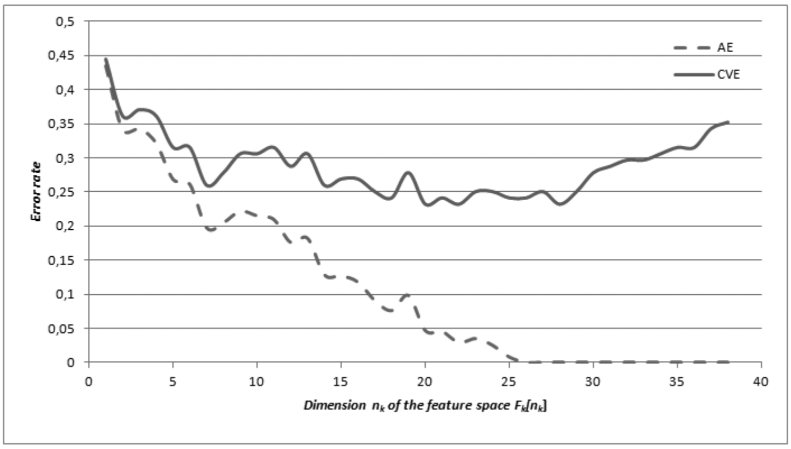

Fig. 5. The classifier error evaluations e_a (46) (AE) and e_{CVE} (CVE) of the model (27) in different feature subspaces $F_k[n_k]$ (37) of the *Adrenocortical carcinoma* data set.

4. Conclusions

Designing prognostic models (27) through exploring linear separability has been examined in the paper, based on examples of genetic data set with censored survival times. The task of the linear regression model designing has been reformulated as the problem of the linear separability of the augmented sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15). The augmented sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) allow to represent both the non-censored, as well as the censored learning sets in the regression analysis.

The exploration of the linear separability of the augmented sets $\mathbf{Z}^+[n+2]$ and $\mathbf{Z}^-[n+2]$ (15) has been performed through minimization (24) of the perceptron (*CPL*) criterion function $\Phi(\mathbf{v}[n+2])$ (22). The parameters $\mathbf{v}^*[n+2] = [\mathbf{w}^*[n]^T, w_0^*, \beta^*]^T$ constituting the minimum (24) of the criterion function $\Phi(\mathbf{v}[n+2])$ (22) have been used in the definition of the optimal prognostic model $t^*(\mathbf{x}[n])$ (27).

The modified *CPL* criterion function $\Psi_\lambda(\mathbf{v}[n+2])$ (30) has been used for generating a sequence of feature subspaces $F_k[n_k]$ (37) in accordance with the relaxed linear separability (*RLS*) method of feature subset selection (model selection).

The prognostic models $t^*(\mathbf{x}[n])$ (27) designed in different feature subspaces $F_k[n_k]$ (37) have been validated through the discrepancy coefficient Q_{CVE} (47) computed in accordance with the cross-validation (*leave one out*) procedure.

The proposed method of the prognostic models designing has been tested both on synthetic data set with the hidden linear key (49), as well as on real genetic data set *Adrenocortical carcinoma* [7] with censored survival times.

The experiments carried out on the genetic data set *Adrenocortical carcinoma* demonstrated, that the *RLS* method allows to find subsets of few genes X_i with good prognostic properties, even if the number of genes X_i is large at the beginning. The selection of optimal subsets of genes X_i was based on the minimal value of the discrepancy coefficient Q_{CVE} (47) computed in accordance with the leave-one-out procedure. The modified *RLS* method based on the discrepancy coefficient Q_{CVE} (47) has been proposed and applied for the purpose of the *CPL* prognostic models selection.

The linear key based on 10 variables X_i (53) was hidden in the synthetic data set composed of $n = 100$ variables (features) X_i. The *RLS* method allowed to find the model (53) hidden in the learning set containing $m = 100$ feature vectors $\mathbf{x}_j[n]$. The model $t(\mathbf{x}[n])$ (42) was approximately identified even when all values t_j of the dependent variable T were censored.

The linear prognostic model (27) have been designed in the reduced feature subspaces $F_k[n_k]$ (37) in a deterministic manner, even though the dimensionality of the genetic data sets was high and the survival times censored.

One of the promising results of the experiments is the possibility to use the discrepancy measure Q_{CVE} (47) in the modified *RLS* method of the *CPL* prognostic models selection. The stop criterion for the sequence (37) of reduced feature subspaces $F_k[n_k]$ can be based on the minimal value of discrepancy coefficient Q_{CVE} (47) (see Figure 4).

5. Acknowledgment

This work was partially supported by the S/WI/2/2018 grant from the Bialystok University of Technology and by the 4.2/St/2018 grant from the Institute of Biocybernetics and Biomedical Engineering PAS.

The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

References

[1] Christopher M Bishop. *Neural networks for pattern recognition.* Oxford University Press, 1995.

[2] Leon Bobrowski. *Data mining based on convex and piecewise linear (CPL) criterion functions (in Polish).* Bialystok University of Technology Press, 2005.

[3] Leon Bobrowski. Prognostic models based on linear separability. *Advances in Data Mining. Applications and Theoretical Aspects*, pages 11–24, 2011.

[4] Leon Bobrowski and Tomasz Łukaszuk. Relaxed linear separability (RLS) approach to feature (gene) subset selection. In *Selected works in bioinformatics*. InTech, 2011.

[5] Leon Bobrowski and Tomasz Łukaszuk. Prognostic modeling with high dimensional and censored data. In *Industrial Conference on Data Mining*, pages 178–193. Springer, 2012.

[6] Leon Bobrowski and Wojciech Niemiro. A method of synthesis of linear discriminant function in the case of nonseparability. *Pattern Recognition*, 17(2):205–210, 1984.

[7] Broad Institute TCGA Genome Data Analysis Center. Analysis overview for adrenocortical carcinoma (primary solid tumor cohort) - 28 january 2016, 2016.

[8] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

[9] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification.* John Wiley & Sons, 2012.

[10] Guadalupe Gómez, Anna Espinal, and Stephen W Lagakos. Inference for a linear regression model with an interval-censored covariate. *Statistics in medicine*, 22(3):409–425, 2003.

[11] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 4. Prentice-Hall New Jersey, 2014.

[12] John P Klein and Melvin L Moeschberger. Survival analysis: techniques for censored and truncated data. 1997.

[13] Charles L Lawson and Richard J Hanson. *Solving least squares problems.* SIAM, 1995.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Leon Bobrowski, Tomasz Łukaszuk

SELEKCJA CECH NA POTRZEBY MODELI PROGNOSTYCZNYCH POPRZEZ LINIOWĄ SEPARACJĘ ZBIORÓW DANYCH GENETYCZNYCH DOTYCZĄCYCH ANALIZY PRZEŻYCIA

Streszczenie W artykule rozważane jest projektowanie modeli regresji opartych na wysokowymiarowych (np. genetycznych) zbiorach danych poprzez badanie problemu separacji liniowej. Projektowanie modelu regresji liniowej zostało tu przeformułowane jako problem separacji liniowej. Eksploracja problemu separacji liniowej opiera się na minimalizacji wypukłej i odcinkowo-liniowej (CPL) funkcji kryterialnej. Minimalizacja funkcji kryterialnej typu CPL została wykorzystana nie tylko do oszacowania parametrów modelu prognostycznego, ale również do skutecznego wyboru podzbioru cech (selekcji modelu) zgodnie z metodą relaksacji separacji liniowej (RLS). Takie podejście do projektowania modeli prognostycznych zostało wykorzystane w eksperymentach zarówno z syntetycznymi danymi wielowymiarowymi, jak i do zbiorów danych genetycznych zawierających cenzurowane wartości zmiennej zależnej. Jakość modeli prognostycznych otrzymywanych w oparciu o postulat liniowej separacji została oceniona przy użyciu miary rozbieżności modelu i szacowanego wskaźnika błędu klasyfikacji. W celu zmniejszenia obciążenia oceny, obliczono wartości rozbieżności modelu i błędu klasyfikacji w różnych podprzestrzeniach cech, zgodnie z procedurą walidacji krzyżowej. Seria nowych eksperymentów opisanych w niniejszym opracowaniu pokazuje, że projektowanie modeli regresji może być oparte na zasadzie separacji liniowej. W szczególności, w procedurze projektowania można użyć wysokowymiarowych zbiorów genetycznych o cenzurowanej zmiennej zależnej. Proponowana miara rozbieżności modelu prognostycznego może być skutecznie wykorzystana w poszukiwaniu optymalnej podprzestrzeni cech i selekcji modelu regresji liniowej.

Słowa kluczowe: eksploracja danych, regresja interwałowa, selekcja modelu, relaksacja separacji liniowej

Artykuł zrealizowano w ramach pracy badawczej S/WI/2/2018.