Chapter

Computing on Vertices in Data Mining

Leon Bobrowski

Abstract

The main challenges in data mining are related to large, multi-dimensional data sets. There is a need to develop algorithms that are precise and efficient enough to deal with big data problems. The Simplex algorithm from linear programming can be seen as an example of a successful big data problem solving tool. According to the fundamental theorem of linear programming the solution of the optimization problem can found in one of the vertices in the parameter space. The basis exchange algorithms also search for the optimal solution among finite number of the vertices in the parameter space. Basis exchange algorithms enable the design of complex layers of classifiers or predictive models based on a small number of multivariate data vectors.

Keywords: data mining, basis exchange algorithms, small samples of multivariate vectors, gene clustering, prognostic models

1. Introduction

Various data mining tools are proposed to extract patterns from data sets [1]. Large, multidimensional data sets impose high requirements as to the precision and efficiency of calculations used to extract patterns (regularities) useful in practice [2]. In this context, there is still a need to develop new algorithms of data mining [3]. New types of patterns are also obtained in result of combining different types of classification or prognosis models [4].

The Simplex algorithm from linear programming is used as an effective big data mining tool [5]. According to the basic theorem of linear programming, the solution to the linear optimization problem with linear constraints can be found at one of the vertices in the parameter space. Narrowing the search area to a finite number of vertices is a source of the efficiency of the Simplex algorithm.

Basis exchange algorithms also look for an optimal solution among a finite number of vertices in the parameter space [6]. The basis exchange algorithms are based on the Gauss - Jordan transformation and, for this reason, are similar to the Simplex algorithm. Controlling the basis exchange algorithm is related to the minimization of convex and piecewise linear (CPL) criterion functions [7].

The perceptron and collinearity criterion functions belong to the family of CPL functions The minimization of the perceptron criterion function allows to check the linear separability of data sets and to design piecewise linear classifiers [8]. Minimizing the collinearity criterion function makes it possible to detect collinear (flat) patterns in data sets and to design multiple interaction models [9].

Data sets consisting of a small number of multivariate feature vectors generate specific problems in data mining [10]. This type of data includes genetic data sets. Minimizing the perceptron criterion function or the collinearity function enables solving problems related to discrimination or regression also in the case of a small set of multidimensional feature vectors by using complex layers of low dimensional linear classifiers or prognostic models [11].

2. Linear separability vs. linear dependence

Let us assume that each of m objects O_j from a given database were represented by the n-dimensional feature vector $\mathbf{x}_j = [x_{j,1},\dots,x_{j,n}]^T$ belonging to the feature space $\boldsymbol{F}[n]$ ($\mathbf{x}_j \in \boldsymbol{F}[n]$). The data set $\boldsymbol{C}$ consists of m such feature vectors $\mathbf{x}_j$:

$$\boldsymbol{C} = \{\mathbf{x}_j\}, \mathit{where}\, j = 1, \dots, m \tag{1}$$

The components $x_{j,i}$ of the feature vector $\mathbf{x}_j$ are numerical values ($x_{j,i} \in R$ or $x_{j,I} \in \{0, 1\}$) of the individual features X_i of the j-th object O_j. In this context, each feature vector $\mathbf{x}_j$ ($\mathbf{x}_j \in \boldsymbol{F}[n]$) represents n features X_i belonging to the feature set $F(n) = \{X_1, \dots, X_n\}$.

The pairs $\{\boldsymbol{G}_k^+, \boldsymbol{G}_k^=\}$ ($k = 1, \dots, K$) of the learning sets $\boldsymbol{G}_k^+$and $\boldsymbol{G}_k^=$ ($\boldsymbol{G}_k^+ \cap \boldsymbol{G}_k^- = \emptyset$) are formed from some feature vectors $\mathbf{x}_j$ selected from the data set $\boldsymbol{C}$ (1):

$$\boldsymbol{G}_k^+ = \{\mathbf{x}_j : j \in J_k^+\}, \mathit{and}\ \boldsymbol{G}_k^- = \{\mathbf{x}_j : j \in J_k^-\} \tag{2}$$

where J_k^+ and J_k^- are non-empty sets of indices j of vectors $\mathbf{x}_j$ ($J_k^+ \cap J_k^- = \emptyset$).

The *positive* learning set $\boldsymbol{G}_k^+$ is composed of m_k^+ feature vectors $\mathbf{x}_j$ ($j \in J_k^+$). Similarly, the *negative* learning set $\boldsymbol{G}_k^-$ is composed of m_k^- feature vectors $\mathbf{x}_j$ ($j \in J_k^-$), where $m_k^+ + m_k^- \le m$.

Possibility of the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) separation using a hyperplane $\boldsymbol{H}(\mathbf{w}_k, \theta_k)$ in the feature space $\boldsymbol{F}[n]$ is investigated in pattern recognition [1]:

$$\boldsymbol{H}(\mathbf{w}_k, \theta_k) = \{\mathbf{x} : \mathbf{w}_k^T\mathbf{x} = \theta_k\} \tag{3}$$

where $\mathbf{w}_k = [w_{k,1},\dots, w_{k,n}]^T \in \boldsymbol{R}^n$ is the weight vector, $\theta_k \in R^1$ is the threshold, and $\mathbf{w}_k^T\mathbf{x} = \Sigma_i\, w_{k,i}\, x_i$ is the scalar product.

Definition 1: The learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (1) are *linearly separable* in the feature space $\boldsymbol{F}[n]$, if and only if there exists a weight vector $\mathbf{w}_k$ ($\mathbf{w}_k \in \boldsymbol{R}^n$), and a threshold θ_k ($\theta_k \in R^1$) that the hyperplane $\boldsymbol{H}(\mathbf{w}_k, \theta_k)$ (3) separates these sets [7]:

$$(\exists \mathbf{w}_k, \theta_k)\ (\forall \mathbf{x}_j \in \boldsymbol{G}_k^+)\ \mathbf{w}_k^T\mathbf{x}_j \ge \theta_k + 1\ \mathit{and} \tag{4}$$

$$(\forall \mathbf{x}_j \in \boldsymbol{G}_k^-)\ \mathbf{w}_k^T\mathbf{x}_j \le \theta_k - 1$$

According to the above inequalities, all vectors $\mathbf{x}_j$ from the learning set $\boldsymbol{G}_k^+$ (2) are located on the positive side of the hyperplane $\boldsymbol{H}(\mathbf{w}_k, \theta_k)$ (3), and all vectors $\mathbf{x}_j$ from the set $\boldsymbol{G}_k^-$ lie on the negative side of this hyperplane.

The hyperplane $\boldsymbol{H}(\mathbf{w}_k, \theta_k)$ (3) separates (4) the sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (1) with the following margin $\delta_{L2}(\mathbf{w}_k)$ based on the Euclidean (L_2) norm which is used in the Support Vector Machines (SVM) method [12]:

$$\delta_{L2}(\mathbf{w}_k) = 2/\|\, \mathbf{w}_k \,\|_{L2} = 2/(\mathbf{w}_k^T\mathbf{w}_k)^{1/2} \tag{5}$$

where $\| \mathbf{w}_k \|_{L2} = (\mathbf{w}_k^T\mathbf{w}_k)^{1/2}$ is the Euclidean length of the weight vector $\mathbf{w}_k$.

The margin $\delta_{L1}(\mathbf{w}_k)$ with the L_1 norm related to the hyperplane $\boldsymbol{H}(\mathbf{w}_k, \theta_k)$ (2), which separates (10) the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) was determined by analogy to (5) as [11]:

$$\delta_{L1}(\mathbf{w}_k) = 2/\| \mathbf{w}_k \|_{L1} = 2/(|w_{k,1}| + \ldots + | w_{k,n}|) \tag{6}$$

where $\| \mathbf{w}_k \|_{L1} = |w_{k,1}| + \ldots +| w_{k,n}|$ is the L_1 length of the weight vector $\mathbf{w}_k$.

The margins $\delta_{\,L2}(\mathbf{w}_k)$ (5) or $\delta_{\,L1}(\mathbf{w}_k)$ (6) are maximized to improve the generalization properties of linear classifiers designed from the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) [7].

The following set of $m_k' = m_k^+ + m_k^-$ linear equations can be formulated on the basis of the linear separability inequalities (4):

$$(\forall j \in J_k^{+})\ \mathbf{x}_j^T\mathbf{w}_k = \theta_k + 1 \textit{ and} \tag{7}$$

$$(\forall j \in J_k^{-})\ \mathbf{x}_j^T\mathbf{w}_k = \theta_k - 1$$

If we assume that the threshold θ_k can be determined latter, then we have n unknown weights $w_{k,i}$ ($\mathbf{w}_k = [w_{k,1},\ldots, w_{k,n}]^T$) in an underdetermined system of $m_k' = m_k^+ + m_k^-$ ($m_k' \leq m_k < n$) linear Eqs. (7). In order to obtain a system of n linear Eqs. (7) with n unknown weights $w_{k,i}$, additional linear equations based on selected n - m_k' unit vectors $\mathbf{e}_i$ ($i \in I_k$) were taken into account [6]:

$$(\forall i \in I_k)\ \mathbf{e}_i^T\mathbf{w}_k = 0 \tag{8}$$

The parameter vertex $\mathbf{w}_k = [w_{k,1},\ldots, w_{k,n}]^T$ can be determined by the linear Eqs. (7) and (8) if the feature vectors $\mathbf{x}_j$ forming the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) are linearly independent [7].

The feature vector $\mathbf{x}_{j'}$ ($\mathbf{x}_{j'} \in \boldsymbol{G}_k^+ \cup \boldsymbol{G}_k^-$ (2)) is a linear combination of some other vectors $\mathbf{x}_{j(i)}$ ($j(i) \neq j'$) from the learning sets (2), if there are such parameters $\alpha_{\,j',i}$ ($\alpha_{j',i} \neq 0$) that the following relation holds:

$$\mathbf{x}_{j'} = \alpha_{j',1}\, \mathbf{x}_{j(1)} + \ldots + \alpha_{j',l}\, \mathbf{x}_{j(l)} \tag{9}$$

Definition 2: Feature vectors $\mathbf{x}_j$ making up the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) are linearly independent if neither of these vectors $\mathbf{x}_{j'}$ ($\mathbf{x}_{j'} \in \boldsymbol{G}_k^+ \cup \boldsymbol{G}_k^-$) can be expressed as a linear combination (9) of l ($l \in \{1, \ldots, m - 1\}$) other vectors $\mathbf{x}_{j(l)}$ from the learning sets.

If the number $m_k' = m_k^+ + m_k^-$ of elements $\mathbf{x}_j$ of the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) is smaller than the dimension n of the feature space $\boldsymbol{F}[n]$ ($m_k^+ + m_k^- \leq n$), then the parameter vertex $\mathbf{w}_k(\theta_k)$ can be defined by the linear equations in the following matrix form [13]:

$$\boldsymbol{B}_k\, \mathbf{w}_k(\theta_k) = \mathbf{1}_k(\theta_k) \tag{10}$$

where

$$\mathbf{1}_k(\theta_k) = [\theta_k + 1, \ldots, \theta_k + 1, \theta_k - 1, \ldots., \theta_k - 1, 0, \ldots., 0]^T \tag{11}$$

and

$$\boldsymbol{B}_k = \left[\mathbf{x}_1, \ldots, \mathbf{x}_{mk'}, \mathbf{e}_{i(mk'+1)}, \ldots., \mathbf{e}_{i(n)}\right]^T \tag{12}$$

The first m_k^+ components of the vector $\mathbf{1}_k(\theta_k)$ are equal to $\theta_k + 1$, the next m_k^- components equal to θ_k - 1, and the last n - $m_k^+ - m_k^-$ components are equal to 0. The first m_k^+ rows of the square matrix $\boldsymbol{B}_k$ (12) are formed by the feature vectors $\mathbf{x}_j$ ($j \in J_k^+$) from the set $\boldsymbol{G}_k^+$ (2), the next m_k^- rows are formed by vectors $\mathbf{x}_j$ ($j \in J_k^-$) from the set $\boldsymbol{G}_k^-$ (2), and the last n - $m_k^+ - m_k^-$ rows are made up of unit vectors $\mathbf{e}_j$ ($i \in I_k$):

If the matrix $\boldsymbol{B}_k$ (12) is non-singular, then there exists the inverse matrix $\boldsymbol{B}_k^{-1}$:

$$\boldsymbol{B}_k^{-1} = \left[\mathbf{r}_1, \ldots, \mathbf{r}_{mk'}, \mathbf{r}_{i(mk'+1)}, \ldots ., \mathbf{r}_{i(n)}\right] \tag{13}$$

In this case, the parameter vertex $\mathbf{w}_k(\theta_k)$ (10) can be defined by the following equation:

$$\begin{aligned}\mathbf{w}_k(\theta_k) = \boldsymbol{B}_k^{-1}\mathbf{1}_k(\theta_k) = (\theta_k + 1)\, \mathbf{r}_k^+ + (\theta_k - 1)\, \mathbf{r}_k^- = \\ = \theta_k\, (\mathbf{r}_k^+ + \mathbf{r}_k^-) + (\mathbf{r}_k^+ - \mathbf{r}_k^-)\end{aligned} \tag{14}$$

where the vector $\mathbf{r}_k^+$ is the sum of the first m_k^+ columns $\mathbf{r}_i$ of the inverse matrix $\boldsymbol{B}_k^{-1}$ (13), and the vector $\mathbf{r}_k^-$ is the sum of the successive m_k^- columns $\mathbf{r}_i$ of this matrix.

The last n - ($m_k^+ + m_k^-$) components $w_{k.i}(\theta_k)$ of the vector $\mathbf{w}_k(\theta_k) = [w_{k,1}(\theta_k), \ldots, w_{k,n}(\theta_k)]^T$ (14) linked to the zero components of the vector $\mathbf{1}_k(\theta_k)$ (11) are equal to zero:

$$(\forall i \in \{m_k^+ + m_k^- + 1, \ldots, n\})\ w_{k.i}(\theta_k) = 0 \tag{15}$$

The conditions $w_{k.i}(\theta_k) = 0$ (15) result from the equations $\mathbf{e}_i^T\mathbf{w}_k(\theta_k) = 0$ (8) at the vertex $\mathbf{w}_k(\theta_k)$ (14).

Length $||\mathbf{w}_k(\theta_k)||_{L1}$ of the weight vector $\mathbf{w}_k(\theta_k)$ (14) in the L_1 norm is the sum of $m_k' = m_k^+ + m_k^-$ components $|w_{k,i}(\theta_k)|$:

$$||\mathbf{w}_k(\theta_k)||_{L1} = |w_{k,1}(\theta_k)| + \ldots + |w_{k,mk'}(\theta_k)| \tag{16}$$

In accordance with the Eq. (14), components $|w_{k,i}(\theta_k)|$ can be determined as follows:

$$(\forall i \in \{1, \ldots, m_k^+ + m_k^-\})\ |w_{k,i}(\theta_k)| = |\ \theta_k\ (r_{k,i}^+ + r_{k,i}^-) + (r_{k,i}^+ - r_{k,i}^-)\ | \tag{17}$$

The length $||\mathbf{w}_k(\theta_k)||_{L1}$ (16) of the vector $\mathbf{w}_k(\theta_k)$ (14) with the L_1 norm is minimized to increase the margin $\delta_{L1}(\mathbf{w}_k(\theta_k))$ (6). The length $||\mathbf{w}_k(\theta_k)||_{L1}$ (16) can be minimized by selecting the optimal threshold value θ_k^* on the basis of the Eq. (14).

$$(\forall \theta_k)\ \delta_{L1}(\mathbf{w}_k(\theta_k^*)) \geq \delta_{L1}(\mathbf{w}_k(\theta_k)) \tag{18}$$

where the optimal vertex $\mathbf{w}_k(\theta_k^*)$ is defined by the Eq. (14).

Theorem 1: The learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) formed by m ($m \leq n$) linearly independent (9) feature vectors $\mathbf{x}_j$ are linearly separable (4) in the feature space $\boldsymbol{F}[n]$ ($\mathbf{x}_j \in \boldsymbol{F}[n]$).

Proof: If the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) are formed by m linearly independent feature.

vectors $\mathbf{x}_j$ then the non-singular matrix $\boldsymbol{B}_k = [\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{e}_{i(m+1)}, \ldots ., \mathbf{e}_{i(n)}]^T$ (12) containing these m vectors $\mathbf{x}_j$ and n - m unit vectors $\mathbf{e}_i$ ($i \in I_k$) can be defined [10]. In this case, the inverse matrix $\boldsymbol{B}_k^{-1}$ (13) exists and can determine the vertex $\mathbf{w}_k(\theta_k)$

(14). The vertex equation $\boldsymbol{B}_k\, \mathbf{w}_k(\theta_k) = \mathbf{1}_k(\theta_k)$ (10) can be reformulated for the feature vectors $\mathbf{x}_j$ (2) as follows:

$$\left(\forall \mathbf{x}_j \in \boldsymbol{G}_k^{\,+}\right)\, \mathbf{w}_k(\theta_k)^T \mathbf{x}_j = \theta_k + 1 \textit{ and } \left(\forall \mathbf{x}_j \in \boldsymbol{G}_k^{\,-}\right)\, \mathbf{w}_k(\theta_k)^T \mathbf{x}_j = \theta_k – 1 \tag{19}$$

The solution of the Eqs. (19) satisfies the linear separability inequalities (4).

It is possible to enlarge the learning sets G_k^+ and G_k^- (2) in such a way, which maintains their linear separability (4).

Lemma 1: Increasing the positive learning set $\boldsymbol{G}_k^+$ (2) by such a new vector $\mathbf{x}_{j'}$ ($\mathbf{x}_{j'} \notin \boldsymbol{G}_k^+$), which is a linear combination with the parameters $\alpha_{j',i}$ (9) of some feature vectors $\mathbf{x}_{j(l)}$ (2) from this set ($\mathbf{x}_{j(l)} \in \boldsymbol{G}_k^+$) preserves the linear separability (4) of the learning sets if the parameters $\alpha_{j',i}$ fulfill the following condition:

$$\alpha_{j',1} + \ldots + \alpha_{j',l} \geq 1 \tag{20}$$

If the assumptions of the lemma are met, then

$$\mathbf{w}_k^{\,T} \mathbf{x}_{j'} = \mathbf{w}_k^{\,T}\left(\alpha_{j',1}\, \mathbf{x}_{j(1)} + \ldots + \alpha_{j',l}\, \mathbf{x}_{j(l)}\right) = \tag{21}$$

$$= \alpha_{j',1}\, \mathbf{w}_k^{\,T} \mathbf{x}_{j(1)} + \ldots + \alpha_{j',l}\, \mathbf{w}_k^{\,T} \mathbf{x}_{j(l)} = \alpha_{j',1}\,(\theta_k + 1) + \ldots + \alpha_{j',l}\,(\theta_k + 1) \geq \theta_k + 1$$

The above inequality means that linear separability connditions (4) still apply after the increasing of the learning set $\boldsymbol{G}_k^+$ (2).

Lemma 2: Increasing the negative learning set $\boldsymbol{G}_k^-$ (2) by such a new vector $\mathbf{x}_{j'}$ ($\mathbf{x}_{j'} \notin G_k^-$), which is a linear combination with the parameters $\alpha_{j',i}$ (9) of some feature vectors $\mathbf{x}_{j(l)}$ (2) from this set ($\mathbf{x}_{j(l)} \in \boldsymbol{G}_k^-$) preserves the linear separability (4) of the learning sets if the parameters $\alpha_{j',\,i}$ fulfill the following condition:

$$\alpha_{j',1} + \ldots + \alpha_{j',l} \leq -1 \tag{22}$$

Justification Lemma 2 may be based on inequality similar to (21).

3. Perceptron criterion function

The minimization the perceptron criterion function allows to assess the degree of linear separabilty (4) of the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) in different feature subspaces $\boldsymbol{F}[n']$ ($\boldsymbol{F}[n'] \subset \boldsymbol{F}[n + 1]$) [6]. When defining the perceptron criterion function, it is convenient to use the following augmented feature vectors $\mathbf{y}_j$ ($\mathbf{y}_j \in \boldsymbol{F}[n + 1]$) and augmented weight vectors $\mathbf{v}_k$ ($\mathbf{v}_k \in \boldsymbol{R}^{n+1}$) [1]:

$$\left(\forall j \in J_k^+ \,(2)\right)\, \mathbf{y}_j = \left[\mathbf{x}_j^{\,T}, 1\right]^T, \tag{23}$$

$$\left(\forall j \in J_k^-(2)\right)\, \mathbf{y}_j = -\left[\mathbf{x}_j^{\,T}, 1\right]^T$$

and

$$\mathbf{v}_k = \left[\mathbf{w}_k^{\,T}, -\theta_k\right]^T = \left[\mathrm{w}_{k,1}, \ldots, \mathrm{w}_{k,n}, -\theta_k\right]^T \tag{24}$$

The augmented vectors $\mathbf{y}_j$ are constructed (23) on the basis of the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2). These learning sets are extracted from the data set $\boldsymbol{C}$ (1) according to some additional knowledge. The linear separability (4) of the learning sets $\boldsymbol{G}_k^+$

and $\boldsymbol{G}_{\mathrm{k}}^{-}$ (2) can be reformulated using the following set of m inequalities with the augmented vectors $\mathbf{y}_{\mathrm{j}}$ (23) [7]:

$$(\exists \mathbf{v}_{\mathrm{k}}) \left(\forall j \in J_{\mathrm{k}}^{+} \cup J_{\mathrm{k}}^{-} \ (2)\right) \mathbf{v}_{\mathrm{k}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{j}} \geq 1 \tag{25}$$

The dual hyperplanes $h_{\mathrm{j}}^{\mathrm{p}}$ in the parameter space $\boldsymbol{R}^{\mathrm{n}\,+\,1}$ ($\mathbf{v} \in \boldsymbol{R}^{\mathrm{n}\,+\,1}$) are defined on the basis of the augmented vectors $\mathbf{y}_{\mathrm{j}}$ [6]:

$$\left(\forall j \in J_{\mathrm{k}}^{+} \cup J_{\mathrm{k}}^{-} \ (2)\right) h_{\mathrm{j}}^{\mathrm{p}} = \left\{\mathbf{v} : \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} = 1\right\} \tag{26}$$

Dual hyperplanes $h_{\mathrm{j}}^{\mathrm{p}}$ (26) divide the parameter space $\boldsymbol{R}^{\mathrm{n}\,+\,1}$ ($\mathbf{v} \in \boldsymbol{R}^{\mathrm{n}\,+\,1}$) into a finite number L of disconnected regions (*convex polyhedra*) $\boldsymbol{D}_{l}^{\mathrm{p}}$ (l = 1, ... , L) [7]:

$$\boldsymbol{D}_{l}^{\mathrm{p}} = \left\{\mathbf{v} : (\forall j \in J_{l}^{+})\ \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} \geq 1 \textit{ and } \ (\forall j \in J_{l}^{-})\ \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} < 1\right\} \tag{27}$$

where J_{l}^{+} and J_{l}^{-} are disjointed subsets ($J_{l}^{+} \cap J_{l}^{+} = \varnothing$) of indices j of feature vectors $\mathbf{x}_{\mathrm{j}}$ making up the learning sets $\boldsymbol{G}_{\mathrm{k}}^{+}$ and $\boldsymbol{G}_{\mathrm{k}}^{-}$ (2).

The perceptron penalty functions $\varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v})$ are defined as follows for each of augmented feature vectors $\mathbf{y}_{\mathrm{j}}$ (23) [6]:

$$\begin{gathered}(\forall j \in J_{\mathrm{k}}) \\ \varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v}) = \begin{matrix} 1 - \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} & \textit{if} & \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} < 1 \\ 0 & \textit{if} & \mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} \geq 1 \end{matrix}\end{gathered} \tag{28}$$

The j - th penalty function $\varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v})$ (28) is greater than zero if and only if the weight vector $\mathbf{v}$ is located on the wrong side ($\mathbf{y}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{v} < 1$) of the j-th dual hyperplane $h_{\mathrm{j}}^{\mathrm{p}}$ (26). The function $\varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v})$ (28) is linear and greater than zero as long as the parameter vector $\mathbf{v} = [\mathrm{v}_{\mathrm{k},1},\ldots, \mathrm{v}_{\mathrm{k},\mathrm{n}\,+\,1}]^{\mathrm{T}}$ remains on the wrong side of the hyperplane $h_{\mathrm{j}}^{\mathrm{p}}$ (26). Convex and piecewise-linear (*CPL*) penalty functions $\varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v})$ (28) are used to enforce the linear separation (8) of the learning sets $\boldsymbol{G}_{\mathrm{k}}^{+}$ and $\boldsymbol{G}_{\mathrm{k}}^{-}$ (2).

The perceptron criterion function $\Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v})$ is defined as the weighted sum of the penalty functions $\varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v})$ (28) [6]:

$$\Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v}) = \Sigma_{\mathrm{j}}\ \alpha_{\mathrm{j}}\ \varphi_{\mathrm{j}}^{\mathrm{p}}(\mathbf{v}) \tag{29}$$

Positive parameters α_{j} ($\alpha_{\mathrm{j}} > 0$) can be treated as prices of individual feature vectors $\mathbf{x}_{\mathrm{j}}$:

$$(\forall j \in J_{\mathrm{k}}^{+}\ (2))\ \alpha_{\mathrm{j}} = 1/(2\, m_{\mathrm{k}}^{+}) \textit{ and } (\forall j \in J_{\mathrm{k}}^{-}\ (2))\ \alpha_{\mathrm{j}} = 1/(2\, m_{\mathrm{k}}^{-}) \tag{30}$$

where m_{k}^{+} (m_{k}^{-}) is the number of elements $\mathbf{x}_{\mathrm{j}}$ in the learning set $\boldsymbol{G}_{\mathrm{k}}^{+}$ ($\boldsymbol{G}_{\mathrm{k}}^{-}$) (2).

The perceptron criterion function $\Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v})$ (29) was built on the basis of the *error correction* algorithm, the basic algorithm in the *Perceptron* model of learning processes in neural networks [14].

The criterion function $\Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v})$ (29) is convex and piecewise-linear (*CPL*) [6]. It means, among others, that the function $\Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v})$ (29) remains linear within each area $\boldsymbol{D}_{l}$ (27):

$$\begin{gathered}(\forall l \in \{1, \ldots, L\}) \\ (\forall \mathbf{v} \in \boldsymbol{D}_{l})\ \Phi_{\mathrm{k}}^{\mathrm{p}}(\mathbf{v}) = \left(\Sigma_{\mathrm{j}}\ \alpha_{\mathrm{j}}\ \mathbf{y}_{\mathrm{j}}\right)^{\mathrm{T}}\end{gathered} \tag{31}$$

where the summation is performed on all vectors $\mathbf{y}_j$ (23) fulfilling the condition $\mathbf{y}_j^T\mathbf{v} < 1$.

The optimal vector $\mathbf{v}_k^*$ determines the minimum value $\Phi_k^P(\mathbf{v}_k^*)$ of the criterion function $\Phi_k^P(\mathbf{v})$ (29):

$$(\exists \mathbf{v}_k^*)\ (\forall \mathbf{v} \in \boldsymbol{R}^{n+1})\ \Phi_k^P(\mathbf{v}) \ge \Phi_k^P(\mathbf{v}_k^*) \ge 0 \tag{32}$$

Since the criterion function $\Phi_k^P(\mathbf{v})$ (29) is linear in each convex polyhedron $\boldsymbol{D}_l$ (27), the optimal point $\mathbf{v}_k^*$ representing the minimum $\Phi_k^P(\mathbf{v}_k^*)$ (32) can be located in selected vertex of some polyhedron $\boldsymbol{D}_{l'}^P$ (27). This property of the optimal vector $\mathbf{v}_k^*$ (32) follows from the *fundamental theorem of linear programming* [5].

It has been shown that the minimum value $\Phi_k^P(\mathbf{v}_k^*)$ (32) of the perceptron criterion function $\Phi_k^P(\mathbf{v})$ (29) with the parameters α_j (30) is normalized as follows [6]:

$$0 \le \Phi_k^P(\mathbf{v}_k^*) \le 1 \tag{33}$$

The below theorem has been proved [6]:

Theorem 2: The minimum value $\Phi_k^P(\mathbf{v}_k^*)$ (32) of the perceptron criterion function $\Phi_k^P(\mathbf{v})$ (29) is equal to zero ($\Phi_k^P(\mathbf{v}_k^*) = 0$) if and only if the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) are linearly separable (4).

The minimum value $\Phi_k^P(\mathbf{v}_k^*)$ (32) is near to one ($\Phi_k^P(\mathbf{v}_k^*) \approx 1$) if the sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2) cover almost completely. It can also be proved that the minimum value $\Phi_k^P(\mathbf{v}_k^*)$ (32) of the perceptron criterion function $\Phi_k^P(\mathbf{v})$ (29) does not depend on invertible linear transformations of the feature vectors $\mathbf{y}_j$ (23) [6]. The perceptron criterion function $\Phi_k(\mathbf{v})$ (29) remains linear inside of each region $\boldsymbol{D}_l$ (27).

The regularized criterion function $\Psi_k^P(\mathbf{v})$ is defined as the sum of the perceptron criterion function $\Phi_k^P(\mathbf{v})$ (29) and some additional penalty functions [13]. These additional *CPL* functions are equal to the costs γ_i ($\gamma_i > 0$) of individual features X_i multiply by the absolute values $|w_i|$ of weighs w_i, where $\mathbf{v} = [\mathbf{w}^T, -\theta]^T = [w_1,\ldots, w_n, -\theta]^T \in \boldsymbol{R}^{n+1}$ (24):

$$\Psi_k^P(\mathbf{v}) = \Phi_k^P(\mathbf{v}) + \lambda\ \Sigma_i\ \gamma_i|w_i| \tag{34}$$

where λ ($\lambda \ge 0$) is the *cost level*. The standard values of the cost parameters γ_i are equal to one ($\forall i \in \{1, \ldots, n\}\ \gamma_i = 1$).

The optimal vector $\mathbf{v}_{k,\lambda}^*$ constitutes the minimum value $\Psi_k^P(\mathbf{v}_{k,\lambda}^*)$ of the *CPL* criterion function $\Psi_k^P(\mathbf{v})$ (34), which is defined on elements $\mathbf{x}_j$ of the learning sets $\boldsymbol{G}_k^+$ and $\boldsymbol{G}_k^-$ (2):

$$(\exists \mathbf{v}_{k,\lambda}^*)\ (\forall \mathbf{v} \in R^{n+1})\ \Psi_k^P(\mathbf{v}) \ge \Psi_k^P(\mathbf{v}_{k,\lambda}^*) > 0 \tag{35}$$

Similarly as in the case of the perceptron criterion function $\Phi_k^P(\mathbf{v})$ (29), the optimal vector $\mathbf{v}_{k,\lambda}^*$ (35) can be located in selected vertex of some polyhedron $\boldsymbol{D}_{l'}$ (27). The minimum value $\Psi_k^P(\mathbf{v}_{k,\lambda}^*)$ (35) of the criterion function $\Psi_k^P(\mathbf{v})$ (34) is used, among others, in the *relaxed linear separability* (*RLS*) method of gene subsets selection [15].

4. Collinearity criterion function

Minimizing the collinearity criterion function is used to extract collinear patterns from large, multidimensional data sets $\boldsymbol{C}$ (1) [7]. Linear models of multivariate interactions can be formulated on the basis of representative collinear patterns [9].

The collinearity penalty functions $\varphi_j(\mathbf{w})$ are determined by individual feature vectors $\mathbf{x}_j = [x_{j,1},\ldots,x_{j,n}]^T$ in the following manner [9]:

$$(\forall \mathbf{x}_j \in \boldsymbol{C}\ (1))$$

$$\varphi_j(\mathbf{w}) = |\,1 - \mathbf{x}_j{}^T\mathbf{w}\,| = \begin{array}{ll} 1 - \mathbf{x}_j{}^T\mathbf{w} & \mathit{if}\ \ \mathbf{x}_j{}^T\mathbf{w} \leq 1 \\ \mathbf{x}_j{}^T\mathbf{w} - 1 & \mathit{if}\ \ \mathbf{x}_j{}^T\mathbf{w} > 1 \end{array} \tag{36}$$

The penalty functions $\varphi_j(\mathbf{w})$ (36) can be related to the following dual hyperplanes $h_j{}^1$ in the parameter (weight) space $\boldsymbol{R}^n$ ($\mathbf{w} \in \boldsymbol{R}^n$):

$$(\forall j = 1, \ldots, m)\ h_j{}^1 = \{\mathbf{w} : \mathbf{x}_j{}^T\mathbf{w} = 1\} \tag{37}$$

The *CPL* penalty $\varphi_j(\mathbf{w})$ (36) is equal to zero ($\varphi_j{}^c(\mathbf{w}) = 0$) in the point $\mathbf{w} = [w_1,\ldots, w_n]^T$ if and only if the point $\mathbf{w}$ is located on the dual hyperplane $h_j{}^1$ (37).

The collinearity criterion function $\Phi_k(\mathbf{w})$ is defined as the weighted sum of the penalty functions $\varphi_j(\mathbf{w})$ (36) determined by feature vectors $\mathbf{x}_j$ forming the data subset $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subset \boldsymbol{C}$ (1)):

$$\Phi_k(\mathbf{w}) = \Sigma_j\ \beta_j\ \varphi_j(\mathbf{w}) \tag{38}$$

where the sum takes into account only the indices J of the set $J_k = \{j: \mathbf{x}_j \in \boldsymbol{C}_k\}$, and the positive parameters β_j ($\beta_j > 0$) in the function $\Phi_k(\mathbf{w})$ (38) can be treated as the *prices* of particular feature vectors $\mathbf{x}_j$. The standard choice of the parameters β_j values is one ($(\forall j \in J_k)\ \beta_j = 1.0$).

The collinearity criterion function $\Phi_k(\mathbf{w})$ (38) is convex and piecewise-linear (*CPL*) as the sum of this type of penalty functions $\varphi_j(\mathbf{w})$ (36) [9]. The vector $\mathbf{w}_k{}^*$ determines the minimum value $\Phi_k(\mathbf{w}_k{}^*)$ of the criterion function $\Phi_k(\mathbf{w})$ (38):

$$(\exists \mathbf{w}_k{}^*)\ (\forall \mathbf{w})\ \Phi_k(\mathbf{w}) \geq \Phi_k(\mathbf{w}_k{}^*) \geq 0 \tag{39}$$

Definition 3: The data subset $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subset \boldsymbol{C}$ (1)) is *collinear* when all feature vectors $\mathbf{x}_j$ from this subset are located on some hyperplane $H(\mathbf{w}, \theta) = \{\mathbf{x}: \mathbf{w}^T\mathbf{x} = \theta\}$ with $\theta \neq 0$.

Theorem 3: The minimum value $\Phi_k{}^P(\mathbf{v}_k{}^*)$ (39) of the collinearity criterion function $\Phi_k(\mathbf{w})$ (38) defined on the feature vectors $\mathbf{x}_j$ constituting a data subset $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subset \boldsymbol{C}$ (1)) is equal to zero ($\Phi_k{}^P(\mathbf{v}_k{}^*) = 0$) when this subset $\boldsymbol{C}_k$ is collinear (*Def.* 3) [9].

Different collinear subsets $\boldsymbol{C}_k$ can be extracted from data set $\boldsymbol{C}$ (1) with a large number m of elements $\mathbf{x}_j$ by minimizing the collinearity criterion function $\Phi_k{}^P(\mathbf{w})$ (38) [9].

The minimum value $\Phi_k{}^P(\mathbf{v}_k{}^*)$ (39) of the collinearity criterion function $\Phi_k(\mathbf{w})$ (38) can be reduced to zero by omitting some feature vectors $\mathbf{x}_j$ from the data subset $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subset \boldsymbol{C}$ (1)). If the minimum value $\Phi_k(\mathbf{w}_k{}^*)$ (39) is greater than zero ($\Phi_k(\mathbf{w}_k{}^*) > 0$) then we can select feature vectors $\mathbf{x}_j$ ($j \in J_k(\mathbf{w}_k{}^*)$) with the penalty $\varphi_j(\mathbf{w}_k{}^*)$ (36) greater than zero:

$$(\forall j \in J_k(\mathbf{w}_k{}^*))\ \varphi_j(\mathbf{w}_k{}^*) = |1 - \mathbf{x}_j{}^T\ \mathbf{w}_k{}^*| > 0 \tag{40}$$

Omitting one feature vector $\mathbf{x}_{j'}$ ($j' \in J_k(\mathbf{w}_k{}^*)$) with the above property results in the following reduction of the minimum value $\Phi_k{}^P(\mathbf{v}_k{}^*)$ (39);

$$\Phi_{k'}(\mathbf{w}_{k'}{}^*) \leq \Phi_k(\mathbf{w}_k{}^*) - \varphi_{j'}(\mathbf{w}_k{}^*) \tag{41}$$

where $\Phi_{k'}(\mathbf{w}_{k'}{}^*)$ is the minimum value (39) of the collinearity criterion function $\Phi_{k'}(\mathbf{w})$ (38) defined on feature vectors $\mathbf{x}_j$ constituting the data subset C_k reduced by the vector $\mathbf{x}_{j'}$.

The regularized criterion function $\Psi_k(\mathbf{w})$ is defined as the sum of the collinearity criterion function $\Phi_k(\mathbf{w})$ (38) and some additional *CPL* penalty functions $\varphi_j{}^0(\mathbf{w})$ [7]:

$$\Psi_k(\mathbf{w}) = \Phi_k(\mathbf{w}) + \lambda\, \Sigma_i\, \chi_i(\mathbf{w}) = \Sigma_j\, \beta_j\, \varphi_j(\mathbf{w}) + \lambda\, \Sigma_i\, \chi_i\, \varphi_i{}^0(\mathbf{w}) \qquad (42)$$

where $\lambda \geq 0$ is the *cost level*. The standard values of the cost parameters γ_i are equal to one $((\forall i \in \{1, \dots, n\})\ \gamma_i = 1)$. The additional *CPL* penalty functions $\varphi_j{}^0(\mathbf{w})$ are defined below [7]:

$$(\forall i = 1, \dots, n) \qquad (43)$$

$$\chi_i(\mathbf{w}) = |\, \mathbf{e}_i{}^T\mathbf{w}\, | = \begin{matrix} -w_j & \mathit{if} & w_j \leq 0 \\ w_j & \mathit{if} & w_j > 0 \end{matrix}$$

The functions $\varphi_j{}^0(\mathbf{w})$ (43) are related to the following dual hyperplanes $h_j{}^0$ in the parameter (*weight*) space $\boldsymbol{R}^n$ ($\mathbf{w} \in \boldsymbol{R}^n$):

$$(\forall i = 1, \dots, n)\ h_j{}^0 = \{\mathbf{w} : \mathbf{e}_j{}^T\mathbf{w} = 0\} = \{\mathbf{w} : \mathbf{w}_j = 0\} \qquad (44)$$

The *CPL* penalty function $\varphi_j{}^0(\mathbf{w})$ (43) is equal to zero ($\varphi_j{}^0(\mathbf{w}) = 0$) in the point $\mathbf{w} = [w_1, \dots, w_n]^T$ if and only if this point is located on the dual hyperplane $h_j{}^0$ (44).

5. Parameter vertices

The perceptron criterion function $\Phi_k{}^P(\mathbf{v})$ (29) and the collinearity criterion function $\Phi_k(\mathbf{w})$ (38) are convex and piecewise-linear (*CPL*). The minimum values of a such *CPL* criterion functions can be located in parameter vertices of some convex polyhedra. We consider the parameter vertices $\mathbf{w}_k$ ($\mathbf{w}_k \in \boldsymbol{R}^n$) related to the collinearity criterion function $\Phi_k(\mathbf{w})$ (38).

Definition 4: The *parameter vertex* $\mathbf{w}_k$ of the *rank* r_k ($r_k \leq n$) in the weight space $\boldsymbol{R}^n$ ($\mathbf{w}_k \in \boldsymbol{R}^n$) is the intersection point of r_k hyperplanes $h_j{}^1$ (37) defined by linearly indepenedent feature vectors $\mathbf{x}_j$ ($j \in J_k$) from the data set $\boldsymbol{C}$ (1) and n - r_k hyperplanes $h_i{}^0$ (44) defined by unit vectors $\mathbf{e}_i$ ($i \in I_k$) [7].

The j-th dual hyperplane $h_j{}^1$ (37) defined by the feature vector $\mathbf{x}_j$ (1) passes through the k-th *vertex* $\mathbf{w}_k$ if the equation $\mathbf{w}_k{}^T\mathbf{x}_j = 1$ holds.

Definition 5: The k-th weight vertex $\mathbf{w}_k$ of the rank r_k is *degenerate* in the parameter space $\boldsymbol{R}^n$ if the number m_k of hyperplanes $h_j{}^1$ (37) passing through this vertex ($\mathbf{w}_k{}^T\mathbf{x}_j = 1$) is greater than the rank r_k ($m_k > r_k$).

The vertex $\mathbf{w}_k$ can be defined by the following set of n linear equations:

$$(\forall j \in J_k(\mathbf{w}_k))\ \mathbf{w}_k{}^T\mathbf{x}_j = 1 \qquad (45)$$

and

$$(\forall i \in I_k(\mathbf{w}_k))\ \mathbf{w}_k{}^T\mathbf{e}_i = 0 \qquad (46)$$

Eqs. (45) and (46) can be represented in the below matrix form [7]:

$$\mathbf{B}_k\, \mathbf{w}_k = \mathbf{1}_k \qquad (47)$$

where $\mathbf{1}_k = [1, \ldots, 1, 0, \ldots, 0]^T$ is the vector with the first r_k components equal to one and the remaining $n - r_k$ components are equal to zero.

The square matrix $\boldsymbol{B}_k$ (47) consists of k feature vectors $\mathbf{x}_j$ ($j \in J_k$ (45)) and $n - k$ unit vectors $\mathbf{e}_i$ ($i \in I_k$ (46)) []:

$$\boldsymbol{B}_k = \left[\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{e}_{i(k+1)}, \ldots, \mathbf{e}_{i(n)}\right]^T \tag{48}$$

where the symbol $\mathbf{e}_{i(l)}$ denotes such unit vector, which is the l-th row of the matrix $\boldsymbol{B}_k$.

Since feature vectors $\mathbf{x}_j$ ($\forall j \in J_k(\mathbf{w}_k)$ (45)) making up r_k rows of the matrix $\boldsymbol{B}_k$ (48) are linearly independent, then the inverse matrix $\boldsymbol{B}_k^{-1}$ exists:

$$\boldsymbol{B}_k^{-1} = \left[\mathbf{r}_1, \ldots, \mathbf{r}_k, \mathbf{r}_{i(k+1)}, \ldots ., \mathbf{r}_{i(n)}\right] \tag{49}$$

The inverse matrix $\boldsymbol{B}_k^{-1}$ (49) can be obtained starting from the unit matrix $\boldsymbol{I} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]^T$ and using the basis exchange algorithm [8].

The non-singular matrix $\mathbf{B}_k$ (48) is the *basis* of the feature space $\boldsymbol{F}[n]$ related to the vertex $\mathbf{w}_k = [w_{k,1}, \ldots, w_{k,n}]^T$. Since the last $n - r_k$ components of the vector $\mathbf{1}_k$ (47) are equal to zero, the following equation holds:

$$\mathbf{w}_k = \boldsymbol{B}_k^{-1}\mathbf{1}_k = \mathbf{r}_1 + \ldots + \mathbf{r}_k \tag{50}$$

According to Eq. (50), the weight vertex $\mathbf{w}_k$ is the sum of the first k columns $\mathbf{r}_i$ of the inverse matrix $\boldsymbol{B}_k^{-1}$ (49).

Remark 1: The $n - k$ components $w_{k.i}$ of the vector $\mathbf{w}_k = [w_{k,1}, \ldots, w_{k,n}]^T$ (50) linked to the zero components of the vector $\mathbf{1}_k = [1, \ldots, 1, 0, \ldots ., 0, 1]^T$ (7) are equal to zero:

$$(\forall i \in \{k+1, \ldots, n\})\ w_{k.i} = 0 \tag{51}$$

The conditions $w_{k.i} = 0$ (51) result from the equations $\mathbf{w}_k^T\mathbf{e}_i = 0$ (46) at the vertex $\mathbf{w}_k$.

The *fundamental theorem of linear programming* shows that the minimum $\Phi_k(\mathbf{w}_k^*)$ (39) of the *CPL* collinearity criterion function $\Phi_k(\mathbf{w})$ (38) can always be located in one of the vertices $\mathbf{w}_k$ (50) [5]. The same property has also the regularized criterion function $\Psi_k(\mathbf{w})$ (42), another function of the *CPL* type [7].

We can see that all such feature vectors $\mathbf{x}_j$ (1) which define hyperplanes h_j^1 (37) passing through the vertex $\mathbf{w}_k$ are located on the hyperplane $\boldsymbol{H}(\mathbf{w}_k, 1) = \{\mathbf{x}: \mathbf{w}_k^T\mathbf{x} = 1\}$ (3) in the feature space $\boldsymbol{F}[n]$. A large number m_k of feature vectors $\mathbf{x}_j$ (1) located on the hyperplane $H(\mathbf{w}_k, 1)$ (3) form the *collinear cluster* $\boldsymbol{C}(\mathbf{w}_k)$ based on the vertex $\mathbf{w}_k$ [8]:

$$\boldsymbol{C}(\mathbf{w}_k) = \left\{\mathbf{x}_j \in \boldsymbol{C}\ (1) : \mathbf{w}_k^T\mathbf{x} = 1\right\} \tag{52}$$

If the vertex $\mathbf{w}_k$ of the rank r_k is degenerate in the parameter space $\boldsymbol{R}^n$ then the collinear cluster $\boldsymbol{C}(\mathbf{w}_k)$ (52) contains more than r_k feature vectors $\mathbf{x}_j$ (1).

The k-th vertex $\mathbf{w}_k = [w_{k,1}, \ldots, w_{k,n}]^T$ in the parameter space $\boldsymbol{R}^n$ ($\mathbf{w}_k \in \boldsymbol{R}^n$) is linked by the Eq. (47) to the non-singular matrix $\mathbf{B}_k$ (48). The rows of the matrix $\mathbf{B}_k$ (48) can form the *basis* of the feature space $\boldsymbol{F}[n]$. The conditions $w_{k.i} = 0$ (51) result from the equations $\mathbf{w}_k^T\mathbf{e}_i = 0$ (46) at the vertex $\mathbf{w}_k$.

$$(\forall i = 1, \ldots, n)\ \textit{if}\ (\mathbf{e}_i \in \boldsymbol{B}_k\ (48)), \textit{then}\ w_{k.i} = 0 \tag{53}$$

Each feature vector $\mathbf{x}_j$ from the data set $\boldsymbol{C}$ (1) represents n features X_i belonging to the feature set $R(n) = \{X_1, \dots , X_n\}$. The k-th *vertexical feature subset* $R_k(r_k)$ consists of r_k features X_i that are connected to the weights $w_{k.i}$ different from zero ($w_{k.i} \neq 0$):

$$R_k(r_k) = \left\{X_{i(1)}, \dots , X_{i(rk)}\right\} \tag{54}$$

The k-th *vertexical subspace* $\boldsymbol{F}_k[r_k]$ ($\boldsymbol{F}_k[r_k] \subset \boldsymbol{F}[n]$) contains the reduced vectors $\mathbf{x}_j[r_k]$ with r_k componets $x_{j,i(l)}$ ($\mathbf{x}_j[r_k] \in \boldsymbol{F}_k[r_k]$) related to the weights $w_{k.i}$ different from zero:

$$(\forall j \in \{1, \dots , m\})\ \ \mathbf{x}_j[r_k] = \left[x_{j,i(1)}, \dots , x_{j,i(rk)}\right]^T \tag{55}$$

The reduced vectors $\mathbf{x}_j[r_k]$ (55) are obtained from the feature vectors $\mathbf{x}_j = [x_{j,1},\dots,x_{j,n}]^T$ belonging to the data set $\boldsymbol{C}$ (1) by omitting the n - r_k components $x_{j,i}$ related to the weights $w_{k.i}$ equal to zero ($w_{k.i} = 0$).

We consider the optimal vertexical subspace $\boldsymbol{F}_k{}^*[r_k]$ ($\boldsymbol{F}_k{}^*[r_k] \subset \boldsymbol{F}[n]$) related to the reduced optimal vertex $\mathbf{w}_k{}^*[r_k]$ which determines the minimum $\Phi_k(\mathbf{w}_k{}^*)$ (39) of the collinearity criterion function $\Phi_k(\mathbf{w})$ (38). The optimal collinear cluster $\boldsymbol{C}(\mathbf{w}_k{}^*[r_k])$ (52) is based on the optimal vertex $\mathbf{w}_k{}^*[r_k] = [w_{k,1}{}^*, \dots , w_{k,rk}{}^*]^T$ with r_k different from zero components $w_{k,i}{}^*$ ($w_{k.i}{}^* \neq 0$). Feature vectors $\mathbf{x}_j$ belonging to the collinear cluster $\boldsymbol{C}(\mathbf{w}_k{}^*)$ (52) satisfy the equations $\mathbf{w}_k{}^*[r_k]^T\mathbf{x}_j[r_k] = 1$, hence:

$$\begin{gathered}\left(\forall \mathbf{x}_j \in \boldsymbol{P}(\mathbf{w}_k{}^*)\right)\\ w_{k.1}{}^*\ x_{j,i(1)} + \dots + w_{k.rk}{}^*\ x_{j,i(rk)} = 1\end{gathered} \tag{56}$$

where $x_{j,i(l)}$ are components of the j-th feature vectors $\mathbf{x}_j$ related to the weights $w_{k.i}$ different from zero ($w_{k.i} \neq 0$).

A large number m_k of feature vectors $\mathbf{x}_j$ (1) belonging to the collinear cluster $\boldsymbol{C}(\mathbf{w}_k{}^*[r_k])$ (52) justifies the following collinear model of interaction between selected features $X_{i(l)}$ which is based on the Eqs. (56) [9]:

$$w_{k.1}{}^*\ X_{i(1)} + \dots + w_{k.rk}{}^*\ X_{i(rk)} = 1 \tag{57}$$

The collinear interaction model (57) allows, inter alia, to design the following prognostic models for each feature $X_{i'}$ from the subset $R_k(r_k)$ (54):

$$(\forall i' \in \{1, \dots , r_k)\ X_{i'} = \alpha_{i',0} + \alpha_{i',1}\ X_{i(1)} + \dots + \alpha_{i',rk}\ X_{i(rk)} \tag{58}$$

where $\beta_{i',0} = 1 / w_{k.i'}{}^*$, $\beta_{i',\ i'} = 0$, and $(\forall\ i(l) \neq i')$ $\beta_{i',i(l)} = w_{k.i(l)}{}^* / w_{k.i'}{}^*$.

Feature $X_{i'}$ is a dependent variable in the prognostic model (58), the remaining m - 1 features $X_{i(l)}$ are independent variables ($i(l) \neq i'$). The family of r_k prognostic models (58) can be designed on the basis of one collinear interaction model (57). Models (58) have a better justification for a large number m_k of feature vectors $\mathbf{x}_j$ (1) in the collinear cluster $\boldsymbol{C}(\mathbf{w}_k{}^*[r_k])$ (52).

6. Basis exchange algorithm

The collinearity criterion function $\Phi(\mathbf{w})$ (38), like other convex and piecewise linear (*CPL*) criterion functions, can be minimized using the basis exchange algorithm [8]. The basis exchange algorithm aimed at minimization of the collinearity criterion function $\Phi(\mathbf{w})$ (38) is described below.

According to the basis exchange algorithm, the optimal vertex $\mathbf{w}_k^*$, which constitutes the minimum value $\Phi_k(\mathbf{w}_k^*)$ (39) of the collinearity function $\Phi_k(\mathbf{w})$ (38), is achieved after a finite number L of the steps l as a result of guided movement between selected vertices $\mathbf{w}_k$ (50) [8]:

$$\mathbf{w}_0 \rightarrow \mathbf{w}_1 \rightarrow \ \dots . \rightarrow \mathbf{w}_L \tag{59}$$

The sequence of vertices $\mathbf{w}_k$ (59) is related by (47) to the following sequence of the inverse matrices $\boldsymbol{B}_k^{-1}$ (49):

$$\boldsymbol{B}_0^{-1} \rightarrow \boldsymbol{B}_1^{-1} \rightarrow \ \dots . \rightarrow \boldsymbol{B}_L^{-1} \tag{60}$$

The sequence of vertices $\mathbf{w}_{k(l)}$ (59) typically starts at the vertex $\mathbf{w}_0 = [0,\dots, 0]^T$ related to the identity matrix $\boldsymbol{B}_0 = \boldsymbol{I}_n = [\mathbf{e}_1,\dots, \mathbf{e}_n]^T$ of the dimension n x n [7]. The final vertex $\mathbf{w}_L$ (59) should assure the minimum value of the collinearity criterion function $\Phi(\mathbf{w})$ (38):

$$(\forall \mathbf{w}) \ \Phi(\mathbf{w}) \geq \Phi(\mathbf{w}_L) \geq 0 \tag{61}$$

If the criterion function $\Phi(\mathbf{w})$ (38) is defined on m ($m \leq n$) linearly independent vectors $\mathbf{x}_j$ ($\mathbf{x}_j \in \boldsymbol{C}$ (1)) then the value $\Phi(\mathbf{w}_L)$ of the collinearity criterion function $\Phi(\mathbf{w})$ (38) at the final vertex $\mathbf{w}_L$ (59) becomes zero ($\Phi(\mathbf{w}_L) = 0$) [8]. The rank r_L (*Def.* 4) of the final vertex $\mathbf{w}_L$ (59) can be equal to the number m of feature vectors $\mathbf{x}_j$ ($r_L = m$) or it can be less than m ($r_L < m$). The rank r_L of the final vertex $\mathbf{w}_L$ (59) is less than m ($r_L < m$) if the final vertex $\mathbf{w}_L$ is degenerate [7].

Consider the reversible matrix $\boldsymbol{B}_k = [\mathbf{x}_1,\dots, \mathbf{x}_k, \mathbf{e}_{i(k + 1)},\dots, \mathbf{e}_{i(n)}]^T$ (48), which determines the vertex $\mathbf{w}_k$ (50) and the value $\Phi_k(\mathbf{w}_k)$ of the criterion function $\Phi_k(\mathbf{w})$ (38) in the k-th step. In the step (l + 1), one of the unit vectors $\mathbf{e}_i$ in the matrix $\mathbf{B}_k$ (48) is replaced by the feature vector $\mathbf{x}_{k + 1}$ and the matrix $\mathbf{B}_{k + 1} = [\mathbf{x}_1,\dots, \mathbf{x}_k, \mathbf{x}_{k + 1}, \mathbf{e}_{i(k + 2)},\dots, \mathbf{e}_{i(n)}]^T$ appears. The unit vector $\mathbf{e}_{i(k + 1)}$ leaving matrix $\boldsymbol{B}_k$ (48) is indicated by an *exit criterion* based on the gradient of the collinearity criterion function $\Phi(\mathbf{w})$ (38) [7]. The exit criterion allows to determine the exit edge $\mathbf{r}_{k + 1}$ (49) of the greatest descent of the collinearity criterion function $\Phi(\mathbf{w})$ (38). As a result of replacing the unit vector $\mathbf{e}_{i(k + 1)}$ with the feature vector $\mathbf{x}_{k + 1,}$ the value $\Phi(\mathbf{w}_k)$ of the collinearity function $\Phi(\mathbf{w})$ (38) decreases (41):

$$\Phi(\mathbf{w}_{k+1}) \leq \Phi(\mathbf{w}_k) - \varphi_{k+1}(\mathbf{w}_k) \tag{62}$$

After a finite number L ($L \leq m$) of the steps k, the collinearity function $\Phi(\mathbf{w})$ (38) reaches its minimum (61) at the final vertex $\mathbf{w}_L$ (59).

The sequence (60) of the inverse matrices $\boldsymbol{B}_k^{-1}$ is obtained in a multi-step process of minimizing the function $\Phi(\mathbf{w})$ (38). During the k-th step, the matrix $\boldsymbol{B}_{k\text{-}1} = [\mathbf{x}_1, \dots , \mathbf{x}_{k\text{-}1}, \mathbf{e}_{i(k)}, \dots ., \mathbf{e}_{i(n)}]^T$ (12) is transformed into the matrix $\boldsymbol{B}_k$ by replacing the unit vector $\mathbf{e}_{i(k)}$ with the feature vector $\mathbf{x}_{k:}$

$$(\forall k \in \{1, \ \dots , L\}) \ \boldsymbol{B}_{k-1} \rightarrow \boldsymbol{B}_k \tag{63}$$

According to the vector Gauss-Jordan transformation, replacing the unit vector $\mathbf{e}_{i(k)}$ with the feature vector $\mathbf{x}_k$ during the k - th stage results in the following modifications of the co + lumns $\mathbf{r}_i(k)$ of the inverse matrix $\boldsymbol{B}_l^{-1} = [\mathbf{x}_1, \dots , \mathbf{x}_l, \mathbf{e}_{i(l+1)}, \dots , \mathbf{e}_{i(n)}]^T$ (49) [6]:

$$\mathbf{r}_{i(l+1)}(l+1) = \left(1/\mathbf{r}_{i(l+1)}(l)^T \mathbf{x}_{l+1}\right) \mathbf{r}_{i(l+1)}(l) \tag{64}$$

and

$$(\forall i \neq i(l+1))\ \mathbf{r}_{\mathrm{i}}(l+1) = \mathbf{r}_{\mathrm{i}}(l) - \left(\mathbf{r}_{\mathrm{i}}(l)^{\mathrm{T}}\mathbf{x}_{l+1}\right)\mathbf{r}_{\mathrm{i(l)}}(l+1) =$$
$$= \mathbf{r}_{\mathrm{i}}(l) - \left(\mathbf{r}_{\mathrm{i}}(l)^{\mathrm{T}}\mathbf{x}_{\mathrm{j}(l+1)}/\mathbf{r}_{\mathrm{i}(l)}(l)^{\mathrm{T}}\mathbf{x}_{l+1}\right)\mathbf{r}_{\mathrm{i}(l)}(l)$$

where $i(l+1)$ is the index of the unit vector $\mathbf{e}_{\mathrm{i}(l+1)}$ leaving the basis $\boldsymbol{B}_l = [\mathbf{x}_1,\dots, \mathbf{x}_l, \mathbf{e}_{\mathrm{i}(l+1)},\dots, \mathbf{e}_{\mathrm{i}(n)}]^{\mathrm{T}}$ during the l-th stage.

Remark 2: The vector Gauss-Jordan transformation (64) resulting from the replacing of the unit vector $\mathbf{e}_{\mathrm{i(k)}}$ with the feature vector $\mathbf{x}_{\mathrm{k}}$ in the basis $\boldsymbol{B}_{\mathrm{k-1}} = [\mathbf{x}_1,\dots, \mathbf{x}_{\mathrm{k-1}}, \mathbf{e}_{\mathrm{i(k)}},\dots, \mathbf{e}_{\mathrm{i}(n)}]^{\mathrm{T}}$ cannot be executed when the below *collinearity condition* is met [7]:

$$\mathbf{r}_{\mathrm{i(k)}}(k)^{\mathrm{T}}\ \mathbf{x}_{\mathrm{k}} = 0 \tag{65}$$

The collinearity condition (65) causes a division by zero in Eq. (64).

Let the symbol $\mathbf{r}_l[k]$ denote the l-th column $\mathbf{r}_l(k) = [\mathrm{r}_{l,1}(k),\dots, \mathrm{r}_{l,\mathrm{n}}(k)]^{\mathrm{T}}$ of the inverse matrix $\boldsymbol{B}_{\mathrm{k}}^{-1} = [\mathbf{r}_1(k), \dots, \mathbf{r}_{\mathrm{k-1}}(k), \mathbf{r}_{\mathrm{k}}(k), \dots., \mathbf{r}_n(k)]$ (49) after the reduction of the last n - k components $\mathrm{r}_{l,\mathrm{i}}(k)$:

$$\mathbf{r}_l[k] = [\mathrm{r}_{l,1}(k),\ \dots, \mathrm{r}_{l,\mathrm{k}}(k)]^{\mathrm{T}} \tag{66}$$

Similarly, the symbol $\mathbf{x}_{\mathrm{j}}[k] = [\mathrm{x}_{\mathrm{j,1}},\dots,\mathrm{x}_{\mathrm{j,k}}]^{\mathrm{T}}$ means the reduced vector obtained from the feature vector $\mathbf{x}_{\mathrm{j}} = [\mathrm{x}_{\mathrm{j,1}},\dots,\mathrm{x}_{\mathrm{j,n}}]^{\mathrm{T}}$ after he reducing of the last n - k components $\mathrm{x}_{\mathrm{j,i}}$:

$$(\forall j \in \{1,\ \dots, m)\ \mathbf{x}_{\mathrm{j}}[k] = \left[\mathrm{x}_{\mathrm{j,1}},\ \dots, \mathrm{x}_{\mathrm{j},k}\right]^{\mathrm{T}} \tag{67}$$

Lemma 3: The collinearity condition (65) appears during the k-th step when the reduced vector $\mathbf{x}_{\mathrm{k}}[k]$ (66) is a linear combination of the basis reduced vectors $\mathbf{x}_{\mathrm{j}}[k]$ (67) with $j < k$:

$$\mathbf{x}_{\mathrm{k}}[k] = \alpha_1\ \mathbf{x}_1[k] +\ \dots\ + \alpha_{\mathrm{k}-1}\ \mathbf{x}_{\mathrm{k}-1}[k] \tag{68}$$

where $(\forall i \in \{1, \dots, k - 1\})\ \alpha_{\mathrm{i}} \in R^1$.

The proof of this lemma results directly from the collinearity condition (65) [7].

7. Small samples of multivariate feature vectors

A small sample of multivariate vectors appears when the number m of feature vectors $\mathbf{x}_{\mathrm{j}}$ in the data set $\boldsymbol{C}$ (1) is much smaller than the dimension n of these vectors ($m << n$). The basis exchange algorithms allows for efficient minimization of the *CPL* criterion functions also in the case of small samples of multivariate vectors [10]. However, for small samples, some new properties of the basis exchange algorithms are more important. In particular, the regularization (42) of the *CPL* criterion functions becomes crucial. New properties of the basis exchange algorithms in the case of a small number m of multidimensional feature vectors $\mathbf{x}_{\mathrm{j}}$ (1) is discussed on the example of the collinearity criterion function $\Phi(\mathbf{w})$ (38) and the regularized criterion function $\Psi(\mathbf{w})$ (42).

Lemma 4: The value $\Phi(\mathbf{w}_{\mathrm{K}})$ of the collinearity criterion function $\Phi(\mathbf{w})$ (38) at the final vertex $\mathbf{w}_L$ (59) is equal to zero if all m linear Eqs. (45) are fulfilled in the vertex

$\mathbf{w}_L$ which is related by the Eq. (47) to the matrix $\boldsymbol{B}_L = [\mathbf{x}_1,\ldots, \mathbf{x}_m, \mathbf{e}_{i(1)},\ldots, \mathbf{e}_{i(n-m)}]^T$ (48) containing the unit vectors $\mathbf{e}_i$ with the indices i from the subset I_L ($i \in I_L$).

Theorem 4: If the feature vectors $\mathbf{x}_j$ constituting the subset $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subset \boldsymbol{C}$ (1)) and used in the definition of the function $\Phi(\mathbf{w})$ (38) are linearly independent (*Def.* 2), then the value $\Phi(\mathbf{w}_L)$ of the collinearity criterion function $\Phi(\mathbf{w})$ at the final vertex $\mathbf{w}_L$ (59) is equal to zero ($\Phi(\mathbf{w}_L) = 0$).

The proof of Theorem 4 can be based on the stepwise inversion of the matrices $\boldsymbol{B}_k$ (48) [16]. The final vertex $\mathbf{w}_L$ (59) can be found by inverting the related matrix $\boldsymbol{B}_L = [\mathbf{x}_1,\ldots, \mathbf{x}_{rk}, \mathbf{e}_{i(1)},\ldots, \mathbf{e}_{i(n-rk)}]^T$ (48).

The final vertex $\mathbf{w}_L$ (59) resetting ($\Phi(\mathbf{w}_L) = 0$) the criterion function $\Phi(\mathbf{w})$ (38) can be related to the optimal matrix $\boldsymbol{B}_L = [\mathbf{x}_1,\ldots, \mathbf{x}_L, \mathrm{e}_{i(L+1)},\ldots, \mathrm{e}_{i(n)}]^T$ (48) built from L ($L \le m$) feature vectors $\mathbf{x}_j$ ($j \in J(\mathbf{w}_L)$ (45)) from the data set $\boldsymbol{C}$ (1) and from $n - L$ selected unit vectors $\mathbf{e}_i$ ($i \in I(\mathbf{w}_L)$ (46)). Different subsets of the unit vectors $\mathbf{e}_i$ in the final matrix $\boldsymbol{B}_L$ (48) result in different positions of the final vertices $\mathbf{w}_{L(l)}$ (59) in the parameter space $\boldsymbol{R}^n$. The criterion function $\Phi(\mathbf{w})$ (38) is equal zero ($\Phi(\mathbf{w}_{L(l)}) = 0$) at each of these vertices $\mathbf{w}_{L(l)}$ (59).

The position of the final vertices $\mathbf{w}_{L(l)}$ (59) in the parameter space $\boldsymbol{R}^n$ depends on which unit vectors $\mathbf{e}_i$ ($i \in I_{L(l)}$) are included in the basis $\mathbf{B}_{L(l)}$ (48), where:

$$(\forall l \in \{1, \ldots, l_{\max}) \; \Phi_k(\mathbf{w}_{L(l)}) = 0 \qquad (69)$$

The maximal number $l_{\max}$ (69) of different vertices $\mathbf{w}_{L(l)}$ (59) can be large when $m << n$:

$$l_{\max} = n!/m!(n\text{–}m)! \qquad (70)$$

The choice between different final vertices $\mathbf{w}_{L(l)}$ (59) can be based on the minimization of the regularized criterion function $\Psi(\mathbf{w})$ (42). The regularized function $\Psi(\mathbf{w})$ (42) is the sum of the collinearity function $\Phi(\mathbf{w})$ (38) and the weighted sum of the cost functions $\varphi_i^0(\mathbf{w})$ (43). If $\Phi(\mathbf{w}_{L(l)}) = 0$ (38), then the value $\Psi(\mathbf{w}_{L(l)})$ of the criterion function $\Psi(\mathbf{w})$ (42) at the final vertex $\mathbf{w}_{L(l)}$ (59) can be given as follows:

$$\begin{aligned} \Psi(\mathbf{w}_{L(l)}) &= \lambda_i \, \Sigma_i \, \chi_i \, \varphi_j{}^0(\mathbf{w}_{L(l)}) = \\ &= \lambda \, \Sigma \, \chi_i \mid \mathrm{w}_{L(l),i} \mid \end{aligned} \qquad (71)$$

where the above sums take into account only the indices i of the subset $I(\mathbf{w}_{L(l)})$ of the non-zero components $\mathrm{w}_{L(l),i}$ of the final vertex $\mathbf{w}_{L(l)} = [\mathrm{w}_{L(l),1}, \ldots, \mathrm{w}_{L(l),n}]^T$ (59):

$$I(\mathbf{w}_{L(l)}) = \{i : \mathbf{e}_i{}^T\mathbf{w}_{L(l)} \neq 0\} = \{i : \mathrm{w}_{L(l),i} \neq 0\} \qquad (72)$$

If the final vertex $\mathbf{w}_{L(l)}$ (59) is not degenerate (*Def.* 5), then the matrix $\boldsymbol{B}_{L(l)}$ (48) is built from all m feature vectors $\mathbf{x}_j$ ($j \in \{1,\ldots., m\}$) making up the data set $\boldsymbol{C}$ (1) and from $n - m$ selected unit vectors $\mathbf{e}_i$ ($i \in I(\mathbf{w}_{L(l)})$ (71)).

$$\boldsymbol{B}_{)m} = [\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{e}_{i(m+1)}, \ldots, \mathbf{e}_{i(n)}]^T \qquad (73)$$

The problem of the constrained minimizing of the regularized function $\Psi(\mathbf{w})$ (71) at the vertices $\mathbf{w}_{L(l)}$ (59) satisfying the conditions $\Phi(\mathbf{w}_{L(l)}) = 0$ (69) can be formulated in the following way:

$$\begin{aligned} &min_l\{\Psi(\mathbf{w}_{L(l)}) \colon \Phi(\mathbf{w}_{L(l)}) = 0\} = \\ &= min_l\{\Sigma_i \, \gamma_i \mid \mathrm{w}_{L(l),i} \mid \colon \Phi(\mathbf{w}_{L(l)}) = 0\} \end{aligned} \qquad (74)$$

According to the above formulation, the search for the minimum of the regularized criterion function $\Psi(\mathbf{w})$ (42) is takes place at all such vertices $\mathbf{w}_{L(l)}$ (59), where the collinearity function $\Phi(\mathbf{w})$ (38) is equal to zero. The regularized criterion function $\Psi(\mathbf{w})$ (42) is defined as follows at the final vertices $\mathbf{w}_{L(l)} = [w_{L(l),1}, \ldots, w_{L(l),n}]^T$ (59), where $\Phi(\mathbf{w}_{L(l)}) = 0$:

$$(\forall \mathbf{w}_{L(l)})\ \Psi'(\mathbf{w}_{L(l)}) = \Sigma\, \gamma_i \mid w_{L(l),i} \mid \tag{75}$$

The optimal vertex $\mathbf{w}_{L(l)}{}^*$ is the minimum value $\Psi'(\mathbf{w}_{L(l)}{}^*)$ of the *CPL* criterion function $\Psi'(\mathbf{w})$ (75) defined on such final vertices $\mathbf{w}_{L(l)}$ (59), where $\Phi(\mathbf{w}_{L(l)}) = 0$ (38):

$$(\exists \mathbf{w}_{L(l)}{}^*)\ (\forall \mathbf{w}_{L(l)} : \Phi(\mathbf{w}_{L(l)}) = 0)\ \Psi'(\mathbf{w}_{L(l)}) \geq \Psi'(\mathbf{w}_{L(l)}{}^*) > 0 \tag{76}$$

As in the case of the minimization of the perceptron criterion function $\Phi_k{}^P(\mathbf{v})$ (29), the optimal vector $\mathbf{w}_{L(l)}{}^*$ (76) may be located at a selected vertex of some convex polyhedron (27) in the parameter space $\boldsymbol{R}^n$ ($\mathbf{w} \in \boldsymbol{R}^n$) [7].

If the cost parameters γ_i (42) have standard values of one ($(\forall i \in \{1, \ldots, n\})\ \gamma_i = 1$), then the constraint minimization problem (74) leads to the optimal vertex $\mathbf{w}_{L(l)}{}^*$ with the smallest L_1 length $\|\ \mathbf{w}_{L(l)}{}^*\ \|_{L1} = |w_{L(l),1}{}^*| + \ \ldots\ + |w_{L(l),n}{}^*|$, where $\Phi(\mathbf{w}_{L(l)}{}^*) = 0$ (38):

$$(\exists \mathbf{w}_{L(l)}{}^*)\ (\forall \mathbf{w}_{L(l)} : \Phi(\mathbf{w}_{L(l)}) = 0)\ \|\mathbf{w}_{L(l)}\| \geq \|\mathbf{w}_{L(l)}{}^*\| \tag{77}$$

Optimal vertex $\mathbf{w}_{L(l)}{}^*$ with the smallest L_1 length $\|\ \mathbf{w}_{L(l)}{}^*\ \|_{L1}$ (77) is related to the largest L_1 margin $\delta_{L1}(\mathbf{w}_{L(l)}{}^*)$ (6) [11]:

$$\delta_{L1}(\mathbf{w}_{L(l)}{}^*) = 2/\|\ \mathbf{w}_{L(l)}{}^*\ \|_{L1} = 2/(|w_{L(l),1}{}^*| + \ldots + |\ w_{L(l),n}{}^*|) \tag{78}$$

The basis exchane algorithm allow to solve the constraint minimization problem (74) and to find the optimal vertex $\mathbf{w}_{L(l)}{}^*$ (77) with the largest L_1 margin $\delta_{L1}(\mathbf{w}_{L(l)}{}^*)$.

Support Vector Machines (*SVM*) is the most popular method for designing linear classifiers or prognostic models with large margins [12]. According to the *SVM* approach, the optimal linear classifier or the prognostic model defined by such an optimal weight vector $\mathbf{w}^*$ that has a maximum margin $\delta_{L2}(\mathbf{w}^*)$ based on the Euclidean (L_2) norm:

$$\delta_{L2}(\mathbf{w}^*) = 2/\|\ \mathbf{w}^*\|_{L2} = 2/\left((\mathbf{w}^*)^T\mathbf{w}^*\right)^{1/2} \tag{79}$$

Maximization of the Euclidean margins $\delta_{L2}(\mathbf{w})$ (79) is performed using quadratic programming [2].

8. Complex layers of linear prognostic models

Complex layers of linear classifiers or prognostic models have been proposed as a scheme for obtaining a general classification or forecasting rules designed on the basis of a small number of multidimensional feature vectors $\mathbf{x}_j$ [11]. According to this scheme, when designing linear prognostic models, averaging over a small number m of feature vectors $\mathbf{x}_j$ of the dimension n ($m << n$) is replaced by averaging on collinear clusters of selected features (genes) X_i. Such an approach to averaging can be linked to the ergodic theory [17].

In the case of a small sample of multivariate vectors, the number m of feature vectors $\mathbf{x}_j$ in the data set $\boldsymbol{C}$ (1) may be much smaller than the dimension n of these vectors ($m << n$). In this case, the collinear cluster $\boldsymbol{C}(\mathbf{w}_k{}^*[r_k])$ (52) may contain all feature vectors $\mathbf{x}_j$ from the set $\boldsymbol{C}$ (1) and the vertex $\mathbf{w}_k{}^*[r_k]$ may have the rank r_k equal to m ($r_k = m$).

As it follows from Theorem 4, if the collinearity criterion function $\Phi(\mathbf{w})$ (38) is defined on linearly independent (*Def.* 2) feature vectors $\mathbf{x}_j$, then the values $\Phi(\mathbf{w}_{m(l)})$ of this function at each final vertex $\mathbf{w}_{m(l)}$ (59) are equal to zero ($\Phi(\mathbf{w}_{m(l)}) = 0$). Each final vertex $\mathbf{w}_{m(l)}$ (59) can be reached in m steps k ($k = 1, \ldots, m$) starting from the vertex $\mathbf{w}_0 = [0,\ldots, 0]^T$ related to the identity matrix $\boldsymbol{B}_0 = \boldsymbol{I}_n = [\mathbf{e}_1,\ldots, \mathbf{e}_n]^T$.

Minimization of the collinearity criterion function $\Phi(\mathbf{w})$ (38), and then minimization of the criterion function $\Psi'(\mathbf{w}_{L(l)})$ (75) at the final vertices $\mathbf{w}_{L(l)}$ (59) allows to determine the optimal vertex $\mathbf{w}_{L(l)}{}^*$ (77) with the largest L_1 margin $\delta_{L1}(\mathbf{w}_{L(l)}{}^*)$ (78). If the feature vectors $\mathbf{x}_j$ (1) are linearly independent, then the optimal vertex $\mathbf{w}_{L(l)}{}^*$ (77) is related to the optimal basis $\boldsymbol{B}_{L(l)}{}^* = [\mathbf{x}_1,\ldots, \mathbf{x}_m, \mathbf{e}_{i(m+1)},\ldots, \mathbf{e}_{i(n)}]^T$ which contains all m feature vectors $\mathbf{x}_j$ (1) and $n - m$ unit vectors $\mathbf{e}_i$ with the indices i belonging to the optimal subset $I(\mathbf{w}_{L(l)}{}^*)$ (71) ($i \in I(\mathbf{w}_{L(l)}{}^*)$).

The optimal basis $\boldsymbol{B}_m{}^* = [\mathbf{x}_1,\ldots, \mathbf{x}_m, \mathbf{e}_{i(m+1)},\ldots, \mathbf{e}_{i(n)}]^T$ (73) is found in two stages. In the first stage, m feature vectors $\mathbf{x}_j$ (1) are introduced into matrices $\boldsymbol{B}_k = [\mathbf{x}_1,\ldots, \mathbf{x}_k, \mathbf{e}_{i(k+1)},\ldots, \mathbf{e}_{i(n)}]^T$ ($k = 0, 1, \ldots, m - 1$). The inverse matrices $\boldsymbol{B}_k{}^{-1}$ (49) are computed in accordance with the vector Gauss-Jordan transformation (64). In the second stage, the unit vectors $\mathbf{e}_{i(l)}$ in the matrices $\boldsymbol{B}_{m(l)}$ (73) are exchanged to minimize the *CPL* function $\Psi'(\mathbf{w}_{m(l)})$ (75) at the final vertices $\mathbf{w}_{m(l)}$ (77). The optimal basis $\boldsymbol{B}_m{}^*$ defines (47) the optimal vertex $\mathbf{w}_{m(l)}{}^*$ (77), which is characterized by the largest margin $\delta_{L1}(\mathbf{w}_{m(l)}{}^*)$ (78).

The vertexical feature subspace $\boldsymbol{F}_1{}^*[m]$ ($\boldsymbol{F}_1{}^*[m] \subset \boldsymbol{F}[n]$ (1)) can be obtained on the basis of the optimal vertex $\mathbf{w}_{m(l)}{}^*$ (77) with the largest margin $\delta_{L1}(\mathbf{w}_{m(l)}{}^*)$ (78). The vertexical subspace $\boldsymbol{F}_1{}^*[m]$ contains the reduced vectors $\mathbf{x}_{1,j}[m]$ with the dimension m [7]:

$$(\forall j \in \{1, \ldots, m\})\ \mathbf{x}_{1,j}[m] \in \ \boldsymbol{F}_1{}^*[m] \tag{80}$$

The reduced vectors $\mathbf{x}_{1,j}[m]$ (80) are obtained from the feature vectors $\mathbf{x}_j = [x_{j,1},\ldots,x_{j,n}]^T$ ($\mathbf{x}_j \in \boldsymbol{F}[n]$) ignoring such components $x_{j,i}$ which are related to the unit vectors $\mathbf{e}_i$ in the optimal basis $\boldsymbol{B}_1{}^* = [\mathbf{x}_1,\ldots, \mathbf{x}_m, \mathbf{e}_{i(m+1)},\ldots, \mathbf{e}_{i(n)}]^T$ (73). The reduced vectors $\mathbf{x}_{1,j}[m]$ are represented by such m features X_i ($X_i \in R_1{}^*$ (54)), which are not linked to the unit vectors $\mathbf{e}_i$ ($i \notin I_{m(l)}{}^*$) in the basis $\boldsymbol{B}_{m(l)}{}^*$ (73) representing the optimal vertex $\mathbf{w}_{m(l)}{}^*$ (77).

$$R_1{}^* = \{X_{i(1)}, \ldots, X_{i(m)} : i(l) \notin I_{m(l)}{}^*\ (72)\} \tag{81}$$

The m features $X_{i(l)}$ belonging to the optimal subset $R_1{}^*$ ($X_{i(l)} \in R_1{}^*$ (81) are related to the weights $w_{k.l}{}^*$ ($\mathbf{w}_k{}^*[m] = [w_{k,1}{}^*, \ldots, w_{k,m}{}^*]^T$) that are not zero ($w_{k.l}{}^* \neq 0$).

The optimal feature subset $R_1{}^*$ (81) consists of m collinear features X_i. The optimal vertex $\mathbf{w}_1{}^*[m]$ ($\Phi(\mathbf{w}_1{}^*[m]) = 0$ (69)) in the reduced parameter space $\boldsymbol{R}^m$ ($\mathbf{w}_1{}^*[m] \in \boldsymbol{R}^m$) is based on these m features X_i. The reduced optimal vertex $\mathbf{w}_1{}^*[m]$ with the largest margin $\delta_{L1}(\mathbf{w}_1{}^*[m])$ (77) is the unique solution of the constrained optimization problem (74). Maximizing the L_1 margin $\delta_{L1}(\mathbf{w}_l{}^*)$ (78) leads to the first reduced vertex $\mathbf{w}_1{}^*[m] = [w_{k,1}{}^*, \ldots, w_{k,m}{}^*]^T$ with non-zero components $w_{k.i}{}^*$ ($w_{k.i}{}^* \neq 0$).

The collinear interaction model between m collinear features $X_{i(l)}$ from the optimal subset $R_1{}^*(m)$ (81) can be formulated as follows (57):

$$w_{k.1}{}^*\ X_{i(1)} + \ldots + w_{k.m}{}^*\ X_{i(m)} = 1 \tag{82}$$

The prognostic models for each feature $X_{i'}$ from the subset $R_1{}^*$ (81) may have the following form (58):

$$(\forall i' \in \{1, \ldots, m) \quad X_{i'} = \alpha_{i',0} + \alpha_{i',1}\, X_{i(1)} + \ldots + \alpha_{i',m}\, X_{i(m)} \tag{83}$$

where $\alpha_{i',0} = 1 / w_{k.i'}{}^*$, $\alpha_{i',\, i'} = 0$, and $(\forall\, i(l) \neq i')\ \alpha_{i',i(l)} = w_{k.i(l)}{}^* / w_{k.i'}{}^*$.

In the case of a data set $\boldsymbol{C}$ with a small number m $(m << n)$ of multidimensional feature vectors $\mathbf{x}_j$ (1), the prognostic models (83) for individual features $X_{i'}$ can be weak. It is know that sets (ensembles) of weak models can have strong generalizing properties [4]. A set of weak prognostic models (83) for a selected feature (dependent variable) $X_{i'}$ can be implemented in the complex layer of L prognostic models (83) [11].

The complex layer can be built on the basis of the sequence of L optimal vertices $\mathbf{w}_l{}^*$ (77) related to m features X_i constituting the subsets $R_l{}^*$ (81), where l = 0, 1,..., L.

$$(\mathbf{w}_1{}^*, R_1{}^*), \ldots, (\mathbf{w}_L{}^*, R_L{}^*) \tag{84}$$

Design assumption: Each subset $R_l{}^*$ (81) in the sequence (84) contains a priori selected feature (dependent variable) $X_{i'}$ and m - 1 other features (independent variables) $X_{i(l)}$. The other features $X_{i(l)}$ ($X_{i(l)} \in R_l{}^*$) should be different in successive subsets $R_l{}^*$ (l = 0, 1,..., L).

The first optimal; vertex $\mathbf{w}_1{}^*$ (77) in the sequence (84) is designed on the basis of m feature vectors $\mathbf{x}_j$ (1), which are represented by all n features X_i constituting the feature set $F(n) = \{X_1, \ldots, X_n\}$. The vertex $\mathbf{w}_1{}^*$ (77) is found by solving the constraint optimization problem (74) according to the procedure with the two stages outlined earlier. The two-stage procedure allows to find the optimal vertex $\mathbf{w}_1{}^*$ (77) with the largest L_1 margin $\delta_{L1}(\mathbf{w}_1{}^*)$ (78).

The second optimal vertex $\mathbf{w}_2{}^*$ (77) in the sequence (84) is obtained on the basis of m reduced feature vectors $\mathbf{x}_j[n - (m - 1)]$ (67), which are represented by n - (m - 1) features X_i constituting the reduced feature subset $F_2(n - (m + 1))$:

$$F_2(n - (m - 1)) = F(n)/R_1{}^* \cup \{X_{i'}\} \tag{85}$$

The l-th optimal vertex $\mathbf{w}_l{}^*$ (77) in the sequence (84) is designed on the basis of m reduced vectors $\mathbf{x}_j[n - l(m - 1)]$ (67), which are represented by n - l(m - 1) features X_i constituting the feature subset $F_l(n - l(m - 1))$:

$$F_l(n - l(m - 1)) = F_{l-1}(n - l(m - 1))/R_{l-1}{}^* \cup \{ X_{i'} \} \tag{86}$$

The sequence (84) of L optimal vertices $\mathbf{w}_l{}^*$ (77) related to the subsets $F_l(n - l(m - 1))$ (86) of features is characterized by decreased L_1 margins $\delta_{L1}(\mathbf{w}_l{}^*)$ (78) [18].

$$\delta_{L1}(\mathbf{w}_1{}^*) \geq \delta_{L1}(\mathbf{w}_2{}^*) \geq \ldots \geq \delta_{L1}(\mathbf{w}_L{}^*) \tag{87}$$

The prognostic models (83) for the dependent feature (variable) $X_{i'}$ are designed for each subset $F_l(n - l(m - 1))$ (86) of features X_i, where l = 0, 1,..., L (84):

$$(\forall l \in \{0, 1, \ldots, L\} \tag{88}$$

$$X_{i'}(l) = \alpha_{i',0}(l) + \alpha_{i',1}(l)\, X_{i(1)}(l) + \ldots + \alpha_{i',m}(l)\, X_{i(m)}$$

The final forecast $X_{i'}{}^\wedge$ for the dependent feature (variable) $X_{i'}$ based on the complex layer of L + 1 prognostic models (88) can have the following form:

$$X_{i'}^{\wedge} = (X_{i'}(1) + \ldots + X_{i'}(L))/(L + 1) \qquad (89)$$

In accordance with the Eq. (89), the final forecast $X_{i(m)}^{\wedge}$ for the feature $X_{i'}$ results from averaging the forecasts of $L + 1$ individual models $X_{i'}(l)$ (88).

9. Concluding remarks

The article considers computational schemes of designing classifiers or prognostic models based on such a data set $\boldsymbol{C}$ (1), which consists of a small number m of high-dimensional feature vectors $\mathbf{x}_j$ ($m < < n$).

The concept of a complex layer composed of many linear prognostic models (88) built in low-dimensional feature subspaces is discussed in more detail. These models (88) are built by using a small number m of collinear features X_i belonging to the optimal feature clusters R_l^* (81). The optimal feature clusters R_l^* (81) are formed by the search for the largest margins $\delta_{L1}(\mathbf{w}_l^*)$ (78) in the L_1 norm.

The averaged prognostic models $X_{i'}^{\wedge}$ (89) are based on the layer of L parallel models $X_{i'}(l)$ (88). In line with the ergodic theory, averaging on a small number m of feature vectors $\mathbf{x}_j$ has been replaced with averaging on L collinear clusters R_l^* (81) of features X_i. Such averaging scheme should allow for a more stable extraction of general patterns from small samples of high-dimensional feature vectors $\mathbf{x}_j$ (1) [11].

Author details

Leon Bobrowski[1,2]

1 Faculty of Computer Science, Białystok University of Technology, Poland

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

*Address all correspondence to: l.bobrowski@pb.edu.pl

References

[1] Duda O. R., Hart P. E., and Stork D. G., *Pattern classification*, J. Wiley, New York, 2001

[2] Hand D., Smyth P., and Mannila H., *Principles of data mining*, MIT Press, Cambridge (2001)

[3] Bishop C. M., *Pattern Recognition and Machine Learning*. Springer Verlag, 2006

[4] Kuncheva L.: *Combining Pattern Classifiers: Methods and Algorithms*, 2nd Edition, J. Wiley, New Jersey (2014).

[5] Simonnard M., *Linear Programming*, Prentice – Hall, New York, Englewood Cliffs, 1966

[6] Bobrowski L., *Data mining based on convex and piecewise linear* (CPL) *criterion functions* (*in Polish*), Białystok University of Technology, 2005

[7] Bobrowski L., *Data Exploration and Linear Separability*, pp. 1 - 172, Lambert Academic Publishing, 2019

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

[9] Bobrowski L., Zabielski P., ″Models of Multiple Interactions from Collinear Patterns″, pp. 153-165 in: *Bioinformatics and Biomedical Engineering* (*IWBBIO* 2018), Eds.: I. Rojas, F. Guzman, LNCS 10208, Springer Verlag, 2018

[10] Bobrowski L., Small Samples of Multidimensional Feature Vectors (*ICCCI 2020*), pp. 87 - 98 in: *Advances in Computational Collective Intelligence*, Eds.: Hernes M, et al., Springer 2020

[11] Bobrowski L., ″Complexes of Low Dimensional Linear Classifiers with L_1 Margins″, pp. *29 - 40 in:* ACIIDS 2021, Springer Verlag, 2021

[12] Boser B. E., Guyon I., Vapnik V. N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth *Annual Workshop of Computational Learning Theory*, 5, 144–152. Pittsburgh, ACM, 1992

[13] Bobrowski L., Łukaszuk T.: Repeatable functionalities in complex layers of formal neurons, *EANN* 2021, *Engineering Applications of Neural Networks*, Springer 2021

[14] Rosenblatt F.: *Principles of neurodynamics*, Spartan Books, Washington, 1962

[15] Bobrowski L., Łukaszuk, T.: Relaxed Linear Separability (*RLS*) Approach to Feature (Gene) Subset Selection, pp. 103 - 118 in: *Selected Works in Bioinformatics*, Edited by: Xuhua Xia, *INTECH*, 2011

[16] Bobrowski L.: ″Large Matrices Inversion Using the Basis Exchange Algorithm″, *British Journal of Mathematics & Computer Science*, 21(1): 1-11, 2017

[17] Petersen K.: *Ergodic Theory* (*Cambridge Studies in Advanced Mathematics*), Cambridge University Press, 1990

[18] Bobrowski L., Zabielski P.: ″Feature (gene) clustering with collinearity models″, ***ICCCI* 2021** (*to appear),* Springer Verlag, 2021