FEATURE SELECTION BASED ON LINEAR SEPARABILITY AND A CPL CRITERION FUNCTION

LEON BOBROWSKI

Faculty of Computer Science, Bialystok University of Technology, Wiejska 45A, 15-351 Bialystok, Poland
and Institute of Biocybernetics and Biomedical Engineering,PAS, Księcia Trojdena 4, 02-109 Warsaw, Poland
leon@ibib.waw.pl

(Received 12 September 2003; revised manuscript received 30 January 2004)

Abstract: Linear separability of data sets is one of the basic concepts in the theory of neural networks and pattern recognition. Data sets are often linearly separable because of their high dimensionality. Such is the case of genomic data, in which a small number of cases is represented in a space with extremely high dimensionality.

An evaluation of linear separability of two data sets can be combined with feature selection and carried out through minimisation of a convex and piecewise-linear (CPL) criterion function. The perceptron criterion function belongs to the CPL family. The basis exchange algorithms allow us to find minimal values of CPL functions efficiently, even in the case of large, multidimensional data sets.

Keywords: linear separability, feature selection, CPL criterion function

1. Introduction

Let us consider a situation when objects in a database are represented as a set of feature vectors of the same dimensionality. Components of these vectors represent particular features, which are numerical results of a given object examination. The feature vectors divided into different categories (classes) constitute the so-called learning sets. A given learning set contains feature vectors related to the same category of objects.

An important practical problem is the extraction of decision rules from learning sets. Estimated rules can be used in the classification process or in the decision support systems [1]. For example, diagnosis support systems may be based on differentiation rules between diseases. Differentiation rules can be extracted from any medical database containing data about patients diagnosed with particular diseases by physicians. Such a principle has been implemented in the *Hepar* computer system which comprises a hepathological database and shell of procedures aimed at multivariate data visualisation, analysis and diagnosis support ([2, 3]). Category models can be designed by means of combining experts' knowledge with data contained in the learning sets. Such an approach has been implemented in modelling liver diseases by means of Bayesian networks in the *Hepar II* system [4].

Not all diagnostic examinations are used in the decision rules. In other words, some features are irrelevant in the classification process and therefore can be neglected. Neglecting unimportant features allows us to improve the quality of classification rules. Extracting sets of the most important features is known as feature selection [1]. A proper feature selection should result in more correct and more general classification rules.

One of the most direct approaches to the feature selection problem is the evaluation of quality of classification rules based on a given feature subset. The most common method of evaluating a classifier is estimating the classification error rate [1]. Estimation of a classification error related to a given feature subset is not efficient from the computational point of view and is difficult to apply to large data sets. Mainly for this reason, a variety of many other methods of feature selection has been proposed [5]. Among them, the Support Vector Machine (SVM) has recently been used for this purpose ([6, 7]).

In this paper we shall discuss the possibility of using the concept of linear separability of learning data sets in the feature selection task [8]. A high dimensionality of data (*long* feature vectors) often results in their linear separability. The convex and piecewise-linear (CPL) functions defined on the learning sets are used for measuring the linear separability of these sets. The perceptron criterion function belongs to the CPL family. Different measures of the linear separability of learning sets can be based on the minimal value of adequately adjusted CPL functions.

2. Linear separabilty of learning sets

Let us assume that m objects O_j contained in a database are represented as feature vectors $\mathbf{x}_j[n]=[x_{j1},\ldots,x_{jn}]^T$ or points in the n-dimensional feature space $F[n]$. The x_i components of vectors $\mathbf{x}_j[n]$ are called features. Features x_{ji} are numerical results of examination of the j^{th} object O_j. We are considering a situation when feature vectors $\mathbf{x}_j[n]$ can be a mixed, qualitative-quantitative type. Components x_{ji} of such vectors $\mathbf{x}_j[n]$ can be binary ($x_i \in \{0,1\}$) or real numbers ($x_i \in R^1$).

Objects O_j are often divided into categories (*classes*), ω_k ($k=1,\ldots,K$). For example, a medical database may contain patients $O_j(k)$ linked by physicians to particular diseases, ω_k, and represented as labelled feature vectors $\mathbf{x}_j(k)$. In such cases, features x_{ji} are numerical results of diagnostic examinations of a given patient, O_j. Learning set C_k contains m_k feature vectors $\mathbf{x}_j(k)$ belonging to the same class, ω_k:

$$C_k=\{\mathbf{x}_j(k)\} \qquad (j\in I_k), \tag{1}$$

where I_k is the set of indexes of the feature vectors $\mathbf{x}_j(k)$ belonging to the ω_k class.

We will consider separation of learning sets C_k by hyperplanes $H(\mathbf{w}_k,\theta_k)$ ($k\in K$) in a feature space:

$$H(\mathbf{w}_k,\theta_k)=\{\mathbf{x}:\langle\mathbf{w}_k,\mathbf{x}\rangle=\theta_k\}, \tag{2}$$

where $\mathbf{w}_k\in R^n$ is the weight vector, $\theta_k\in R^1$ is the threshold, and $\langle\mathbf{w}_k,\mathbf{x}\rangle$ is the inner product.

Feature vector $\mathbf{x}$ is situated on the *positive* (*negative*) *side* of hyperplane $H(\mathbf{w}_l,\theta_l)$ if and only if $\langle \mathbf{w}_k,\mathbf{x}_j\rangle > \theta_l$ ($\langle \mathbf{w}_k,\mathbf{x}_j\rangle < \theta_l$).

DEFINITION 1: Learning sets (1) are *linearly separable* if each of the C_k sets can be fully separated from the sum of the remaining C_i sets by some $H(\mathbf{w}_k,\theta_k)$ hyperplane (3):

$$\begin{aligned}(\forall k \in \{1,\ldots,K\})\ (\exists \mathbf{w}_k,\theta_k)\ (\forall \mathbf{x}_j \in C_k) \qquad & \langle \mathbf{w}_k,\mathbf{x}_j\rangle > \theta_k,\\ \text{and } (\forall \mathbf{x}_j \in C_i, i \neq k) \qquad & \langle \mathbf{w}_k,\mathbf{x}_j\rangle < \theta_k.\end{aligned} \tag{3}$$

In accordance with relation (3), the entire C_k learning set is situated on the positive side of the $H(\mathbf{w}_k,\theta_k)$ hyperplane (2) and all the $\mathbf{x}_j(i)$ feature vectors belonging to the sum of the remaining C_i sets are situated on the negative side of this hyperplane.

Let the symbol $F_l[n']$ stand for the n'-dimensional subspace of the n-dimensional feature space $F[n]$ ($F_l[n'] \subset F[n], n' \leq n$). The $F_l[n']$ subspace is constituted of n'-dimensional vectors $\mathbf{x}' = [x_{i(1)},\ldots,x_{i(n')}]^T$ with the $i(j)$ indices of n' features $x_{i(j)}$ belonging to the $I_l[n']$ set:

$$F_l[n'] = \{\mathbf{x}' = [x_{i(1)},\ldots,x_{i(n')}]^T : i(j) \in I_l[n']\}. \tag{4}$$

The $\mathbf{x}' = [x_{i(1)},\ldots,x_{i(n')}]^T$ vectors constitute of the n-dimensional feature $\mathbf{x} = [x_1,\ldots,x_n]^T \in F[n]$ vectors as a result of neglecting the features, x_i, with the indices i outside the set $I_l[n'] (i \notin I_l[n'])$.

The separability property (3) depends on the feature space, $F_l[n']$. The C_k (1) learning sets can be linearly separable in one feature space, $F_l[n']$, and not separable in another space, $F_k[n']$.

REMARK 1 (monotonocity property): If C_k (1) learning sets are linearly separable in one feature space, F_l, then they are also linearly separable in a greater feature space, F_k ($F_l \subset F_k$).

In accordance with the above remarks, enlargement of the feature space cannot eliminate the linear separability of the learning sets. In order to prove that linear separability is preserved, it is enough to mention that any enlargement of feature space F_l by some x_i components can be linked to the enlargement of the weight vector, $\mathbf{w}_k$ (3), by some w_{ki} components equal to zero. As a result, relation (3) is fulfilled in a greater feature space, F_k. We can also mention that linear separability of learning sets C_k (1) can always be achieved by means of a sufficient enlargement of the feature space, $F_l[1]$. This subject is discussed in greater detail in the following section.

3. Positive and negative sets

Let us take into consideration two disjoined sets: *positive* (G^+) and *negative* (G^-) composed of m^+ and m^- feature vectors $\mathbf{x}_j(k)$ (1), so that:

$$G^+ \cap G^- = \varnothing. \tag{5}$$

It is convenient to assume that entire learning sets C_k (1) have been allocated to the positive (G^+) or the negative (G^-) set. In other words, the learning sets, C_k, are not divided during this allocation:

$$\begin{aligned}(\forall k \in \{1,\ldots,K\})\ (\forall j)\ & \mathbf{x}_j(k) \in G^+ \Rightarrow C_k \subset G^+,\\ & \mathbf{x}_j(k) \in G^- \Rightarrow C_k \subset G^-.\end{aligned} \tag{6}$$

We are interested in finding such a $H(\mathbf{w}_1,\theta_1)$ hyperplane (3) that would separate the G^+ and G^- sets. It may mean that the largest possible number of points $\mathbf{x}_j$ from the first set, G^+, should be situated on the positive side of the $H(\mathbf{w}_1,\theta_1)$*hyperplane*$(\langle\mathbf{w}_1,\mathbf{x}_j\rangle > \theta_1)$ and at the same time the largest possible number of points $\mathbf{x}_j$ from the second set, G^-, should be situated on the negative side $(\langle\mathbf{w}_1,\mathbf{x}_j\rangle < \theta_1)$. The G^+ and G^- sets are *linearly separable* if there exist such parameters $\mathbf{w}_1$ and θ_1 that all points $\mathbf{x}_j$ from these sets are properly allocated:

$$\begin{aligned} (\exists \mathbf{w}_1,\theta_1)\ (\forall \mathbf{x}_j \in G^+) \qquad & \langle\mathbf{w}_1,\mathbf{x}_j\rangle > \theta_1, \\ \text{and } (\forall \mathbf{x}_j \in G^-) \qquad & \langle\mathbf{w}_1,\mathbf{x}_j\rangle < \theta_1. \end{aligned} \tag{7}$$

We are searching for such a $H(\mathbf{w}_1,\theta_1)$ hyperplane that would separate these sets.

It is convenient to use *augmented* feature vectors, $\mathbf{x}_j = [1,(\mathbf{x}_j)^T]^T$, in dealing with linear separability:

$$\begin{aligned} (\exists \mathbf{v}_1)\ (\forall \mathbf{y}_j \in G^+) \qquad & \langle\mathbf{v}_1,\mathbf{y}_j\rangle > 0, \\ \text{and } (\forall \mathbf{y}_j \in G^-) \qquad & \langle\mathbf{v}_1,\mathbf{y}_j\rangle < 0, \end{aligned} \tag{8}$$

where $\mathbf{v} = [-\theta,\mathbf{w}^T]^T$ is the *augmented* weight vector [1].

Inequalities (8) can be represented as:

$$\begin{aligned} (\exists \mathbf{v}_1)\ (\forall \mathbf{y}_j \in G^+) \qquad & \langle\mathbf{v}_1,\mathbf{y}_j\rangle \geq \varepsilon, \\ \text{and } (\forall \mathbf{y}_j \in G^-) \qquad & \langle\mathbf{v}_1,\mathbf{y}_j\rangle \leq -\varepsilon, \end{aligned} \tag{9}$$

where $\varepsilon > 0$.

The ε parameter could be chosen as:

$$\varepsilon = \min_j \varepsilon_j, \tag{10}$$

where $(\forall \mathbf{y}_j \in G^+)\ \varepsilon_j = \langle\mathbf{v}_1,\mathbf{y}_j\rangle$ and $(\forall \mathbf{y}_j \in G^-)\ \varepsilon_j = -\langle\mathbf{v}_1,\mathbf{y}_j\rangle$.

REMARK 2 (linear separability): The G^+ and G^- sets are *linearly separable* (8) **if and only if** the following inequalities are fulfilled:

$$\begin{aligned} (\exists \mathbf{v}_2)\ (\forall \mathbf{y}_j \in G^+) \qquad & \langle\mathbf{v}_2,\mathbf{y}_j\rangle \geq 1, \\ \text{and } (\forall \mathbf{y}_j \in G^-) \qquad & \langle\mathbf{v}_2,\mathbf{y}_j\rangle \leq -1. \end{aligned} \tag{11}$$

To prove equivalence between Equation (9) and Equation (11) we can take:

$$\mathbf{v}_2 = \mathbf{v}_1/\varepsilon. \tag{12}$$

REMARK 3 (sufficient condition for linear separability): The G^+ and G^- sets are *linearly separable* (8) **if** the following equalities are fulfilled:

$$\begin{aligned} (\exists \mathbf{v}_2)\ (\forall \mathbf{y}_j \in G^+) \qquad & \langle\mathbf{v}_2,\mathbf{y}_j\rangle = 1, \\ \text{and } (\forall \mathbf{y}_j \in G^-) \qquad & \langle\mathbf{v}_2,\mathbf{y}_j\rangle = -1. \end{aligned} \tag{13}$$

Equalities (13) constitute a part of condition (10).

The set of equalities (13) can be represented in the matrix form:

$$(\exists \mathbf{v}_2)\ \mathbf{A}\mathbf{v}_2 = \mathbf{1}', \tag{14}$$

where $\mathbf{A}$ is the matrix of dimension $m \times (n+1)$, $m = m^+ + m^-$, and $\mathbf{1}'$ is the vector of dimension m. The rows of matrix $\mathbf{A}$ constitute of augmented feature vectors $\mathbf{y}_{j(i)}$.

Vector $\mathbf{y}_{j(i)}$ constitutes the i^{th} row of matrix $\mathbf{A}$. The i^{th} component of vector $\mathbf{1}'$ is equal to 1 if $\mathbf{y}_j \in G^+$ and equal to -1 if $\mathbf{y}_j \in G^-$.

Let us suppose that $m \leq n+1$ and that matrix $\mathbf{A}$ contains non-singular submatrix $\mathbf{B}$ of dimension $m \times m$ obtained from m independent columns of $\mathbf{A}$. In other words, matrix $\mathbf{B}$ is composed of m independent vectors $\mathbf{y}'_{j(i)}$ of dimension m. Vectors $\mathbf{y}'_j$ are obtained from feature vectors $\mathbf{y}_j$ by neglecting the same components x_i. It is clear in this case that the equation below:

$$\mathbf{B}\mathbf{v}'_2 = \mathbf{1}' \tag{15}$$

has the following solution:

$$\mathbf{v}'_2 = \mathbf{B}^{-1}\mathbf{1}'. \tag{16}$$

Let us remark that the $\mathbf{v}_2$ solution of Equation (13) also exists in this case. The $\mathbf{v}_2$ solution of Equation (13) can be extracted from Equation (15) by means of enlargement of vector $\mathbf{v}'_2$ by additional components equal to zero. The new components are put where the neglected x_i components of vectors $\mathbf{y}_j$ have been situated. The existence of the $\mathbf{v}_2$ solution of Equation (13) means that the G^+ and G^- sets are linearly separable (8).

LEMMA 1: If non-empty G^+ and G^- sets (5) contain no more than $(n+1)$ $(m \leq n+1)$ independent, $(n+1)$-dimensional feature vectors $\mathbf{y}_j$, then these sets are linearly separable (8).

The proof of this lemma can be based on the equations listed above (Equation (13) and Equation (16)) and on the above remarks.

Let us also remark that matrix $\mathbf{A}$ (14) may contain many non-singular submatrices, $\mathbf{B}_l$, of dimension $m \times m$ based on different feature subspaces, $F_l[m]$ (4) (Figure 1). Each of these feature subspaces $F_l[m]$ provides linear separability (8) of the G_l^+ and G_l^- sets, where the G_l^+ and G_l^- symbols stand for sets (5) of feature vectors $\mathbf{y}'_j$ which are constituted only of features x_i with indices i from the $I_l[m]$ set (4).

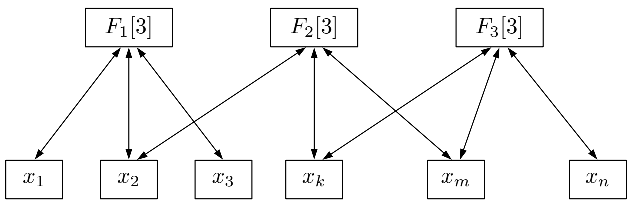

Figure 1. Each of the feature subspaces ($F_1[3]$, $F_2[3]$, and $F_3[3]$) provides linear separability (8) of the G_l^+ and G_l^- sets $(m=3)$

4. Convex and piecewise linear (CPL) criterion functions

Convex and piecewise linear (CPL) criterion functions are used to find optimal parameters $\mathbf{v}^*$ of the $H(\mathbf{v}^*) = \{\mathbf{y} : \langle \mathbf{v}^*, \mathbf{y} \rangle = 0\}$ hyperplane (2) separating the G^+ and G^- sets (5).

The *perceptron criterion function*, $\Psi(\mathbf{v})$, belongs to the CPL family [1, 5]. $\Psi(\mathbf{v})$ is the sum of the convex and piecewise linear penalty functions, $\varphi_j^+(\mathbf{v})$ and $\varphi_j^-(\mathbf{v})$ (Figure 2):

$$\varphi_j^+(\mathbf{v}) = \begin{cases} 1-\langle \mathbf{v},\mathbf{y}_j\rangle & \text{if } \langle \mathbf{v},\mathbf{y}_j\rangle < 1, \\ 0 & \text{if } \langle \mathbf{v},\mathbf{y}_j\rangle \geq 1 \end{cases} \tag{17}$$

and

$$\varphi_j^-(\mathbf{v}) = \begin{cases} 1+\langle \mathbf{v},\mathbf{y}_j\rangle & \text{if } \langle \mathbf{v},\mathbf{y}_j\rangle > -1, \\ 0 & \text{if } \langle \mathbf{v},\mathbf{y}_j\rangle \leq -1. \end{cases} \tag{18}$$

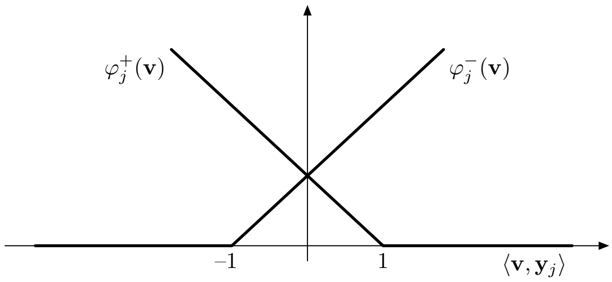

Figure 2. The penalty functions $\varphi_j^+(\mathbf{v})$ and $\varphi_j^-(\mathbf{v})$

Number 1 in Equations (16) and (17) represents the margins, $\delta_j(\delta_j = 1)$. If $\delta_j = 0$, then the penalty functions are related to the *error correction* algorithm used in the *Perceptron* [1]. The $\varphi_j^+(\mathbf{v})$ function is equal to zero if and only if vector $\mathbf{y}_j(\mathbf{y}_j \in G^+)$ is situated on the positive side of hyperplane $H(\mathbf{v})$ (5) and is not too close to it. Similarly, $\varphi_j^+(\mathbf{v})$ is equal to zero if vector $\mathbf{y}_j(\mathbf{y}_j \in G^-)$ is situated on the negative side of hyperplane $H(\mathbf{v})$ and is not too close to it.

The perceptron criterion function, $\Psi(\mathbf{v})$, can be defined on the G^+ and G^- sets (5) as follows:

$$\Psi(\mathbf{v}) = \sum_{\mathbf{x}_j \in G^+} \alpha_j \varphi_j^+(\mathbf{v}) + \sum_{\mathbf{x}_j \in G^-} \alpha_j \varphi_j^-(\mathbf{v}), \tag{19}$$

where non-negative parameters α_j determine the relative importance (*price*) of particular feature vectors $\mathbf{x}_j(k)$. Let us remark that the positive penalty functions, $\varphi_j^+(\mathbf{v})$, are defined on elements $\mathbf{y}_j$ of the G^+ set (5), while and the negative functions, $\varphi_j^-(\mathbf{v})$, are defined on the elements of the G^- set.

The perceptron criterion function, $\Psi(\mathbf{v})$, in its *standard form* has the following parameters:

$$\alpha_j = \begin{cases} 1/(2m^+) & \text{if } \mathbf{x}_j(k) \in G^+, \\ 1/(2m^-) & \text{if } \mathbf{x}_j(k) \in G^-. \end{cases} \tag{20}$$

We are interested in parameters $\mathbf{v}^*$ constituting the minimum of the $\Psi(\mathbf{v})$ function:

$$\Psi^* = \Psi(\mathbf{v}^*) = \min \Psi(\mathbf{v}). \tag{21}$$

Minimisation of the criterion function, $\Psi(\mathbf{v})$, results in minimisation of the penalty functions, $\varphi_j^+(\mathbf{v})$ and $\varphi_j^-(\mathbf{v})$. It has been proved that Ψ^* is equal to zero ($\Psi^* = 0$) if and only if the G^+ and G^- sets (5) are linearly separable (10):

$$(\Psi^* = 0) \Leftrightarrow (G^+ \text{ and } G^- \text{ are linearly separable}). \tag{22}$$

If the G^+ and G^- sets (5) are linearly separable, then the entire set G^+ is situated on the positive side of the $H(\mathbf{v}^*)$ hyperplane, and the entire set G^- is situated on the negative side of $H(\mathbf{v}^*)$.

The basic exchange algorithm allows us to find the minimum (20) efficiently, even if the multidimensional data sets G^+ and G^- (5) are large [9].

If the dimensionality of feature space F_l (4) is high, then there may exist many feature subsets $F_l[n']$ (4), which provide the linear separability of the G^+ and G^- sets (5). This possibility can be seen on the basis of relations (13) and (15).

Let us introduce an additional penalty function, $\phi_i(\mathbf{v})$, to the criterion function $\Psi(\mathbf{v})$, Equation (18), in order to compare the linear separability of sets G^+ and G^- in different feature subsets F_l (4). The $\phi_i(\mathbf{v})(i=1,\ldots,n+1)$ functions are defined as the absolute values, $|v_i|$, of weights v_i (Figure 3):

$$\phi_i(\mathbf{v}) = \begin{cases} -v_i & \text{if } v_i < 0, \\ v_i & \text{if } v_i \geq 0. \end{cases} \tag{23}$$

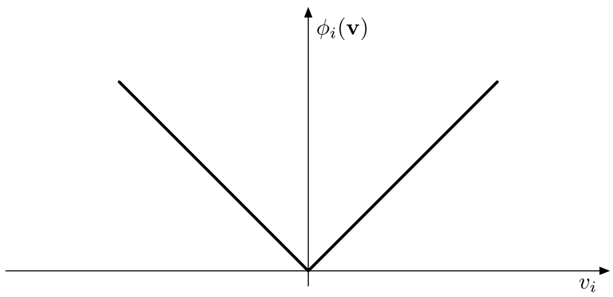

Figure 3. The penalty function $\phi_i(\mathbf{v})$

We can see that $\phi_i(\mathbf{v})$ are convex and piecewise-linear (CPL) functions. The penalty function $\phi_i(\mathbf{v})$ can be represented in the form similar to Equation (16) or Equation (17) by using unit vectors $\mathbf{e}_i = [0,\ldots,0,1,0,\ldots,0]^T(i=1,\ldots,n+1)$ with all components but i ones equal to zero and the i component equal to one:

$$\phi_i(\mathbf{v}) = \begin{cases} -\langle \mathbf{e}_i, \mathbf{v} \rangle & \text{if } \langle \mathbf{e}_i, \mathbf{v} \rangle < 0, \\ \langle \mathbf{e}_i, \mathbf{v} \rangle & \text{if } \langle \mathbf{e}_i, \mathbf{v} \rangle \geq 0. \end{cases} \tag{24}$$

Let us introduce modified criterion function, $\Phi_\lambda(\mathbf{v})$:

$$\Phi_\lambda(\mathbf{v}) = \Psi(\mathbf{v}) + \lambda \sum_{i \in I} \gamma_i \phi_i(\mathbf{v}), \tag{25}$$

where $\lambda \geq 0$, $\gamma_i > 0$, $I = \{1,\ldots,n+1\}$.

The $\Phi_\lambda(\mathbf{v})$ function is the sum of the perceptron criterion function $\Psi(\mathbf{v})$ (19) in the standard form (20) and the $\phi_i(\mathbf{v})$ penalty functions multiplied by positive parameters γ_i. The γ_i parameters represent the *costs* of particular features x_i. These costs can be chosen a priori, according to our additional knowledge.

The $\Phi_\lambda(\mathbf{v})$ criterion function (25) is a convex and piecewise-linear (CPL) function, as the sum of the CPL penalty functions $\alpha_j \varphi_j^+(\mathbf{v})$ (17), $\alpha_j \varphi_j^-(\mathbf{v})$ (18), and

$\lambda\gamma_i\phi_i(\mathbf{v})$ (24). As previously (21), we are looking for the minimal value of the $\Phi_\lambda(\mathbf{v})$ criterion function:

$$\Phi_\lambda(\mathbf{v}_\lambda^*) = \min_v \Phi_\lambda(\mathbf{v}). \tag{26}$$

The basic exchange algorithms allow us to efficiently solve the above minimisation problem.

Let us remark that if $\lambda = 0$, then $\Phi_\lambda(\mathbf{v}) = \Psi(\mathbf{v})$ and:

$$\mathbf{v}_0^* = \mathbf{v}^*, \tag{27}$$

where $\mathbf{v}^*$ is the minimum point (21) of the $\Psi(\mathbf{v})$ function (19). Otherwise, it can be proven that:

$$\mathbf{v}_\infty^* = \mathbf{0}, \tag{28}$$

where the symbol $\mathbf{v}_\infty^*$ means "the minimum point of the $\Phi_\lambda(\mathbf{v})$ function (25), with a very large value of the λ parameter".

If the i^{th} component of the optimal vector $\mathbf{v}_\lambda^*$ equals zero ($\mathbf{v}_{\lambda i}^* = 0$), then the i^{th} feature x_i can be neglected in the $\mathbf{y}_j(k)$ vectors without affecting the separation of the G^+ and G^- sets (5) by the optimal $H(\mathbf{v}_\lambda^*)$ hyperplane (2):

$$\{w_{\lambda i}^* = 0\} \Rightarrow \{\text{the } i^{\text{th}} \text{ feature } x_i \text{ can be neglected in all the } \mathbf{x}_j(k) \text{ vectors}\}. \tag{29}$$

Solution (28) means that none of the x_i features is used in designing the $H(\mathbf{v}_\infty^*)$ hyperplane (2). In other words, the $H(\mathbf{v}^*)$ hyperplane (2) cannot be designed by minimizing the $\Phi_\infty(\mathbf{v})$ criterion function because all weights $w_{\infty i}^*$ equal zero. The minimization of the $\Phi_0(\mathbf{v})$ criterion function is equivalent to taking into account the possibility that all features x_i could be used in the construction of the separating hyperplane, $H(\mathbf{v}_0^*)$. For some intermediate values of the λ parameter, some components $v_{\lambda i}^*$ of the optimal vector $\mathbf{v}_\lambda^*$ will equal zero, but others will be different from zero.

5. Cost sensitive measures of the data sets' linear separability

Let us assume that the G^+ and G^- sets (5) are linearly separable (8). Under this assumption it can be proved that, for sufficiently small values of the λ parameter, the optimal hyperplane $H(\mathbf{v}_\lambda^*) = H(\mathbf{w}_\lambda^*, \theta_\lambda^*)$ (2), (26) separates the G^+ and G^- sets:

$$\begin{aligned} (\exists\lambda^+)\ (\forall\lambda \in [0,\lambda^+])\ (\forall \mathbf{y}_j \in G^+) \qquad & \langle \mathbf{v}_\lambda^*, \mathbf{y}_j \rangle > 0 \\ \text{and } (\forall \mathbf{y}_j \in G^-) \qquad & \langle \mathbf{v}_\lambda^*, \mathbf{y}_j \rangle < 0, \end{aligned} \tag{30}$$

where $\mathbf{v}_\lambda^*$ is the weight vector constituting the minimum (26) of the $\Phi_\lambda(\mathbf{v})$ criterion function (25) with parameter λ. λ^+ is the maximum value of the λ parameter which still allows for the separation of the G^+ and G^- sets by the $H(\mathbf{v}_\lambda^*)$ optimal hyperplane.

Definition 2: The measure, Φ^*, of linear separability of the linearly separable G^+ and G^- sets (5) is equal to the minimal value, $\Phi_\lambda(\mathbf{v}_\lambda^*)$ (26), of the $\Phi_\lambda(\mathbf{v})$ criterion function (25) with the λ parameter equal to λ^+, Equation (30):

$$\Phi^* = \Phi_{\lambda^+}(\mathbf{v}_\lambda^*). \tag{31}$$

Let us remark that the G_l^+ and G_l^- sets (5) may be linearly separable in different feature subspaces $F_l[n']$ (4) (Figure 1). As a result, the measure of linear separability, Φ_l^* (31), may depend on the feature subset $F_l[n']$ used in the definition of the $\Phi_\lambda(\mathbf{v})$

criterion function (25). In other words, the measure of linear separability, Φ_l^* (31), may also allow for evaluation and comparison of such feature subsets $F_l[n']$ (4) that assure linear separability of the G_l^+ and G_l^- sets (5).

If the G_l^+ and G_l^- sets (5) are linearly separable (8) in feature space $F_l[n']$ (4), then the Φ_l^* measure can be expressed as:

$$\Phi_l^* = \Phi_{\lambda^+}(\mathbf{v}_\lambda^*) = \lambda_l^+ \sum_{i \in I_l} \gamma_i |v_i^*|, \tag{32}$$

where v_i^* are components of the optimal vector $\mathbf{v}_\lambda^*$ (26) in feature space $F_l[n']$ (4), λ_l^+ is the maximal value of the λ parameter (25) which still allows for linear separability (30) of the G_l^+ and G_l^- sets in $F_l[n']$ (4) by optimal hyperplane $H_l(\mathbf{v}_\lambda^*)$.

Let us introduce another measure, Γ_l^*, of linear separability in order to remove the dependence of the Φ_l^* measure on the λ_l^+ parameters (32):

$$\Gamma_l^* = \Phi_l^* / \lambda_l^+ = \sum_{i \in I_l} \gamma_i |v_i^*|. \tag{33}$$

If costs γ_i (25) are equal to one, then the Γ_l^* measure can be expressed as:

$$\Gamma_l^* = \sum_{i \in I_l} |v_i^*|. \tag{34}$$

Let us notice a similarities between the Γ_l^* measure (34) and the criterion used in Support Vector Machines (SVM) [6]. In the case of the linearly separable sets G_l^+ and G_l^- (5) the SVM criterion can be expressed as:

$$\min\{||\mathbf{v}_1||_2 : \mathbf{v}_1 \text{separates linearly (30) the } G_l^+ \text{ and } G_l^- \text{ sets (5)}\}, \tag{35}$$

where $||\mathbf{v}||_2 = \langle \mathbf{v}, \mathbf{v} \rangle^{1/2}$ is the Euclidean norm of the parameter vector $\mathbf{v}$.

We can see similarities between criterion (34) and the SVM criterion (35). The SVM criterion (35) is roughly aimed at such a parameter vector $\mathbf{v}_1^*$ that separates the G_l^+ and G_l^- sets and has the minimal Euclidean norm $||\mathbf{v}_1^*||_2$. Similarly, criterion (34) is aimed at such a parameter vector $\mathbf{v}_1'$ that separates the G_l^+ and G_l^- sets and has the minimal L_1 norm $||\mathbf{v}_1'||_1$.

Let us finally mention that minimization of an adequately adjusted criterion function, $\Phi_\lambda(\mathbf{v})$ (25), offers the possibility of tackling the following problems:

- Linear separability problem I:
  Find the smallest feature subset $F_l[n']$ (4) which still allows for linear separation (8) of sets G_l^+ and G_l^- (5).
- Linear separability problem II:
  Find such a feature subset $F_l[n']$ (4) with the number of elements no greater than n_0, which gives the largest distance between the linearly separable sets G_l^+ and G_l^- (5).

Solutions to the above feature selection problem could have very important applications in gene selection from genomic data [7].

6. Concluding remarks

The cost sensitive criterion Φ_l^* (32), which is based on the minimisation of the CPL function $\Phi_\lambda(\mathbf{v})$ (25), constitutes the general framework for the feature selection problem. By an adequate choice of costs γ_i (25) we are able to formulate a variety of specifications of the feature selection problem. It is also possible to find optimal hyperplanes $H_l(\mathbf{v}^*_\lambda)$ which best separate the G^+ and G^- sets. In particular, a special choice (34) of costs γ_i (25) allows us to find a solution similar to the SVM solution (35).

The basic exchange algorithms allow us to find efficiently the minimal value (26) of the $\Phi_\lambda(\mathbf{v})$ criterion function (25) with fixed parameter λ and fixed costs γ_i [9]. This technique allows us not only to compute the value of the separability measures for a given feature subspace $F_l[n']$ (4), but also to compare different subspaces $F_l[n']$ (4) providing linear separabilty.

The technique of feature selection based on the minimization of CPL functions has been applied by us in the *Hepar* medical diagnosis support system [2]. The future applications of this technique to gene selection problems appear to be very promising [7].

Acknowledgements

This work was partly supported by the grant W/II/1/2004 from the Bialystok University of Technology and by the grant 16/St/2004 from the Institute of Biocybernetics and Biomedical Engineering PAS.

References

[1] Duda O R, Hart P E and Stork D G 2001 *Pattern Classification*, J. Wiley, New York
[2] Bobrowski L (Ed) 1992 *Hepar – Computer System for Diagnossis Support and Data Analysis*, Internal Reports, IBIB PAS, Warsaw **31** (in Polish)
[3] Bobrowski L and Wasyluk H 2001 *Proc. 10th World Congress on Medical Informatics, MEDINFO 2001* (Patel V L, Rogers R and Haux R, Eds.), IMIA, IOS Press, Amsterdam, pp. 1309–1313
[4] Oniśko A, Druzdzel M J and Wasyluk H 2000 *Advances in Soft Computing* (Klopotek M, Michalewicz M and Wierzchon S T, Eds.), Physica-Verlag, Heidelberg, New York, pp. 303–313
[5] Bobrowski L 2000 *Biocybernetics and Biomedical Engineering* (Nałęcz M, Ed.), Akademicka Oficyna Wydawnicza Exit, Warsaw, **6** pp. 295–321 (in Polish)
[6] Vapnik V N 1998 *Statistical Learning Theory*, J. Wiley, New York
[7] Guyon I, Weston J, Barnhill S and Vapnik V 2002 *Machine Learning* **46** 389
[8] Bobrowski L, Wasyluk H and Niemiro W 1984 *Computers in Biology and Medicine* **14** (2) 237
[9] Bobrowski L 1991 *Pattern Recognition* **24** (9) 863