Selection of the Linearly Separable Feature Subsets

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

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

Abstract. We address a situation when more than one feature subset allows for linear separability of given data sets. Such situation can occur if a small number of cases is represented in a highly dimensional feature space.
The method of the feature selection based on minimisation of a special criterion function is here analysed. This criterion function is convex and piecewise-linear (*CPL*). The proposed method allows to evaluate different feature subsets enabling linear separability and to choose the best one among them. A comparison of this method with the *Support Vector Machines* is also included. ([3])

1 Introduction

The linear separability of data sets is one of the basic concepts in neural networks and pattern recognition [1]. This concept provided fundamentals for the Perceptron's theory [2], [3]. More recently, the linear separability is intensively explored in the method of the *Support Vector Machines* [4].

The feature selection in pattern recognition means neglecting such measurements (features) which have no significant influence on the final decisions. The feature selection is particularly important when the data sets are composed of a small number of elements in a highly dimensional feature space. The situation when a small number of elements is represented in a highly dimensional feature space (*long feature vectors*) usually leads to the linear separability of data sets. The genomic data sets contain examples of the "long feature vectors".

The measures of linear separability of two data sets can be based on the minimal value of the convex and piecewise-linear (*CPL*) criterion functions [5]. The perceptron criterion function belongs to the *CPL* family in question. The linear separability measures with different properties can be achieved through modification of the *CPL* criterion functions. Recently proposed CPL criterion function allows to compare different feature subsets enabling linear separability and to choose the best one among them [6]. This criterion function contains the *CPL* penalty functions reflecting the costs of the particular features.

[3] This work was partially 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.

The minimal value of the *CPL* functions can be found efficiently through applying the basis exchange algorithms, which can be treated as special methods for the linear programming [7]. The Support Vector Machines are based on the algorithms of the quadratic programming [4].

This paper is an analysis of the properties of the feature selection based on the modified *CPL* criterion function. Particular attention is paid to the comparison of the *CPL* criterion functions to the *Support Vector Machine* approach.

2 Linear Separability of Data Sets

Let us consider data represented as the feature vectors $\boldsymbol{x}_j[n] = [x_{j1}, ..., x_{jn}]^T$ $(j = 1, ..., m)$ of the same dimensionality n or as points in the n-dimensional feature space $F[n]$. The components x_i of the vectors $\boldsymbol{x}_j[n]$ are called features. We are considering a situation, when the data can be a mixed (a qualitative-quantitative) type. Some components x_{ji} of the vectors $\boldsymbol{x}_j[n]$ can be the binary $(x_i \in \{0, 1\})$ and others the real numbers $(x_i \in \mathbf{R}^1)$.

Let us take into consideration two disjoined sets G^+ and G^- composed of m feature vectors $\boldsymbol{x}_j$:

$$G^+ \cap G^- = \emptyset \ . \tag{1}$$

The *positive set* G^+ contains m^+ vectors $\boldsymbol{x}_j$ and the *negative set* G^- contains m^- vectors $(m = m^+ + m^-)$. We are considering the separation of the sets G^+ and G^- by the hyperplane $H(\boldsymbol{w}, \theta)$ in the feature space $F[n]$

$$H(\boldsymbol{w}, \theta) = \{\boldsymbol{x} : \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \theta\} \tag{2}$$

where $\boldsymbol{w} = [w_1, ..., w_n]^T \in \mathbf{R}^n$ is the weight vector, $\theta \in \mathbf{R}^1$ is the threshold, and $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ is the inner product.

Definition 1. *The feature vector $\boldsymbol{x}$ is situated on the* positive side *of the hyperplane $H(\boldsymbol{w}, \theta)$ if and only if $\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle > \theta$ and the vector $\boldsymbol{x}$ is situated on the* negative side *of $H(\boldsymbol{w}, \theta)$ iff $\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle < \theta$.*

Definition 2. *The sets G^+ and G^- are* linearly separable *if and only if they can be fully separated by some hyperplane $H(\boldsymbol{w}, \theta)$ (2):*

$$(\exists \boldsymbol{w}, \theta) \quad (\forall \boldsymbol{x}_j \in G^+) \langle \boldsymbol{w}, \boldsymbol{x}_j \rangle > \theta \quad \textbf{\textit{and}} \quad (\forall \boldsymbol{x}_j \in G^-) \langle \boldsymbol{w}, \boldsymbol{x}_j \rangle < \theta \ . \tag{3}$$

In accordance with the relation (3), all the vectors $\boldsymbol{x}_j$ belonging to the set G^+ are situated on the positive side of the hyperplane $H(\boldsymbol{w}, \theta)$ (2) and all the feature vectors $\boldsymbol{x}_j$ from the set G^- are situated on the negative side of this hyperplane. It is convenient to replace the feature vectors $\boldsymbol{x}_j$ by the *augmented* vectors $\boldsymbol{y}_j$, where

$$\boldsymbol{y}_j = [1, \boldsymbol{x}_j^T]^T = [1, x_{j1}, ..., x_{jn}]^T \ . \tag{4}$$

The inequalities (3) can be represented now as

$$(\exists \boldsymbol{v}) \quad (\forall \boldsymbol{y}_j \in G^+) \langle \boldsymbol{v}, \boldsymbol{y}_j \rangle > 0 \quad \textbf{and} \quad (\forall \boldsymbol{y}_j \in G^-) \langle \boldsymbol{v}, \boldsymbol{y}_j \rangle < 0 \tag{5}$$

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

3 From Linear Independence to Linear Separability

The linear separability of the sets G^+ and G^- can be defined equivalently to (5) in the following manner:

$$(\exists \boldsymbol{v}_1) \quad (\forall \boldsymbol{y}_j \in G^+)\langle \boldsymbol{v}_1, \boldsymbol{y}_j \rangle \geq 1 \quad \textbf{and} \quad (\forall \boldsymbol{y}_j \in G^-)\langle \boldsymbol{v}_1, \boldsymbol{y}_j \rangle \leq -1 \tag{6}$$

Remark 1. (*sufficient condition for linear separability*). The sets G^+ and G^- are *linearly separable* (6), if the following matrix equality is fulfilled:

$$(\exists \boldsymbol{v}_2)\boldsymbol{A}\boldsymbol{v}_2 = \textbf{1'} \tag{7}$$

where $\boldsymbol{A}$ is the matrix of dimension $m \times (n+1), \quad m = m^+ + m^-$, and **1'** is the vector of dimension m. The rows of the matrix A constitute of the augmented feature vectors $\boldsymbol{y}_{j(i)}$. The vector yj(i) constitutes the i-th row of the matrix A. The i-th component of the vector **1'** is equal to 1 if $\boldsymbol{y}_{j(i)} \in G^+$ and equal to -1 if $\boldsymbol{y}_{j(i)} \in G^-$.

Remark 2. If the m vectors $\boldsymbol{y}_{j(i)}$ constituting the matrix $\boldsymbol{A}$ are linearly independent, then there exists at least one nonsingular submatrix $\boldsymbol{B}$ of dimension $m \times m$ made of m independent columns of $\boldsymbol{A}$.

In other words, the matrix $\boldsymbol{B}$ is composed of m independent vectors $\boldsymbol{y'}_{j(i)}$ of dimension m. The vectors $\boldsymbol{y'}_j$ are constructed from the feature vectors $\boldsymbol{y}_j$ by means of neglecting of the same components x_i. In this case, the below equation

$$\boldsymbol{B}\boldsymbol{v'}_2 = \textbf{1'} \tag{8}$$

has the following solution:

$$\boldsymbol{v'}_2 = \boldsymbol{B}^{-1}\textbf{1'} \; . \tag{9}$$

Let us remark that the solution $\boldsymbol{v}_2$ of the equation (7) also exists in this case. The solution $\boldsymbol{v}_2$ (7) can be derived from (8) by means of enlarging the vector $\boldsymbol{v'}_2$ with additional components equal to zero. The new components are put in those places, where the neglected components x_i of the vectors $\boldsymbol{y}_j$ have been situated. The existence of the solution $\boldsymbol{v}_2$ of the equation (7) means that the sets G^+ and G^- are linearly separable (9). The above remarks allow to prove the following Lemma.

Lemma 1. *The sets G^+ and G^- (8) composed of m linearly independent feature vectors $\boldsymbol{y}_j$ are linearly separable in at least one m-dimensional feature subspace $F_k[m]$ ($F_k[m] \subset F[n], m \leq n$).*

The Lemma 1 points out an important fact, that the linear separability of the sets G^+ and G^- (5) may result from the linear independence of the feature vectors $\boldsymbol{y}_j$ constituting these sets. Such case often occurs in practice, when the number m of the vectors $\boldsymbol{y}_j$ in the sets G^+ and G^- is no greater than dimensionality $(n+1)$ of these vectors $(m \leq n+1)$.

4 Convex and Piecewise Linear (*CPL*) criterion function $\Phi_{\lambda}(\boldsymbol{v})$

The criterion function $\Phi_{\lambda}(\boldsymbol{v})$ is based on the *CPL* penalty functions $\varphi_j^+(\boldsymbol{v})$ or $\varphi_j^-(\boldsymbol{v})$ and $\phi_i(\boldsymbol{v})$. The functions $\varphi_j^+(\boldsymbol{v})$ are defined on the feature vectors $\boldsymbol{y}_j$ from the set G^+. Similarly $\varphi_j^-(\boldsymbol{v})$ are based on the elements $\boldsymbol{y}_j$ of the set G^-.

$$\begin{array}{lllll} \textbf{if} & (\boldsymbol{y}_j \in G^+) & \textbf{and} & (\langle \boldsymbol{v}, \boldsymbol{y}_j \rangle < 1) & \textbf{then} \quad \varphi_j^+(\boldsymbol{v}) = 1 - \langle \boldsymbol{v}, \boldsymbol{y}_j \rangle \\ \textbf{if} & (\boldsymbol{y}_j \in G^+) & \textbf{and} & (\langle \boldsymbol{v}, \boldsymbol{y}_j \rangle \geq 1) & \textbf{then} \quad \varphi_j^+(\boldsymbol{v}) = 0 \end{array} \tag{10}$$

and

$$\begin{array}{lllll} \textbf{if} & (\boldsymbol{y}_j \in G^-) & \textbf{and} & (\langle \boldsymbol{v}, \boldsymbol{y}_j \rangle > -1) & \textbf{then} \quad \varphi_j^-(\boldsymbol{v}) = 1 + \langle \boldsymbol{v}, \boldsymbol{y}_j \rangle \\ \textbf{if} & (\boldsymbol{y}_j \in G^-) & \textbf{and} & (\langle \boldsymbol{v}, \boldsymbol{y}_j \rangle \leq -1) & \textbf{then} \quad \varphi_j^-(\boldsymbol{v}) = 0 \end{array} \tag{11}$$

The penalty functions $\phi_i(\boldsymbol{v}) = |v_i|$ are related to particular features x_i.

$$\begin{array}{lll} \textbf{if} & (\langle \boldsymbol{e}_i, \boldsymbol{v} \rangle < 0) & \textbf{then} \quad \phi(\boldsymbol{v}) = -\langle \boldsymbol{e}_i, \boldsymbol{v} \rangle \\ \textbf{if} & (\langle \boldsymbol{e}_i, \boldsymbol{v} \rangle \geq 0) & \textbf{then} \quad \phi(\boldsymbol{v}) = \quad \langle \boldsymbol{e}_i, \boldsymbol{v} \rangle \end{array} \tag{12}$$

where $\boldsymbol{e}_i = [0, ..., 0, 1, 0, ..., 0]^T$ are the unit vectors $(i = 1, ..., n+1)$.
The criterion function $\Phi_{\lambda}(\boldsymbol{v})$ can be given in the following form:

$$\Phi_{\lambda}(\boldsymbol{v}) = \sum_{\boldsymbol{y}_j \in G^+} \alpha_j \varphi_j^+(\boldsymbol{v}) + \sum_{\boldsymbol{y}_j \in G^-} \alpha_j \varphi_j^-(\boldsymbol{v}) + \lambda \sum_{i \in I} \gamma_i \phi_i(\boldsymbol{v}) \tag{13}$$

where $\alpha_j \geq 0$, $\lambda \geq 0$, $\gamma_i > 0$, $I = \{1, ..., n+1\}$.
The nonnegative parameters α_j determine relative importance (*price*) of particular feature vectors $\boldsymbol{x}_j(k)$. The parameters γ_i. represent the *costs* of particular features x_i.We are using the minimal value of the criterion function $\Phi_{\lambda}(\boldsymbol{v})$:

$$\Phi_{\lambda}(\boldsymbol{v}*) = \min_{\boldsymbol{v}} \Phi_{\lambda}(\boldsymbol{v}) \tag{14}$$

The criterion function $\Phi_{\lambda}(\boldsymbol{v})$ (13) is the convex and piecewise linear (*CPL*) function as the sum of the *CPL* penalty functions $\alpha_j \varphi_j^+(\boldsymbol{v})$ (11), $\alpha_j \varphi_j^-(\boldsymbol{v})$ (12) and $\lambda \gamma_i \phi_i(\boldsymbol{v})$ (13). The basis exchange algorithm allows to find the minimum (18) efficiently, even in the case of large multidimensional data sets G^+ and G^- (1) [7]. The following *Lemma* can be proved:

Lemma 2. *If the sets G^+ and G^- (1) are linearly separable (5), and the prices γ_i are equal to 1 ($(\forall i \in I)\gamma_i = 1$), then there exists such value λ^+ that for a positive parameter λ which is no greater than λ^+ ($\forall \lambda \in (0, \lambda^+)$), the optimal vector $\boldsymbol{v}*$ (14) separates (5) these sets and*

$$\Phi_{\lambda}(\boldsymbol{v}*) = \lambda \sum_{i \in I} |v_i * | = \lambda \| \boldsymbol{v}* \|_{L1} \tag{15}$$

where $\boldsymbol{v}* = [v_1*, ..., v_n*]^T$ and $\| \boldsymbol{v}* \|_{L1} = \sum |v_i*|$ is the L_1 norm of the vector $\boldsymbol{v}*$.

The proof of this Lemma is based on the fact, that for sufficiently small parameter λ the minimal value $\Phi_\lambda(\boldsymbol{v}*)$ (14) of the function $\Phi_\lambda(\boldsymbol{v})$ (13) defined on the linearly separable sets G^+ and G^- (1) is equal to

$$\Phi_\lambda(\boldsymbol{v}*) = \lambda \sum_{i \in I} \gamma_i \phi_i(\boldsymbol{v}*) \tag{16}$$

The above equality results from the property, that the values of all the penalty functions $\varphi_j^+(\boldsymbol{v})$ and $\varphi_j^-(\boldsymbol{v})$ are equal to zero in the optimal point $\boldsymbol{v}*$ for the linearly separable case.

As it results from the *Lemma 2*, in the case of linearly separable sets G^+ and G^- (1) minimisation of the function $\Phi_\lambda(\boldsymbol{v})$ (13) with a small parameter λ leads to the optimal vector $\boldsymbol{v}*$ which not only separates these sets, but also has the minimal value of the L_1 norm of this vector.

5 Comparisons of the *Support Vector Machines* with the *CPL* Approach

The linear separability of the sets G^+ and G^- (5) by the vector $\boldsymbol{v}*$ (14) can be formulated as:

$$\begin{array}{rll} & (\forall \boldsymbol{y}_j \in G^+) & \langle \boldsymbol{v}*/\| \boldsymbol{v}* \|, \boldsymbol{y}_j \rangle \geq 1/\| \boldsymbol{v}* \| \\ \textbf{and} & (\forall \boldsymbol{y}_j \in G^-) & \langle \boldsymbol{v}*/\| \boldsymbol{v}* \|, \boldsymbol{y}_j \rangle \leq 1/\| \boldsymbol{v}* \| \end{array} \tag{17}$$

If the Euclidean norm ($\| \boldsymbol{v}* \| = \langle \boldsymbol{v}*, \boldsymbol{v}* \rangle$) is used, the inequalities (17) mean that the sets G^+ and G^- (10) are separated by the hyperplane $H(\boldsymbol{v}*) = \{\boldsymbol{y} : \langle \boldsymbol{v}*, \boldsymbol{y} \rangle = 0\}$ (2) with the *margin* $\delta = 2/\| \boldsymbol{v}* \|$. Minimization of the norm $\| \boldsymbol{v}* \|$ means that the margin δ between the sets G^+ and G^- (10) becomes maximal. Such approach has been adopted in the *Support Vector Machine (SVM)* method in order to optimize location of the separating hyperplane $H(\boldsymbol{v}*)$ (2) [7]. The quadratic programming is applied in order to find the minimal value of the margin $2/\| \boldsymbol{v}* \|$ under the condition of the linear separability (17).

Let the symbols $G_l^+[m]$ and $G_l^-[m]$ stand for the positive and negative sets (1) composed of the m-dimensional feature vectors $\boldsymbol{y}_j[m]$ from the subspace $F_k[m]$ ($F_k[m] \subset F[n]$). The sets $G_k^+[m]$ and $G_k^-[m]$ can be linearly separable (5) in the subspace $F_l[m]$. The minimal value $\Phi_\lambda(\boldsymbol{v}_k^*[m])$ (14) of the *CPL* criterion function $\Phi_\lambda(\boldsymbol{v}[m])$ (13) defined on the vectors $\boldsymbol{y'}_j[m]$ can be used as the measure of the linear separability of the subspace $F_k[m]$. In other words, minimisation of the criterion function $\Phi_\lambda(\boldsymbol{v})$ (13) allows to compare different feature subspaces $F_k[m]$ and to choose the best one $F_k^*[m]$ from them.

The basis exchange algorithm adjusted to minimisation of the *CPL* criterion functions $\Phi_k(\boldsymbol{v}[m])$ (13) in different subspaces $F_k[m]$ has been designed and implemented. This algorithm allows to find the best feature subspace $F_k^*[m]$ through the sequence of the below type:

$$F_1[m] \rightarrow F_2[m] \rightarrow \ldots \cdots \rightarrow F_k[m] = F_k^*[m] \tag{18}$$

where

$$\Phi_1^*(\boldsymbol{v}[m]) \geq \Phi_2^*(\boldsymbol{v}[m]) \geq \ldots \cdots \geq \Phi_k^*(\boldsymbol{v}[m]) = \Phi_k^*(\boldsymbol{v}[m]) \tag{19}$$

In accordance with the above relations, the sequence of the linearly separable feature subspaces $F_k[m]$ is designed in a such manner, that the minimal values $\Phi_k^*(\boldsymbol{v}[m])$ of the criterion functions $\Phi_k(\boldsymbol{v}[m])$ (13) in the successive subspaces $F_k[m]$ is decreasing. Each feature subspace $F_k[m]$ assures linear separability of the sets $G_k^+[m]$ and $G_k^-[m]$. In this case, the decreasing of the minimal values $\Phi_k^*(\boldsymbol{v}[m])$ means the decreasing of the L_1 type distance (15), (17) between the sets $G_k^+[m]$ and $G_k^-[m]$.

6 Concluding Remarks

The proposed method of the selection of the optimal feature subspace $F_k^*[m]$ is based on directed search among linearly separable feature subspace $F_k[m]$. This search can be implemented as an efficient basis exchange procedure based on the sequence (18) with the property (19).

Selection of the feature subspaces $F_k^*[m]$ with best linear separability may be applied in solving many problems. One of the most interesting possibilities is gene extraction [8]. Another group of important applications is related to designing hierarchical neural networks and multivariate decision trees on the basis of the learning sets G_k (1) with a "long feature vectors". The ranked and the dipolar designing strategies can be combined with the procedure proposed here of the optimal feature subspace $F_k^*[m]$ selection [9].

References

1. Duda, O.R., Hart, P.E., Stork, D.G.: Pattern Classification, J.Wiley, New York (2001)
2. Bishop, Ch.M.: Neural Networks for Pattern Recognition, Clarendon Press, Oxford (1995)
3. Ripley, B.D.: Pattern Recognition and Neural Networks, Cambridge Univ. Press (1996)
4. Vapnik, V.N.: Statistical Learning Theory, J.Wiley, New York (1998)
5. Bobrowski, L.: Piecewise-Linear Classifiers, Formal Neurons and Separability of the Learning Sets, Proceedings of ICPR'96, 13th International Conference on Pattern Recognition, Wienna, Austria (August 25-29, 1996) 224–228
6. Bobrowski, L., The Method of the Feature Selection Based on the Linearly Separable Learning Sets, Proceedings of the 13th Internal Scientific Conference Biocybernetics and Biomedical Engineering, Edited by A. Nowakowski, Gdansk (2003) 237–242 (in Polish)
7. Bobrowski, L.: Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, Pattern Recognition, 24(9), (1991) 863–870
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, 46, (2002) 389–422
9. Bobrowski, L.: Strategies of Designing Neural Networks, Neural Networks, Vol. 6 in monography: Biocybernetics and Biomedical Engineering, Edited by M. Nalecz, Academic Publishing House Exit, Warsaw (2000) 295–321 (in Polish)