ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2007
Informatyka – Zeszyt 2

Leon Bobrowski[1,2]

SEPARABLE DATA AGGREGATION BY LAYERS OF ELEMENTARY CLASSIFIERS

Abstract: Data exploration or data mining goals can be reached by using variety of methods such as the fuzzy set theory or the rough sets theory. An interesting group of data exploration methods is based on minimization of convex and piecewise linear (*CPL*) criterion functions. This method originated from the theory of neural networks (multilayer *Perceptrons*). Powerful methods of data mining based on the support vector machines (*SVM*) can be also linked to this concept.

Hierarchical networks of formal neurons or multivariate decision trees can be induced from learning sets through minimization *CPL* criterion functions specified for classification problem. Another type of the *CPL* criterion functions can be used for designing visualizing data transformations. Separability of the transformed learning sets is a fundamental concept in the *CPL* approach to designing data mining tools.

Keywords: data transformations, data aggregation, separable data sets, elementary classifiers, convex and piecewise linear (*CPL*) criterion function

1. Introduction

Data exploration is aimed at discovering regularities (*patterns*) in data sets and at designing such data models which take into account these patterns. Many data exploration goals can be reached through the process of pattern recognition [1], [2], [3]. In this approach each object or event is represented as a feature vector or as a point in a multidimensional feature space. The pattern recognition process includes three basic stages: feature selection, feature extraction and classification. The primary goal of classification is proper allocation of each object or event into one of the classes (categories). This goal is often achieved by using variety of tools originating, among others, from case based reasoning techniques, neural networks models, decision trees approach. Feature selection stage is aimed at reducing dimensionality of the feature space through neglecting such features which are not

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

important in the classification process. The dimensionality reduction can be achieved also during feature extraction stage in result of feature space transformations.

Many concepts in the theory and applications of artificial neural networks and pattern recognition has his beginning in the model of the Perceptron [4]. Hierarchical layers of formal neurons (multilayer perceptrons) still belong to the most fundamental models of neural networks [1]. Designing a neural network relates to the choice of a neural network structure (e.g. the number of layers and the number of elements in particular layers) and the weights of connections between elements of successive layers. The above designing tasks can be performed through minimization of the convex and piecewise linear (*CPL*) criterion functions deriving from the Perceptron model [4]. Linear separability of learning sets in a selected feature space is a big issue of the perceptron theory and plays a central role in applications the perceptron *CPL* criterion function. Similar *CPL* criterion functions can be used, for example, in designing decision trees, designing data transformation for feature extraction, feature selection, or data visualization. These topics are discussed more closely in the presented paper. Particular attention is paid to problems of designing separable layers of elementary classifiers.

2. Separable learning sets

Let us assume that each of the m analysed objects O_j ($j = 1,...,m$) is represented as the feature vector $\mathbf{x}_j = [x_{j1},...,x_{jn}]^T$ or as a point in the n-dimensional *feature space* $F[n]$ ($\mathbf{x}_j \in F[n]$). The components (*features*) x_{ij} of the vector $\mathbf{x}_j$ are supposed to be numerical results of a variety of examinations of the given object O_j. The feature vectors $\mathbf{x}_j$ can be of mixed, qualitative-quantitative type with n binary or real components x_{ij} ($x_{ij} \in \{0,1\}$ or $x_{ij} \in R$).

We assume that the database contains descriptions $\mathbf{x}_j(k)$ of m objects $O_j(k)$ ($j = 1,...,m$) labelled according to their *category* (*class*) ω_k ($k = 1,...,K$). The learning sets C_k can be created on this basis. One learning set C_k contains m_k feature vectors $\mathbf{x}_j(k)$ assigned to the k-th category ω_k

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

where I_k is the set of indices j of such feature vectors $\mathbf{x}_j(k)$ from the class ω_k which belong to the set C_k.

Definition 1: The learning sets C_k (1) are *separable* in the feature space $F[n]$, if they are disjunctive in this space ($C_k \cap C_{k'} = \phi$ for $k \neq k'$). It means that feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k')$ from different learning sets C_k and $C_{k'}$ cannot be equal:

$$(k \neq k') \Rightarrow \quad (\forall j \in I_k) \quad and \quad (\forall j' \in I_{k'}) \quad \mathbf{x}_j(k) \neq \mathbf{x}_{j'}(k') \tag{2}$$

We are also considering the separation of the sets C_k (1) by the hyperplanes $H(\mathbf{w}_k, \theta_k)$ in the feature space $F[n]$

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

where $\mathbf{w}_k = [w_{k1}, ..., w_{kn}]^T \in R^n$ is the weight vector, $\theta_k \in R^1$ is the threshold, and $\mathbf{w}_k{}^T \mathbf{x}$ is the inner product.

Definition 2: The learning sets (1) are *linearly separable* in the *n*-dimensional feature space $F[n]$ if each of these sets C_k can be fully separated by some hyperplane $H(\mathbf{w}_k, \theta_k)$ (3) from the sum $\cup C_i$ ($i \neq k$) of the remaining sets C_i:

$$\begin{aligned} &(\exists k \in \{1,...,K\})(\exists \mathbf{w}_k, \theta_k) \quad (\forall \mathbf{x}_j(k) \in C_k) \quad \mathbf{w}_k{}^T \mathbf{x}_j(k) > \theta_k \\ &\qquad and \quad (\forall \mathbf{x}_{j'}(k') \in C_{k'}, k' \neq k) \quad \mathbf{w}_k{}^T \mathbf{x}_{j'}(k') < \theta_k \end{aligned} \tag{4}$$

In accordance with relation (4), all the vectors $\mathbf{x}_j(k)$ belonging to the learning set C_k are situated on the positive side ($\mathbf{w}_k{}^T \mathbf{x}_j(k) > \theta_k$) of the hyperplane $H(\mathbf{w}_k, \theta_k)$ (3) and all the feature vectors $\mathbf{x}_{j'}(k')$ from the remaining sets C_i are situated on the negative side ($\mathbf{w}_k{}^T \mathbf{x}_{j'}(k') < \theta_k$) of this hyperplane.

The separation of data sets C_k by the hyperplanes $H(\mathbf{w}_k, \theta_k)$ (3) can by linked to data transformation by a layer of K formal neurons $FN(\mathbf{w}_k, \theta_k)$. The formal neuron $FN(\mathbf{w}_k, \theta_k)$ is defined by the threshold decision rule $q(\mathbf{w}_k, \theta_k; \mathbf{x})$:

$$q = q(\mathbf{w}_k, \theta_k; \mathbf{x}) = \begin{cases} 1 & if \quad \mathbf{w}_k{}^T \mathbf{x} \geq \theta_k \\ 0 & if \quad \mathbf{w}_k{}^T \mathbf{x} < \theta_k \end{cases} \tag{5}$$

where q is the output, $\mathbf{w}_k = [w_{k1},...,w_{kn}]^T \in R^n$ is the weight vector, $\theta_k \in R^1$ is the threshold and $\mathbf{x} = [x_1,...,x_n]^T$ is the input feature vector.

The feature vector $\mathbf{x}$ activates ($r = 1$) the formal neuron $FN(\mathbf{w}_k, \theta_k)$ if and only if $\mathbf{x}$ is situated on the positive side of the hyperplane $H(\mathbf{w}_k, \theta_k)$ ($\mathbf{w}_k{}^T \mathbf{x} \geq \theta_k$).

Layer of K formal neurons $FN(\mathbf{w}_i, \theta_i)$ transforms the feature vectors $\mathbf{x}$ into the binary vectors $\mathbf{q} = \mathbf{q}(\mathbf{x})$, where $\mathbf{q} = [q_1,...,q_K]$, $q_i = q(\mathbf{w}_i, \theta_i; \mathbf{x})$ (5). Such a layer can be used as the classifier with the allocation rule given below

$$if \quad (q(\mathbf{w}_k, \theta_k; \mathbf{x}) = 1) \quad and \quad (\forall i \neq k) q(\mathbf{w}_i, \theta_i; \mathbf{x}) \quad then \quad (\mathbf{x} \in \omega_k) \tag{6}$$

A vector $\mathbf{x}$ is allocated to the class ω_k if only one neuron $FN(\mathbf{w}_k, \theta_k)$ in this layer is activated. We can remark that if the learning sets C_k (1) are linearly separable (4), then the layer of K formal neurons $FN(\mathbf{w}_i, \theta_i)$ with the rule (6) can allocate properly all the feature vectors $\mathbf{x}_j(k)$ (1).

3. Elementary classifiers

Let us take into account a layer of L *elementary classifiers* $Q_i = Q_i(\mathbf{v}_i)$ ($i = 1,..., L$) with the binary outputs q_i ($q_i \in \{0,1\}$). Each classifier Q_i is defined on the feature vectors $\mathbf{x}$ by an individual decision rule $q_i = q_i(\mathbf{v}_i; \mathbf{x})$:

$$q_i = q_i(\mathbf{v}_i; \mathbf{x}) \quad (i = 1,..., L) \tag{7}$$

where $\mathbf{v_i} = [v_{i1},..., v_{in'}]^T$ is n' - dimensional vector of parameters.

The classifier Q_i is activated by the feature vectors $\mathbf{x}$ if and only if $q_i(\mathbf{v}_i; \mathbf{x}) = 1$. Formal neurons $FN(\mathbf{w}_k, \theta_k)$ (5) can be used as the elementary classifiers Q_i.

Definition 3: The *activation field* S_i of the elementary classifier $Q_i = Q_i(\mathbf{v}_i)$ is defined as the set of such feature vectors $\mathbf{x}$, which activates ($q_i(\mathbf{v}_i;\mathbf{x}) = 1$) this classifier.

$$S_i = \{\mathbf{x} : q_i(\mathbf{v}_i;\mathbf{x}) = 1\} \tag{8}$$

The layer of L elementary classifiers Q_i transforms each feature vector $\mathbf{x}_j(k)$ from the sets C_k (1) into the vector $\mathbf{q}_j(k)$ with L binary components $q_i = q_i(\mathbf{v}_i;\mathbf{x}_j(k))$.

$$\mathbf{q}_j(k) = [q_1(\mathbf{v}_1;\mathbf{x}_j(k)),...,q_L(\mathbf{v}_L;\mathbf{x}_j(k))]^T \tag{9}$$

where $\mathbf{v_i} = [v_{i1},...,v_{in'}]^T$ is a vector of parameters.

Vectors $\mathbf{q}_j(k)$ form new data representation which can be useful in designing valuable decision rules for classification purpose [2]. The decision rules could be designed more efficiently on the basis of the transformed vectors $\mathbf{q}_j(k)$ than on the basis of the feature vectors $\mathbf{x}_j(k)$. An example of the decision rule is given by (6). The transformed vectors $\mathbf{q}_j(k)$ form the sets D_k:

$$D_k = \{\mathbf{q}_j(k)\} \quad (j \in I_K) \tag{10}$$

One of the fundamental goals in designing layers of elementary classifiers Q_i could be the separability (2) or the linear separability (4) of the transformed sets D_k. Additionally, we could demand the *separable aggregation* of the learning sets C_k (1).

Definition 4: The transformation (9) results in the *separable aggregation* of the learning sets C_k (1) if and only if the transformed sets D_k (10) are separable (2), each feature vector $\mathbf{x}_j(k)$ (1) activates at least one elementary classifiers Q_i (7) of the layer, and the number m' of different vectors $\mathbf{q}_j(k)$ (9) in the sets D_k is less than m.

Classification postulate I: The transformation (9) defined by the layer of L' elementary classifiers Q_i should result in the separable (2) sets D_k (10) with a low number m' of different vectors $\mathbf{q}_j(k)$ (9) and a low dimensionality L' of these vectors.

Few examples of the elementary classifiers Q_i are given below:

Example 1: Formal neurons $FN(\mathbf{w}_i,\theta_i)$ (5) can be treated as the elementary classifiers Q_i (8). In this case, the decision rule $q_i = q_i(\mathbf{v}_i;\mathbf{x}_j(k))$ (7) is based on the vector of parameters $\mathbf{v}_i$ given below:

$$\mathbf{v}_i = [\mathbf{w}_i^T, \theta_i]^T \tag{11}$$

The activation field S_i (8) of the formal neuron $FN(\mathbf{w}_i,\theta_i)$ (5) is the positive half-space defined by the hyperplane $H(\mathbf{w}_i,\theta_i)$ (3).

Example 2: The number n of inputs $\mathbf{x}_j$ to formal neuron $FN(\mathbf{w}_i,\theta_i)$ (5) can be reduced to one. Such reduced neuron will be called as *logical element* $LE(w_i,\theta_i)$. The elementary classifiers Q_i are determined in this case by the vector of parameters $\mathbf{v}_i = [w_i,\theta_i]^T$ (11) with only two components w_i and θ_i. The decision rule $q_i = q_i(\mathbf{v}_i;\mathbf{x}_j(k))$ (8) can be reduced to the below form:

$$if \quad (w_i x_{ji}(k) \geq \theta_i) \quad then \quad (q_i(\mathbf{v}_i;\mathbf{x}_j(k)) = 1) \quad else \quad (q_i(\mathbf{v}_i;\mathbf{x}_j(k)) = 0) \tag{12}$$

The activation field S_i (8) of the logical element $LE(w_i,\theta_i)$ is the positive half-space defined by such hyperplanes $H(\mathbf{w}_k,\theta_k)$ $H(\mathrm{w_k},\theta_\mathrm{k})$ (3), which are parallel to all but one axis of the feature space $F[n]$.

Example 3: The elementary classifier Q_i (7) can be based on the Euclidean ball $K_E(\mathbf{w}_i,\rho_i)$ centered in the point $\mathbf{w}_i$ and with the radius ρ_i in the feature space $F[n]$:

$$K_E(\mathbf{w}_i,\rho_i) = \{\mathbf{x} : (\mathbf{x}-\mathbf{w}_i)^T(\mathbf{x}-\mathbf{w}_i) \leq \rho_i\} \tag{13}$$

The vector of parameters $\mathbf{v}_i = [\mathbf{w}_i^T, \rho_i]^T$ defines the decision rule $q_i = q_i(\mathbf{v}_i;\mathbf{x}_j(k))$ (7) in the below manner

$$\begin{gathered} if \quad ((\mathbf{x}_j(k)-\mathbf{w}_i)^T(\mathbf{x}_j(k)-\mathbf{w}_i) \leq \rho_i) \quad then \quad (q_i(\mathrm{v}_i;\mathbf{x}_j(k)) = 1) \\ else \quad (q_i(\mathrm{v}_i;\mathbf{x}_j(k)) = 0) \end{gathered} \tag{14}$$

The activation field S_i (8) of such elementary classifier Q_i (7) is the ball $K_E(\mathbf{w}_i, \rho_i)$ (13).

Example 4: The $L1$ ball $K_{L1}(\mathbf{w}_i, \rho_i)$ centered in the point $\mathbf{w}_i$ and with the radius ρ_i in the feature space $F[n]$ also can serve as an elementary classifier Q_i (8).

$$K_{L1}(\mathbf{w}_i, \rho_i) = \{\mathbf{x} : |x_1 - w_1| + ... + |x_n - w_n| \leq \rho_i\} \tag{15}$$

The decision rule $q_i = q_i(\mathbf{v}_i; \mathbf{x}_j(k))$ (7) has now the following form:

$$\begin{aligned} &if \quad (|x_1 - w_1| + ... + |x_n - w_n| \leq \rho_i) \quad then \quad (q_i(\mathrm{v}_i; \mathbf{x}_j(k)) = 1) \\ &\qquad\qquad else \quad (q_i(\mathrm{v}_i; \mathbf{x}_j(k)) = 0) \end{aligned} \tag{16}$$

The activation field S_i (8) of this elementary classifier Q_i (7) is the ball $K_{L1}(\mathbf{w}_i, \rho_i)$ (15).

Example 5: The $L1$ ball $K_{L1}(\mathbf{w}_i, \rho_i)$ (15) can be generalized to the ball $K_P(\mathbf{w}_i, \rho_i)$

$$K_{L1}(\mathbf{w}_i, \alpha_{i,} \rho_i) = \{\mathbf{x} : \alpha_{i1}|x_1 - w_1| + ... + \alpha_{in}|x_n - w_n| \leq \rho_i\} \tag{17}$$

where $\alpha_i = [\alpha_{i1}, ..., \alpha_{in'}]^T$ is the vector of *features costs* α_{ik}.

The decision rule $q_i = q_i(\mathbf{v}_i; \mathbf{x}_j(k))$ (16) is generalised with the parameters $\mathrm{v}_i = [\mathbf{w}_i^T, \alpha_i^T, \rho_i]^T$ to:

$$\begin{aligned} &if \quad (\alpha_{i1}|x_{j1}(k) - w_1| + ... + \alpha_{in}|x_{jn}(k) - w_n| \leq \rho_i) \\ &\quad then \quad (q_i(\mathbf{v}_i; \mathbf{x}_j(k)) = 1) \quad else \quad (q_i(\mathbf{v}_i; \mathbf{x}_j(k)) = 0) \end{aligned} \tag{18}$$

Let us remark that the number of parameters in the above rule has been increased to ($2n+1$) in comparison to ($n+1$) parameters used in the rule (16).

4. Dipolar strategy of separable layers designing

Lets us take into consideration the problem of designing separable layers of elementary classifiers Q_i ($i = 1, ..., L$) (7). “Separable layer”is such a layer of L

elementary classifiers Q_i (7) which results in the separability (2) of the transformed sets D_k (10). The dipolar and the ranked strategies of designing separable layers of formal neurons were proposed earlier [6], [7]. Now, we will generalize these strategies to the layers of elementary classifiers Q_i (7). We will start with the description of the dipolar strategy. This strategy is based on the concept of clear and mixed dipoles [5].

Definition 5: A pair of different feature vectors $(\mathbf{x}_j(k), \mathbf{x}_{j'}(k'))$ ($\mathbf{x}_j(k) \neq \mathbf{x}_{j'}(k')$) constitutes a *mixed dipole* if and only if these vectors belong to different classes ω_k ($k \neq k'$). Similarly, a pair of different feature vectors from the same class ω_k constitutes the *clear dipole* $(\mathbf{x}_j(k), \mathbf{x}_{j'}(k'))$.

Definition 6: The elementary classifier Q_i (7) *separates* (*divides*) the dipole $(\mathbf{x}_j(k), \mathbf{x}_{j'}(k'))$ if only one feature vector $\mathbf{x}_j(k)$ or $\mathbf{x}_{j'}(k')$ from this pair activates this element ($q_i(\mathbf{v}_i; \mathbf{x}_j(k)) = 1$ and $q_i(\mathbf{v}_i; \mathbf{x}_{j'}(k')) = 0$ or $q_i(\mathbf{v}_i; \mathbf{x}_j(k)) = 0$ and $q_i(\mathbf{v}_i; \mathbf{x}_{j'}(k')) = 1$).

Lemma 1: The necessary and sufficient condition for the separability (*Def. 1*) of the sets D_k (10) transformed by the layer (9) is the separation of each mixed dipole $(\mathbf{x}_j(k), \mathbf{x}_{j'}(k'))$ by at least one elementary classifier Q_i (7) of the layer.

The proof of similar result for layer of formal neurons $FN(\mathbf{w}, \theta)$ (4) has been given in [5], [8]. In accordance with the *Lemma 1*, a layer which divides all mixed dipoles transforms separable sets C_k (1) into separable sets D_k (10). In order to preserve the chance for correct classification of all feature vectors $\mathbf{x}_j(k)$ (1), an additional postulate is introduced:

*Classification postulat*e *II*: Each feature vector $\mathbf{x}_j(k)$ (1) should activate ($q_i(\mathbf{v}_i; \mathbf{x}_j(k)) = 1$) at least one elementary classifier Q_i (7) of a given layer (9) .

5. Ranked strategy of separable layers designing

Let us take into consideration the ranked strategy of designing separable layers of elementary classifiers Q_i ($i = 1,\ldots,L$) (7). This strategy uses a fixed order between elementary classifiers Q_i of the layer which is based on the indexing of these classifiers. The relation "*prior to*" is defined between any two elementary classifiers Q_l and Q_i of the layer on the base of the indices l and i in the below manner.

Definition 7: The classifier Q_i (7) is *prior to* the classifier Q_l if and only if $i < l$.

Definition 8: The l-th *ranked field* R_l ($l = 1,...,L$) of the layer of L elementary Q_i classifiers is a set of such feature vectors $\mathbf{x}_j(k)$ (1) which activate the l-th classifier Q_l and do not activate any of the prior classifiers Q_i.

$$R_l = \{\mathbf{x}_j(k) : q_l(\mathrm{v}_i;\mathbf{x}_j(k)) = 1 \quad \textit{and} \quad (\forall i < l) q_i(\mathrm{v}_i;\mathbf{x}_j(k)) = 0\} \tag{19}$$

Definition 9: The *r*anked field R_i (19) is *deterministically admissible* if and only if it contains feature vectors $\mathbf{x}_j(k)$ from only one learning set C_k (1).

Definition 10: The *r*anked field R_i (19) is *statistically admissible* at the level α ($0 < \alpha < 0.5$) if and only if it contains feature vectors $\mathbf{x}_j(k)$ not only from the dominant set C_k but also from other sets C_i (1) in a fraction f_i less than α ($f_i < \alpha$).

The fraction f_i of elements $\mathbf{x}_j(l)$ from non-dominant sets C_l is defined by the expression below:

$$f_i = \frac{m_i'(k)}{m_i(k) + m_i'(k)} \tag{20}$$

where $m_i(k)$ is the number of elements $\mathbf{x}_j(k)$ from the dominant set C_k in the ranked field R_i (19) and $m_i'(k)$ is the number of elements $\mathbf{x}_j(l)$ in this field from all non-dominant sets C_l (1) ($m_i(k) > m_i'(k)$.

The layer of L elementary classifiers Q_i (7) with admissible ranked fields R_i (19) will be called an admissible one (deterministically or statistically admissible). It can be seen that the number L of classifiers R_i in an admissible layer fulfills below condition.

$$K \le L \le m \tag{21}$$

where K is the number of the learning sets C_k (1), and m is the number of feature vectors $\mathbf{x}_j(k)$ in these sets.

The lowest possible number $L = K$ appears when the ranked fields R_i (19) are extremely large and contains whole learning sets C_k ($\forall k \in \{1,...,K\}$ $R_k = C_k$ (1). The highest possible number $l = m$ appears when each ranked field R_i (19) contains only one feature vector $\mathbf{x}_j(l)$. It can be expected that layer of classifiers Q_i with large ranked fields R_i (19) should have greater generalizing power than the layer with small active fields.

Definition 11: The layer of elementary classifiers Q_i (7) with deterministically admissible (*Def. 7*) ranked fields R_i (19) forms the *ranked layer* if and only if each feature vector $\mathbf{x}_j(k)$ from the sets C_k (1) belongs to one of this fields.

The ranked layer of L elementary classifiers Q_i transforms each feature vector $\mathbf{x}_j(k)$ into the vector $\mathbf{q}_j(k)$ (9) with L binary components $q_i = q_i(\mathbf{v}_i;\mathbf{x}_j(k))$. The separability (2) of the sets C_k (1) is preserved during the transformation by the ranked layer as it is proven below.

Lemma 2: If the sets C_k (1) are separable (2), then the sets D_k (10) at the output of the ranked layers are also separable.

Proof: The sufficient condition for the sets D_k (10) separability has the form (2).

$$(k \neq k') \Rightarrow (\forall j \in I_k) \quad and \quad (\forall j' \in I_{k'}) \mathbf{q}_j(k) \neq \mathbf{q}_{j'}(k') \tag{22}$$

The above condition results directly from the definition of the ranked fields R_i (19). Two vectors $\mathbf{q}_j(k)$ and $\mathbf{q}_{j'}(k')$ related to the ranked fields R_j and $R_{j'}$ are linked to different classes ω_k and $\omega_{k'}$. So, these vectors cannot be equal ($\mathbf{q}_j(k) \neq \mathbf{q}_{j'}(k')$).

Theorem 1: The ranked layer (*Def. 9*) of L elementary classifiers Q_i with the decision rules $q_i(\mathbf{v}_i;\mathbf{x})$ (7) transforms the separable sets C_k (1) into linearly separable (4) sets D_k (10).

Proof: Let us assign the following parameter α_i to each ranked field R_i (19).

$$(\forall i \in \{1,...,L\}) \quad \alpha_i = \frac{1}{2^i} \tag{23}$$

The hyperplane $H(\mathbf{z}_k, \theta_k)$ (3) used for separation of the set D_k (10) from the sum $\cup D_i$ of the remaining sets D_i ($i \neq k$) can be defined by the weight vector $\mathbf{z}_k = [z_{k1},..., z_{kL}]^T$ with the following components z_{ki}

$$\begin{aligned}(\forall i \in \{1,...,L\}) \quad & \textit{if} \quad R_i \in C_k \quad \textit{then} \quad z_{ki} = \alpha_i \\ & \textit{and} \quad \textit{if} \quad R_i \notin C_k \quad \textit{then} \quad z_{ki} = -\alpha_i\end{aligned} \tag{24}$$

By direct computations we can verify the inequalities below.

$$\begin{aligned}(\exists k \in \{1,...,K\})(\forall \mathbf{q}_j(k) \in D_k) \quad & \mathbf{z}_k{}^T \mathbf{q}_j(k) > 0 \\ \textit{and} \quad (\forall \mathbf{q}_j(k) \in D_k, i \neq k) \quad & \mathbf{z}_k{}^T \mathbf{q}_j(i) < 0\end{aligned} \tag{25}$$

where $\mathbf{z}_k$ is the weight vector with the components z_{ki} (24). The inequalities (25) mean that the sets D_k (10) are linearly separable (4).

The considerations above are similar to the proof given in the paper [7]. The notions used in *Theorem 1* are illustrated by the below Figure.

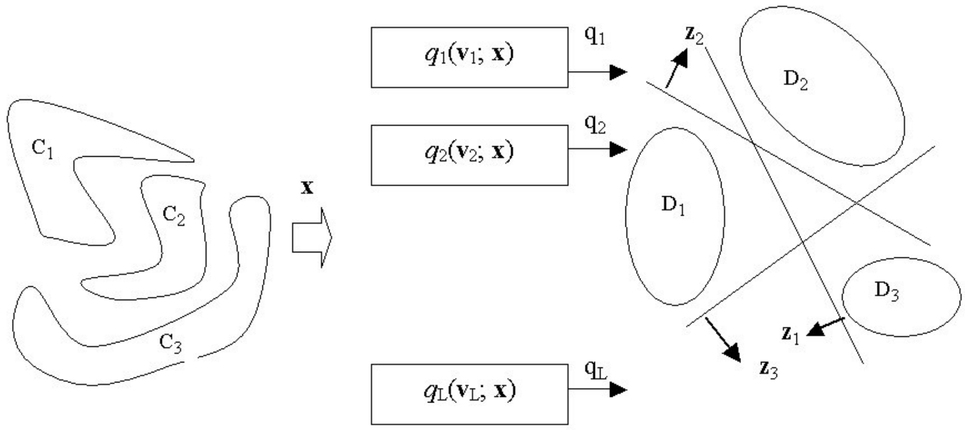

Fig. 1. Linearization (25) of three data sets C_k by the ranked layer of *L* elementary classifiers Q_i (7)

Linearization (26) of data sets C_k (1) by the ranked layer of L elementary classifiers Q_i has such consequence, that the second layer of K formal neurons $FN(\mathbf{w}_k, \theta_k)$ defined on the vectors $\mathbf{q}_j(k)$ (9) can separate exactly the sets D_k (10). As a consequence, each feature vector $\mathbf{x}_j(k)$ (1) can be correctly classified by the hierarchical network formed by such two layers.

6. Optimization of separable layers

A layer of L elementary classifiers Q_i (7) transforms each feature vector $\mathbf{x}_j(k)$ (1) into the *output vector* $\mathbf{q}_1 = [q_1,...,q_L]^T$ (9) with L binary components q_i ($q_i \in \{0,1\}$). Many feature vectors $\mathbf{x}_j(k)$ can be transformed into the same vector $\mathbf{q}_1$ ($(l \neq l') \Rightarrow \mathbf{q}_1(k) \neq \mathbf{q}_{1'}(k')$) in this manner. The set of such feature vectors $\mathbf{x}_j(k)$ is called as the l-th *activation field* S_l of the layer of elementary classifiers.

$$S_l = \{\mathbf{x}_j(k) : [q_1(\mathbf{v}_1, \mathbf{x}_j(k)),..., q_L(\mathbf{v}_L, \mathbf{x}_j(k))]^T = \mathbf{q}_l\} \tag{26}$$

where $(\forall l \neq l')$ $\mathbf{q}_1(k) \neq \mathbf{q}_{1'}(k)$.

Definition 12: The set S_l (26) will be called the *clear activation field* if all feature vectors $\mathbf{x}_j(k)$ (1) from this set ($\mathbf{x}_j(k) \in S_l$) belong to the same class ω_k. Similarly, the set S_l is the *mixed activation field* if it contains feature vectors $\mathbf{x}_j(k)$ (1) from different classes ω_k.

The field $S_l(k)$ and the output vector $\mathbf{q}_l(k)$ will be linked to the k-th class ω_k if and only if the most of the labeled feature vectors $\mathbf{x}_j(k)$ from the set S_l (26) is labeled to the class ω_k.

All feature vectors $\mathbf{x}_j(k)$ from the l-th activation field S_l are *aggregated* by the layer of elementary classifiers into one vector $\mathbf{q}_l$. In other words, the vector $\mathbf{q}_l$ *generalizes* all feature vectors $\mathbf{x}_j(k)$ from the field S_l (26). It can be expected that the layer of elementary classifiers with large and clear activation fields S_l (26)

will have a great *generalization power*. Such layer could be used also as a classifier with the following decision rule

$$\text{if} \quad \mathbf{x}_0 \in S_l(k) \quad \text{then} \quad \mathbf{x}_0 \in \omega_k \tag{27}$$

where $S_l(k)$ is such activation field (26) that most of the labeled feature vectors (1) from this field belong to the class ω_k.

A quality of the decision rule can be evaluated by the *error rate er* [9]. The classification error rate *er* is often evaluated as

$$\hat{er} = \frac{m_e}{m} \tag{28}$$

where m_e is the number of such feature vector $\mathbf{x}_j(k)$ from the sets C_k (1) which are wrongly allocated by the decision rule (27). The error rate evaluation (28) is positively biased (*optimistic bias*). The unbiased error rate *er* evaluations are based on such technique as cross-validation or on using testing sets [1].

Optimization problem I: To design such a layer of L of elementary classifiers Q_i (7) which will produce the decision rule (27) with the minimal *error rate er*.

Definition 13: A layer L of elementary classifiers Q_i (7) will be called *separable* if each feature vector $\mathbf{x}_j(k)$ from the sets C_k (1) belongs to some clear activation field S_l (26).

Optimization problem II: To design a separable layer of L of elementary classifiers Q_i (7) with minimal number L' of activation fields $S_l(k)$ or the output vectors $\mathbf{q}_l$ (26).

The minimal number L' of the activation fields S_l (26) can not be less than the number K of the classes ω_k ($L' \geq K$).

One can see, that a separable layer of elementary classifiers Q_i (7) with the decision rule (27) allocates correctly all feature vectors $\mathbf{x}_j(k)$ from the learning sets C_k (1). In this case, the estimator (28) of the error rate *er* is equal to zero. The classifiers which have error rate evaluation (28) on the sets C_k (1) equal to zero are often overfittning to these sets. As a consequence, such classifier can have a low generalisation power and the classification of new objects $\mathbf{x}$ might often be wrong. So, the classification (27) based only on the clear activation fields S_l (26) could be far from optimal. In order to improve the classification rule (26), the clear activation fields S_l (26) should be replaced by such fields S_l (26) which can be "slightly" mixed.

Definition 14: A layer of L elementary classifiers Q_i (7) will be called the ε - s*eparable,* if and only if the ratio m_e / m (28) of the wrongly classified feature vectors $\mathbf{x}_j(k)$ by the rule is no greater than ε ($m_e / m \le \varepsilon$), where ε is a positive parameter ($\varepsilon > 0$).

Optimization problem III: Design a ε-separable layer of L of elementary classifiers Q_i (7) with minimal number L' of activation fields $S_l(k)$ (26).

A separable layer of L elementary classifiers Q_i (*Def. 9*) can serve also in data aggregation. Let us define the *aggregation coefficient* η_a f such layer a in the following manner

$$\eta_a = \frac{m - m'}{m - K} \tag{29}$$

where m is the number of the feature vectors $\mathbf{x}_j(k)$ from the sets C_k (1), m' is the number of different output vectors $\mathbf{q}_l$ (26) from a separable layer, and K is the number of the classes ω_k or the learning sets C_k (1).

The minimal number m' of the output vectors $\mathbf{q}_l$ (26) from a separable layer is equal to K ($m' = K$). The aggregation coefficient η_a (29) takes the maximal value equal to one ($\eta_a = 1$) in this ideal situation. The aggregation coefficient η_a (29) of a layer of formal neurons $FN(\mathbf{w}_i, \theta_i)$ (5) can take the maximal value $\eta_a = 1$ if and only if the learning sets C_k (1) are linearly separable. The maximal value of the number m' is equal to m. There is no aggregation in this case and the aggregation coefficient η_a (29) takes the minimal value equal to 0 ($\eta_a = 0$). As a result.

$$0 \le \eta_a \le 1 \tag{30}$$

It can be noted that a solution of the *Optimization problem II* leads to the maximisation of the aggregation coefficient η_a (29).

In some cases, the above optimization problems can be solved through minimisation of the convex and piecewise linear (*CPL*) criterion functions [7]. We will pay particular attention to the perceptron criterion function (*CPL*). This function is linked to the beginning of the theory of neural networks.

7. Convex and piecewise linear criterion function (CPL)

Let us consider designing a separable layer of the formal neurons $FN(\mathbf{w}_i, \theta_i)$ (5) or the logical elements $LE(\mathbf{w}_i, \theta_i)$ (12). In this case, the designing procedure can be based on a sequence of minimisation of the convex and piecewise linear (*CPL*) criterion functions $\Psi_l(\mathbf{w}, \theta)$ ([3], [4]). The perceptron criterion function $\Psi_l(\mathbf{w}, \theta)$ belongs to the *CPL* family. It is easy to define the functions $\Psi_l(\mathbf{w}, \theta)$ by using the positive G_l^+ and the negative G_l^- sets of the feature vectors $\mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]^T$ (1).

$$G_l^+ = \{\mathbf{x}_j\} \quad j \in J_l^+ \quad and \quad G_l^- = \{\mathbf{x}_j\} \quad j \in J_l^- \tag{31}$$

Each element $\mathbf{x}_j$ of the set G_l^+ defines the positive penalty function $\varphi_j^+(\mathbf{w}, \theta)$.

$$\varphi_j^+(\mathbf{w}, \theta) = \begin{matrix} 1 - \mathbf{w}^T\mathbf{x}_j + \theta & if \quad \mathbf{w}^T\mathbf{x}_j - \theta \leq 1 \\ 0 & if \quad \mathbf{w}^T\mathbf{x}_j - \theta > 1 \end{matrix} \tag{32}$$

Similarly, each element $\mathbf{x}_j$ of the set G_l^- defines the negative penalty function $\varphi_j^-(\mathbf{w}, \theta)$.

$$\varphi_j^-(\mathbf{w}, \theta) = \begin{matrix} 1 + \mathbf{w}^T\mathbf{x}_j - \theta & if \quad \mathbf{w}^T\mathbf{x}_j - \theta \geq -1 \\ 0 & if \quad \mathbf{w}^T\mathbf{x}_j - \theta < -1 \end{matrix} \tag{33}$$

The penalty function $\varphi_j^+(\mathbf{w}, \theta)$ is aimed at positioning the vector $\mathbf{x}_j$ ($\mathbf{x}_j \in G_l^+$) on the positive side of the hyperplane $H(\mathbf{w}_k, \theta_k)$ (3). Similarly, the function $\varphi_j^-(\mathbf{w}, \theta)$ should set the vector $\mathbf{x}_j$ ($\mathbf{x}_j \in G_l^-$) on the negative side of this hyperplane.

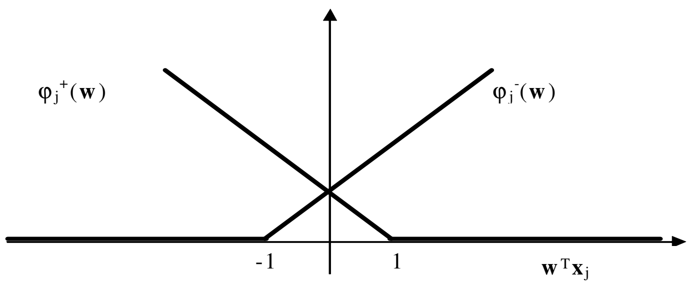

Fig. 2. The penalty functions $\varphi_j^+(w)$ (32) and $\varphi_j^-(w)$ (33).

The criterion function $\Psi_l(\mathbf{w},\theta)$ is the positively weighted sum of the penalty functions $\varphi_j^+(\mathbf{w},\theta)$ and $\varphi_j^-(\mathbf{w},\theta)$.

$$\Psi_l(\mathbf{w},\theta) = \sum_{j\in J_l^+} \alpha_j^+ \varphi_j^+(\mathbf{w},\theta) + \sum_{j\in J_l^-} \alpha_j^- \varphi_j^-(\mathbf{w},\theta) \qquad (34)$$

where α_j^+ ($\alpha_j^+ > 0$) and α_j^- ($\alpha_j^- > 0$) are the positive parameters (*prices*).

The criterion function $\Psi_l(\mathbf{w},\theta)$ belongs to the family of the convex and piecewise linear (*CPL*) criterion functions. Minimization of the function $\Psi_l(\mathbf{w},\theta)$ allows to find optimal parameters $(\mathbf{w}_l^*,\theta_l^*)$.

$$\Psi_l^* = \Psi_l(\mathbf{w}_l^*,\theta_l^*) = \min \Psi_l(\mathbf{w},\theta) > 0 \qquad (35)$$

The basis exchange algorithms which are similar to linear programming allow to find the minimum of the criterion function $\Psi_l(\mathbf{w},\theta)$ efficiently, even in the case of large, multidimensional data sets G_l^+ and G_l^- (29) [5].

It has been proved that the minimal value Ψ_l^* of the peceptron criterion function $\Psi_l(\mathbf{w},\theta)$ (32) is equal to zero ($\Psi_l^* = 0$) if and only if the positive G_l^+ and the negative G_l^- sets (29) are linearly separable (4). In this case, all elements $\mathbf{x}_j$ of the set G_l^+ (29) are located on the positive side of the hyperplane $H(\mathbf{w}_l^*,\theta_l^*)$ (3) and all elements $\mathbf{x}_j$ of the set G_l^- are located on the negative side:

$$\begin{array}{ll} & (\forall \mathbf{x}_j \in G_l^+) \quad (\mathbf{w}_l^*)^T \mathbf{x}_j > \theta_l^* \\ \textit{and} & (\forall \mathbf{x}_{j'} \in G_l^-) \quad (\mathbf{w}_l^*)^T \mathbf{x}_{j'} < \theta_l^* \end{array} \qquad (36)$$

If the sets G_l^+ and G_l^- (22) are not linearly separable (4), then $\Psi_l^* > 0$ and the inequalities (34) are fulfilled only partly, not by all, but by the majority of the elements $\mathbf{x}_j$ of the sets (22).

Minimization of the function $\Psi_l(\mathbf{w},\theta)$ (32) allows one to find optimal parameters $(\mathbf{w}_l^*,\theta_l^*)$ which define such hyperplane $H(\mathbf{w}_l^*,\theta_l^*)$ (3), which separates relatively well two sets G_l^+ and G_l^- (22). The parameters $(\mathbf{w}_l^*,\theta_l^*)$ can be also used in defining the l-th element $FN(\mathbf{w}_l^*,\theta_l^*)$ (5) of a neural layer.

The perceptron criterion function $\Psi_l(\mathbf{w},\theta)$ (32) can be used in designing separable layers of formal neurons $FN(\mathbf{w}_l,\theta_l)$ (5) both in accordance with the dipolar strategy described in Paragraph 4 as well as in accordance with the ranked strategy described in Paragraph 5. Specification of the criterion function $\Psi_l(\mathbf{w},\theta)$ (32) to particular strategy is achieved through an adequate choice of the sets G_l^+ and G_l^- (22) and the prices α_j^+ or α_j^- of the feature vectors $\mathbf{x}_j(k)$.

Designing separable layers of the formal neurons $FN(\mathbf{w}_l,\theta_l)$ (5) or the logical elements $LE(\mathbf{w}_i,\theta_i)$ (12) can be done in a sequential manner. During the l-th stage the l-th element $FN(\mathbf{w}_l^*,\theta_l^*)$ (5) or $LE(\mathbf{w}_i,\theta_i)$ (12) of the layer is designed through minimization of the criterion function $\Psi_l(\mathbf{w},\theta)$ (32).

Both the dipolar and the ranked strategy of separable layer designing can be optimised in accordance with the postulates described in Paragraph 6. In order to obtain a layer with large activation fields S_l (26) (*Optimization problem II*) the following postulate has been formulated in the framework of the sequential dipolar strategy:

"… First neuron should be designed in such a manner that its hyperplane divides the greatest number possible of mixed dipoles and a possibly low number of the clear dipoles. Second neuron should divide the greatest number possible of mixed dipoles undivided by the first neuron, and so on. The procedure is stopped after all mixed dipoles are divided…." [5]

Similar goal in the framework of the ranked strategy is realized through the postulate of large *r*anked fields R_i (19). These postulates are aimed at achieving a separable layer with a large generalization power. Such layer should allow for considerable data aggregation (29) or for classification rules (27) with a low error rate.

8. Concluding remarks

Designing separable layers from different types of elementary classifiers Q_i (7) was discussed in this paper. The dipolar and the ranked strategy of separable layers designing was described. The dipolar strategy allows for preserving separability (2) of the learning sets C_k (1) by the design layer of elementary classifiers Q_i Ranked layers have a fundamental property of linearization of learning sets. This means that the separable data sets C_k (1) are transformed by the ranked layer into linearly separable (4) sets D_k (12). A simplified representation of a classification problem can be reached as a result of such transformation. Linearization of data sets by the ranked layers could find important applications also in the methods originating from Support Vector Machines (*SVM*) [3]. Both the dipolar, as well as the ranked layers, can be used as a tool for separable data aggregation.

The deterministic version of the dipolar and the ranked strategies was discussed in this paper. The deterministic approach has a constraint in the form of data overfitting. It can be expected that the statistical approach towards designing ranked layers (e.g. *Def.* 8) combined with feature selection techniques will increase the chance of obtaining accurate classifiers with a large discriminative power.

The dipolar and the ranked strategies of designing separable layers of the formal neurons $FN(\mathbf{w}_l, \theta_l)$ (5) or the logical elements $LE(\mathbf{w}_l, \theta_l)$ (12) can be done in a sequential manner. The optimisation of the parameters $(\mathbf{w}_l, \theta_l)$ (5) or $(\mathbf{w}_l, \theta_l)$ (12) during the l-th stage of designing can be done through minimisation of the convex and piecewise linear (*CPL*) criterion functions $\Psi_l(\mathbf{w}, \theta)$ (32). The basis exchange algorithms which are similar to linear programming allow to find the minimum of the criterion functions $\Psi_l(\mathbf{w}, \theta)$ [5]. Designing separable layers from such elementary classifiers Q_i (7) which are based on the Euclidean balls $K_E(\mathbf{w}_i, \rho_i)$ (13) demand other types of algorithms. For example, the genetic algorithms can be used in the designing process.

References

[1] Duda, O.R., Heart, P.E., Stork D.G.: *Pattern Classification*, Wydanie drugie, zmienione, John Wiley & Sons, 2001.

[2] Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, Academic Press 1990.

[3] Vapnik, V.N.: *Statistical Learning Theory*, J. Wiley, New York 1998.
[4] Rosenblatt, F.: *Principles of neurodynamics*, Spartan Books, Washington 1962.
[5] Bobrowski, L.: *Piecewise-Linear Classifiers, Formal Neurons and separability of the Learning Sets*, Proceedings of ICPR'96, pp. 224-228, (13th International Conference on Pattern Recognition", August 25-29, 1996, Wienna, Austria).
[6] Bobrowski, L,: *Design of piecewise linear classifiers from formal neurons by some basis exchange technique*, Pattern Recognition, 24(9), pp. 863-870, 1991.
[7] Bobrowski, L.: *The ranked neuronal networks*, Biocybernetics and Biomedical Engineering, Vol. 12, No. 1-4, pp. 61-75, 1992.
[8] Bobrowisk, L.: *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych*, Politechnik Białostocka, 2005.

SEPAROWALNA AGREGACJA DANYCH W WARSTWACH KLASYFIKATORÓW ELEMENTARNYCH

Streszczenie: Cele eksploracji danych mogą być osiągnięte przy użyciu różnorodnych metod, takich jak teoria zbiorów rozmytych lub teoria zbiorów przybliżonych. Interesująca grupa metod eksploracji danych bazuje na minimalizacji wypukłych i odcinkowo-liniowych (CPL) funkcji kryterialnych. Metody te wywodzą się z teorii sieci neuropodobnych (wielowarstwowy perceptron). Do tej grupy mogą być także zaliczone silne obliczeniowo metody eksploracji danych bazujące na maszynach wektorów podpierających (SVM).
Hierarchiczne sieci neuronów formalnych lub wielowymiarowe drzewa decyzyjne mogą być zbudowane na podstawie zbiorów uczących poprzez minimalizację funkcji kryterialnych typu CPL dostosowanych do problemu klasyfikacji. Inny typ funkcji kryterialnych CPL może być użyty do projektowania wizualizacyjnych transformacji danych. Podstawą w omawianym podejściu CPL do projektowania narzędzi eksploracji danych jest separowalność transformowanych zbiorów uczących.

Słowa kluczowe: transformacje danych, agregacja danych, separowalne zbiory danych, klasyfikatory elementarne, wypukła i odcinkowo-liniowa (CPL) funkcja kryterialna

Artykuł zrealizowano w ramach projektu badawczego „Temporalna reprezentacja wiedzy i jej implementacja w informatycznych systemach wspomagania postępowania medycznego", nr 3 T11F 011 30.