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 \( O_j \ ( j = 1, \ldots, m ) \) is represented as the feature vector \( \mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]^T \) 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 \( x_{ij} \ ( x_{ij} \in \{0,1\} \text{ or } x_{ij} \in R ) \).

We assume that the database contains descriptions \( \mathbf{x}_j(k) \) of \( m \) objects \( O_j(k) \ ( j = 1, \ldots, m ) \) labelled according to their category (class) \( \omega_k \ ( k = 1, \ldots, K ) \). 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 \( \omega_k \)

\[
C_k = \{ \mathbf{x}_j(k) \} \quad ( j \in I_k )
\]
where $I_k$ is the set of indices $j$ of such feature vectors $x_j(k)$ from the class $a_k$ 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 ($C_k \cap C_{k'} = \emptyset$ for $k \neq k'$). It means that feature vectors $x_j(k)$ and $x_j(k')$ from different learning sets $C_k$ and $C_{k'}$ cannot be equal:

$$ \forall j \in I_k \quad \text{and} \quad \forall j' \in I_{k'} \quad x_j(k) \neq x_j(k') \quad (2)$$

We are also considering the separation of the sets $C_k$ (1) by the hyperplanes $H(w_k, \theta_k)$ in the feature space $F[n]$:

$$ H(w_k, \theta_k) = \{ x : w_k^T x = \theta_k \} \quad (3) $$

where $w_k = [w_{k1}, \ldots, w_{kn}]^T \in \mathbb{R}^n$ is the weight vector, $\theta_k \in \mathbb{R}$ is the threshold, and $w_k^T x$ 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 $H(w_k, \theta_k)$ (3) from the sum $\bigcup C_i \ (i \neq k)$ of the remaining sets $C_i$:

$$ \exists k \in \{1, \ldots, K\} \exists w_k, \theta_k \ \ (\forall x_j(k) \in C_k) \ w_k^T x_j(k) > \theta_k $$

and

$$ \forall x_j(k') \in C_{k'}, k' \neq k \quad w_k^T x_j(k') < \theta_k \quad (4) $$

In accordance with relation (4), all the vectors $x_j(k)$ belonging to the learning set $C_k$ are situated on the positive side ($w_k^T x_j(k) > \theta_k$) of the hyperplane $H(w_k, \theta_k)$ (3) and all the feature vectors $x_j(k')$ from the remaining sets $C_i$ are situated on the negative side ($w_k^T x_j(k') < \theta_k$) of this hyperplane.

The separation of data sets $C_k$ by the hyperplanes $H(w_k, \theta_k)$ (3) can by linked to data transformation by a layer of $K$ formal neurons $FN(w_k, \theta_k)$. The formal neuron $FN(w_k, \theta_k)$ is defined by the threshold decision rule $q(w_k, \theta_k; x)$:
\[ q = q(w_k, \theta_k; x) = \begin{cases} 1 & \text{if } w_k^T x \geq \theta_k \\ 0 & \text{if } w_k^T x < \theta_k \end{cases} \]  \quad (5)

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

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

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

\[ \text{if } (q(w_k, \theta_k; x) = 1) \text{ and } (\forall i \neq k)q(w_i, \theta_i; x) \text{ then } (x \in \omega_k) \]  \quad (6)

A vector \( x \) is allocated to the class \( \omega_k \) if only one neuron \( FN(w_k, \theta_k) \) 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 \( FN(w_i, \theta_i) \) with the rule \((6)\) can allocate properly all the feature vectors \( x_j(k) \) \((1)\).

3. **Elementary classifiers**

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

\[ q_i = q_i(v_i; x) \quad (i = 1, ..., L) \]  \quad (7)

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

The classifier \( Q_i \) is activated by the feature vectors \( x \) if and only if \( q_i(v_i; x) = 1 \). Formal neurons \( FN(w_k, \theta_k) \) \((5)\) can be used as the elementary classifiers \( Q_i \).
**Definition 3:** The activation field $S_i$ of the elementary classifier $Q_i = Q_i(v_i)$ is defined as the set of such feature vectors $x$, which activates ($q_i(v_i; x) = 1$) this classifier.

$$S_i = \{ x : q_i(v_i; x) = 1 \} \quad (8)$$

The layer of $L$ elementary classifiers $Q_i$ transforms each feature vector $x_j(k)$ from the sets $C_k (1)$ into the vector $q_j(k)$ with $L$ binary components $q_i = q_i(v_i; x_j(k))$.

$$q_j(k) = [q_i(v_i; x_j(k)), \ldots, q_L(v_L; x_j(k))]^T \quad (9)$$

where $v_i = [v_{i1}, \ldots, v_{in}]^T$ is a vector of parameters.

Vectors $q_j(k)$ 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 $q_j(k)$ than on the basis of the feature vectors $x_j(k)$. An example of the decision rule is given by (6). The transformed vectors $q_j(k)$ form the sets $D_k$:

$$D_k = \{ q_j(k) \} \quad (j \in I_k) \quad (10)$$

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 $x_j(k)$ (1) activates at least one elementary classifiers $Q_i$ (7) of the layer, and the number $m'$ of different vectors $q_j(k)$ (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 $q_j(k)$ (9) and a low dimensionality $L'$ of these vectors.

23
Few examples of the elementary classifiers \( Q_i \) are given below:

**Example 1:** Formal neurons \( FN(w_i, \theta_i) \) (5) can be treated as the elementary classifiers \( Q_i \) (8). In this case, the decision rule \( q_i = q_i(v_i; x_j(k)) \) (7) is based on the vector of parameters \( v_i \) given below:

\[
v_i = [w_i^T, \theta_i]^T
\]

(11)

The activation field \( S_i \) (8) of the formal neuron \( FN(w_i, \theta_i) \) (5) is the positive half-space defined by the hyperplane \( H(w_i, \theta_i) \) (3).

**Example 2:** The number \( n \) of inputs \( x_j \) to formal neuron \( FN(w_i, \theta_i) \) (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 \( v_i = [w_i, \theta_i]^T \) (11) with only two components \( w_i \) and \( \theta_i \). The decision rule \( q_i = q_i(v_i; x_j(k)) \) (8) can be reduced to the below form:

\[
\text{if } (w_i x_j(k) \geq \theta_i) \text{ then } (q_i(v_i; x_j(k)) = 1) \text{ else } (q_i(v_i; x_j(k)) = 0)
\]

(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 \( H(w_k, \theta_k) H(w_k, \theta_k) \) (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 \( K_E(w_i, \rho_i) \) centered in the point \( w_i \) and with the radius \( \rho_i \) in the feature space \( F[n] \):

\[
K_E(w_i, \rho_i) = \{ x : (x - w_i)^T (x - w_i) \leq \rho_i \}
\]

(13)

The vector of parameters \( v_i = [w_i^T, \rho_i]^T \) defines the decision rule \( q_i = q_i(v_i; x_j(k)) \) (7) in the below manner

\[
\text{if } ((x_j(k) - w_i)^T (x_j(k) - w_i) \leq \rho_i) \text{ then } (q_i(v_i; x_j(k)) = 1) \text{ else } (q_i(v_i; x_j(k)) = 0)
\]

(14)
The activation field $S_i (8)$ of such elementary classifier $Q_i (7)$ is the ball $K_E(w_i, \rho_i) (13)$.

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

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

The decision rule $q_i = q_i(v_j; x_j(k)) (7)$ has now the following form:

$$\text{if} \ (|x_1 - w_1| + ... + |x_n - w_n| \leq \rho_i) \ \text{then} \ (q_i(v_j; x_j(k)) = 1)$$

$$\text{else} \ (q_i(v_j; x_j(k)) = 0)$$ (16)

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

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

$$K_{L1}(w_i, \alpha, \rho_i) = \{ x : \alpha_i |x_1 - w_1| + ... + \alpha_m |x_n - w_n| \leq \rho_i \}$$ (17)

where $\alpha_i = [\alpha_{i1}, ..., \alpha_{im}]^T$ is the vector of features costs $\alpha_i$.

The decision rule $q_i = q_i(v_j; x_j(k)) (16)$ is generalised with the parameters $v_i = [w_i^T, \alpha_i^T, \rho_i]^T$ to:

$$\text{if} \ (|\alpha_i x_j(k) - w_1| + ... + |\alpha_m x_j(k) - w_n| \leq \rho_i)$$

$$\text{then} \ (q_i(v_j; x_j(k)) = 1) \ \text{else} \ (q_i(v_j; x_j(k)) = 0)$$ (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

Let 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 $(x_j(k), x_j(k'))$ $(x_j(k) \neq x_j(k'))$ constitutes a **mixed dipole** if and only if these vectors belong to different classes $\omega_k$ ($k \neq k'$). Similarly, a pair of different feature vectors from the same class $\omega_k$ constitutes the **clear dipole** $(x_j(k), x_j(k'))$.

Definition 6: The elementary classifier $Q_i$ (7) **separates** (divides) the dipole $(x_j(k), x_j(k'))$ if only one feature vector $x_j(k)$ or $x_j(k')$ from this pair activates this element ($q_i(v_i;x_j(k)) = 1$ and $q_i(v_i;x_j(k')) = 0$ or $q_i(v_i;x_j(k)) = 0$ and $q_i(v_i;x_j(k')) = 1$).

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 $(x_j(k), x_j(k'))$ by at least one elementary classifier $Q_i$ (7) of the layer.

The proof of similar result for layer of formal neurons $FN(w, \theta)$ (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 $x_j(k)$ (1), an additional postulate is introduced:

Classification postulate II: Each feature vector $x_j(k)$ (1) should activate ($q_i(v_i;x_j(k)) = 1$) 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 $Q_i$ ($i = 1, ..., L$) (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 $Q_i$ and $Q_l$ 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 $R_i (l = 1, ..., L)$ of the layer of $L$ elementary $Q_i$ classifiers is a set of such feature vectors $x_j(k)$ (1) which activate the $l$-th classifier $Q_l$ and do not activate any of the prior classifiers $Q_i$.

$$R_i = \{x_j(k) : q_i(v_jx_j(k)) = 1 \text{ and } (\forall i < l)q_i(v_jx_j(k)) = 0\} \quad (19)$$

**Definition 9:** The ranked field $R_i$ (19) is deterministically admissible if and only if it contains feature vectors $x_j(k)$ from only one learning set $C_k (1)$.

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

The fraction $f_i$ of elements $x_j(l)$ from non-dominant sets $C_i$ is defined by the expression below:

$$f_i = \frac{m'_i(k)}{m_i(k) + m'_i(k)} \quad (20)$$

where $m_i(k)$ is the number of elements $x_j(k)$ from the dominant set $C_k$ in the ranked field $R_i$ (19) and $m'_i(k)$ is the number of elements $x_j(l)$ in this field from all non-dominant sets $C_i (1)$ ($m_i(k) > m'_i(k)$).

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.

$$K \leq L \leq m \quad (21)$$
where \( K \) is the number of the learning sets \( C_k \) (1), and \( m \) is the number of feature vectors \( x_j(k) \) 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 \( x_j(l) \). 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 \( (\text{Def. 7}) \) ranked fields \( R_i (19) \) forms the ranked layer if and only if each feature vector \( x_j(k) \) 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 \( x_j(k) \) into the vector \( q_j(k) \) (9) with \( L \) binary components \( q_i = q_i(v_j; x_j(k)) \). 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).

\[
(k \neq k') \Rightarrow (\forall j \in I_k) \text{ and } (\forall j' \in I_{k'}) q_j(k) \neq q_{j'}(k')
\] (22)

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

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

\[28\]
Proof: Let us assign the following parameter $\alpha_i$ to each ranked field $R_i$ (19).

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

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

$$\forall i \in \{1,...,L\} \quad \text{if } R_i \in C_k \text{ then } z_{ki} = \alpha_i$$
$$\text{and if } R_i \notin C_k \text{ then } z_{ki} = -\alpha_i$$

By direct computations we can verify the inequalities below.

$$\exists k \in \{1,...,K\}(\forall j, (k) \in D_k) \quad z_k^T q_j(k) > 0$$
$$\text{and } (\forall j, (k) \in D_k, i \neq k) \quad z_k^T q_j(i) < 0$$

where $z_k$ 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 $FN(w_k, \theta_k)$ defined on the vectors $q_j(k)$ (9) can separate exactly the sets $D_k$ (10). As a consequence, each feature vector $x_j(k)$ (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 $x_j(k)$ (1) into the output vector $q_l = [q_{1i}, ..., q_{Li}]^T$ (9) with $L$ binary components $q_i$ ($q_i \in \{0,1\}$). Many feature vectors $x_j(k)$ can be transformed into the same vector $q_l$ ($(l \neq l') \Rightarrow q_l(k) \neq q_{l'}(k')$) in this manner. The set of such feature vectors $x_j(k)$ is called as the $l$-th activation field $S_l$ of the layer of elementary classifiers.

$$S_l = \{ x_j(k) : [q_{1i}(v_1, x_j(k)), ..., q_{Li}(v_L, x_j(k))]^T = q_i \} \quad (26)$$

where $l \neq l'$.

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

The field $S_j(k)$ and the output vector $q_j(k)$ will be linked to the $k$-th class $\omega_k$ if and only if the most of the labeled feature vectors $x_j(k)$ from the set $S_j$ (26) is labeled to the class $\omega_k$.

All feature vectors $x_j(k)$ from the $l$-th activation field $S_l$ are aggregated by the layer of elementary classifiers into one vector $q_l$. In other words, the vector $q_l$ generalizes all feature vectors $x_j(k)$ 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

\[ \text{if } x_0 \in S_i(k) \text{ then } x_0 \in \omega_k \]  \hspace{1cm} (27)

where \( S_i(k) \) is such activation field (26) that most of the labeled feature vectors (1) from this field belong to the class \( \omega_k \).

A quality of the decision rule can be evaluated by the error rate \( \text{er} \) [9]. The classification error rate \( \text{er} \) is often evaluated as

\[ \text{er} = \frac{m_e}{m} \]  \hspace{1cm} (28)

where \( m_e \) is the number of such feature vector \( x_j(k) \) 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 \( \text{er} \) 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 \( \text{er} \).

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

**Optimization problem II:** To design a separable layer of \( L \) of elementary classifiers \( Q_i(7) \) with minimal number \( L' \) of activation fields \( S_i(k) \) or the output vectors \( q_i \) (26).

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

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

**Definition 14:** A layer of \( L \) elementary classifiers \( Q_i \) (7) will be called the \( \varepsilon \)-separable, if and only if the ratio \( m_e / m \) (28) of the wrongly classified feature vectors \( x_j(k) \) by the rule is no greater than \( \varepsilon \) \((m_e / m \leq \varepsilon)\), where \( \varepsilon \) is a positive parameter \((\varepsilon > 0)\).

**Optimization problem III:** Design a \( \varepsilon \)-separable layer of \( L \) of elementary classifiers \( Q_i \) (7) with minimal number \( L' \) of activation fields \( S_j(k) \) (26).

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

\[
\eta_a = \frac{m - m'}{m - K}
\]

where \( m \) is the number of the feature vectors \( x_j(k) \) from the sets \( C_k \) (1), \( m' \) is the number of different output vectors \( q_i \) (26) from a separable layer, and \( K \) is the number of the classes \( \omega_k \) or the learning sets \( C_k \) (1).

The minimal number \( m' \) of the output vectors \( q_i \) (26) from a separable layer is equal to \( K \) \((m' = K)\). The aggregation coefficient \( \eta_a \) (29) takes the maximal value equal to one \((\eta_a = 1)\) in this ideal situation. The aggregation coefficient \( \eta_a \) (29) of a layer of formal neurons \( FN(w_i, \theta_i) \) (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 \( \eta_a \) (29) takes the minimal value equal to 0 \((\eta_a = 0)\). As a result,

\[
0 \leq \eta_a \leq 1
\]

It can be noted that a solution of the Optimization problem II leads to the maximisation of the aggregation coefficient \( \eta_a \) (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 $FN(w_j, \theta_j)$ (5) or the logical elements $LE(w_j, \theta_j)$ (12). In this case, the designing procedure can be based on a sequence of minimisation of the convex and piecewise linear (CPL) criterion functions $\Psi_j(w, \theta)$ ([3], [4]). The perceptron criterion function $\Psi_j(w, \theta)$ belongs to the CPL family. It is easy to define the functions $\Psi_j(w, \theta)$ by using the positive $G^+_i$ and the negative $G^-_i$ sets of the feature vectors $x_j = [x_{j1}, ..., x_{jn}]^T$ (1).

$$G^+_i = \{x_j \} \quad j \in J^+_i \quad \text{and} \quad G^-_i = \{x_j \} \quad j \in J^-_i$$

(31)

Each element $x_j$ of the set $G^+_i$ defines the positive penalty function $\phi^+_j(w, \theta)$.

$$\phi^+_j(w, \theta) = \begin{cases} 
1 - w^T x_j + \theta & \text{if} \quad w^T x_j - \theta \leq 1 \\
0 & \text{if} \quad w^T x_j - \theta > 1
\end{cases}$$

(32)

Similarly, each element $x_j$ of the set $G^-_i$ defines the negative penalty function $\phi^-_j(w, \theta)$.

$$\phi^-_j(w, \theta) = \begin{cases} 
1 + w^T x_j - \theta & \text{if} \quad w^T x_j - \theta \geq -1 \\
0 & \text{if} \quad w^T x_j - \theta < -1
\end{cases}$$

(33)

The penalty function $\phi^+_j(w, \theta)$ is aimed at positioning the vector $x_j$ ($x_j \in G^+_i$) on the positive side of the hyperplane $H(w_k, \theta_k)$ (3). Similarly, the function $\phi^-_j(w, \theta)$ should set the vector $x_j$ ($x_j \in G^-_i$) on the negative side of this hyperplane.
The criterion function $\Psi_i(w, \theta)$ is the positively weighted sum of the penalty functions $\varphi_j^+(w, \theta)$ and $\varphi_j^-(w, \theta)$.

$$\Psi_i(w, \theta) = \sum_{j \in I^+_i} \alpha_j^+ \varphi_j^+(w, \theta) + \sum_{j \in I^-_i} \alpha_j^- \varphi_j^-(w, \theta)$$  \hspace{1cm} (34)

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

The criterion function $\Psi_i(w, \theta)$ belongs to the family of the convex and piecewise linear (CPL) criterion functions. Minimization of the function $\Psi_i(w, \theta)$ allows to find optimal parameters $(w_i^+, \theta_i^+)$.

$$\Psi_i^* = \Psi_i(w_i^+, \theta_i^+) = \min \Psi_i(w, \theta) > 0$$  \hspace{1cm} (35)

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

It has been proved that the minimal value $\Psi_i^*$ of the perceptron criterion function $\Psi_i(w, \theta)$ (32) is equal to zero ($\Psi_i^* = 0$) if and only if the positive $G_i^+$ and the negative $G_i^-$ sets (29) are linearly separable (4). In this case, all elements $x_j$ of the set $G_i^+$ (29) are located on the positive side of the hyperplane $H(w_i^+, \theta_i^+)$ (3) and all elements $x_j$ of the set $G_i^-$ are located on the negative side:

$$\left( \forall x_j \in G_i^+ \right) (w_i^+)^T x_j > \theta_i^+$$

and $$\left( \forall x_j \in G_i^- \right) (w_i^-)^T x_j < \theta_i^-$$  \hspace{1cm} (36)
If the sets $G_i^+$ and $G_i^-$ (22) are not linearly separable (4), then $\Psi_i^* > 0$ and the inequalities (34) are fulfilled only partly, not by all, but by the majority of the elements $x_j$ of the sets (22).

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

The perceptron criterion function $\Psi_i(w, \theta)$ (32) can be used in designing separable layers of formal neurons $FN(w_j, \theta_j)$ (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 $\Psi_i(w, \theta)$ (32) to particular strategy is achieved through an adequate choice of the sets $G_i^+$ and $G_i^-$ (22) and the prices $\alpha^+_j$ or $\alpha^-_j$ of the feature vectors $x_j(k)$.

Designing separable layers of the formal neurons $FN(w_i, \theta_i)$ (5) or the logical elements $LE(w_j, \theta_j)$ (12) can be done in a sequential manner. During the $l$-th stage the $l$-th element $FN(w_j, \theta_j)$ (5) or $LE(w_j, \theta_j)$ (12) of the layer is designed through minimization of the criterion function $\Psi_i(w, \theta)$ (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 $S_i$ (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 ranked 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 \( FN(w_j, \theta_j) \) (5) or the logical elements \( LE(w_j, \theta_j) \) (12) can be done in a sequential manner. The optimisation of the parameters \( (w_j, \theta_j) \) (5) or \( (w_j, \theta_j) \) (12) during the \( l \)-th stage of designing can be done through minimisation of the convex and piecewise linear (CPL) criterion functions \( \Psi_l(w, \theta) \) (32). The basis exchange algorithms which are similar to linear programming allow to find the minimum of the criterion functions \( \Psi_l(w, \theta) \) [5]. Designing separable layers from such elementary classifiers \( Q_i \) (7) which are based on the Euclidean balls \( K_{\rho_i} \) (13) demand other types of algorithms. For example, the genetic algorithms can be used in the designing process.

References

Separable data aggregation by layers of elementary classifiers


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. Interesującą grupą 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 zaliczane 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.