Activity of the McLeod Instistitute of Simulation Sciences (MISS Center during accademic year 2005/2006 at the Computer Science Faculty Bialystok Technical University (BTU) in Poland

The MISS Center at the BTU Computer Science Faculty includes actually the following researchers:

- Prof. Leon Bobrowski (dean of the CS Faculty)
- Prof. Ralph Huntsinger
- Dr. Zenon Sonowski (vice-dean of the CS Faculty)
- Dr. Walenty Oniszczuk

The Computer Science (CS) Faculty at BTU obtained recently the title of the Center of Excellence in the field of information society and knowledge based economy. The MISS Center activity at the CS Faculty is coordinated with the activity of the Polish Society of Computer Simulation (PSCS). Prof. Bobrowski is actually the president of the PSCS and Dr. Sonowski is the treasure of this Society. Among others, the MISS has been represented at PSCS research workshops.

The research activites of the MISS Center at the CS Faculty includes th
following topics:

- modeling and simulation of finite-source systems
- simulation of networks with blocking
- investigation and modeling congestion problems in modeling
- data exploration methods originated from bionics concepts
- inference models based on fuzzy methodology

Among CS Faculty research projects the following ones are related to the
MISS Center activity:

- *Mathematical (analytical) and simulation models of information system with priority scheduling and blocking*. (directed by Dr. Oniszczuk)

Knowledge exploration in data bases with using of bionics concepts (directed by Prof. Bobrowski)

- *Temporal knowledge representation in computer support of medical diagnosis* (directed by Prof. Bobrowski)

Educational activities at the CS Faculty includes the following university c

- *Mathematical modeling and its application in informatics*

- *Simulation methodology*

- *Continuous Systems Simulation*

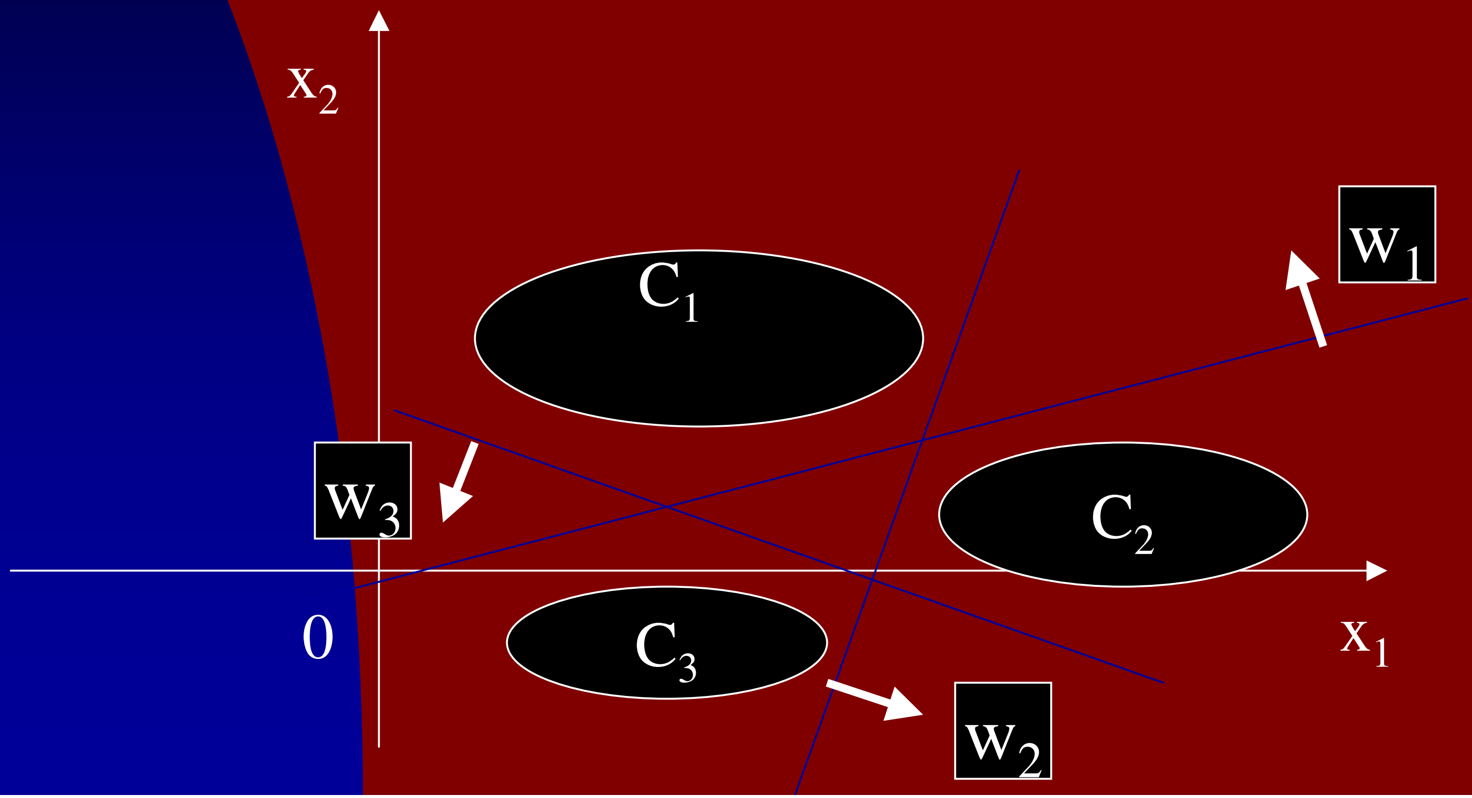
$\Phi(w)$ $\Phi(w)$

Thank you for your attention.

Separability of the learning sets C_k

Linearly separable learning sets C_k

Each data set C_k can be separated from the sum (*union*) of the remaining sets C_l by some hyperplane $H(\mathbf{w}_k,\theta_k) = \{\mathbf{x}: \langle\mathbf{w}_k,\mathbf{x}\rangle = \theta_k\}$.



From *linear independence* to *linear separability*

Augmented vectors $\mathbf{y}_j[n+1] = [\mathbf{x}_j[n]^T, 1]^T$ are linearly independent ($j = 1, \ldots\ldots, m$)

$$\Rightarrow$$

The learning sets C_k constituted by *m feature* vectors $\mathbf{x}_j[n]$ are linearly separable

Dipoles in the learning sets C_k

Definition 2: The pair $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k)\}$ ($j<j'$) of the feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k)$ constitutes the *pure dipole*, if these vectors belong to the same learning sets C_k. Similarly, two feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k')$ from different learning sets C_k and $C_{k'}$, constitute the *mixed dipole* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$.

If the Euclidean distance is used, then the length $\delta_x(j,j')$ of the dipole $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$ is defined by:

$$\delta_x(j,j') = (\mathbf{x}_j(k) - \mathbf{x}_{j'}(k'))^T(\mathbf{x}_j(k) - \mathbf{x}_{j'}(k'))^{1/2}$$

Separability postulates in designing data transformations

Examples:

I. The transformation $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ *should shorten clear dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k)\}$ *and lengthen mixed dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$.

II. The linear transformation $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ *should shorten clear dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k)\}$ *below the marigin* δ^- *and lengthen mixed dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$ *upper the marigin* δ^+.

III. The linear transformation $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ *should shorten clear dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k)\}$ *with the the length* $\delta_x(j,j')$ *less than the marigin* δ^- *and lengthen mixed dipoles* $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$ *with the the length* $\delta_x(j,j')$ *greater than the margin* δ^+.

SEPARABLE TRANSFORMATIONS OF THE LEARNIG SETS C_k

The parameters ($\mathbf{w}_k^*$,θ_k^*) ($k = 1,\ldots,n$) defining the separable transformation $y_k = (\mathbf{w}_k^*)^T \mathbf{x}$ of the learning sets C_k can be found through minimisation of the convex and picewise linear (*CPL*) criterion functions.

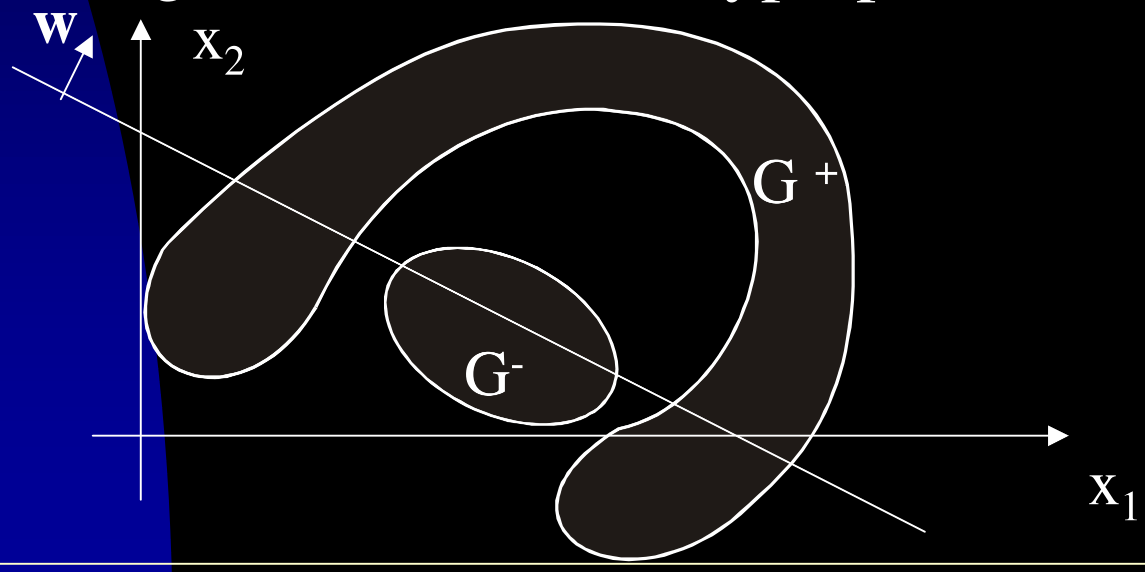
The *perceptron criterion function* belongs to the *CPL* family and could be used for this purpose.

Separation of the data sets G^+ and G^- by the hyperplane $H(\mathbf{w},\theta)$

- $H(\mathbf{w},\theta)$ - hyperplane in the feature space:

$$H(\mathbf{w},\theta) = \{\mathbf{x}: \langle \mathbf{w},\mathbf{x} \rangle = \theta\}$$

- Elements of the set G^+ should be situated on the *positive side* and elements of the set G^- should be on the *negative side* of the hyperplane $H(\mathbf{w},\theta)$.



Linearly separable data sets G^+ and G^-

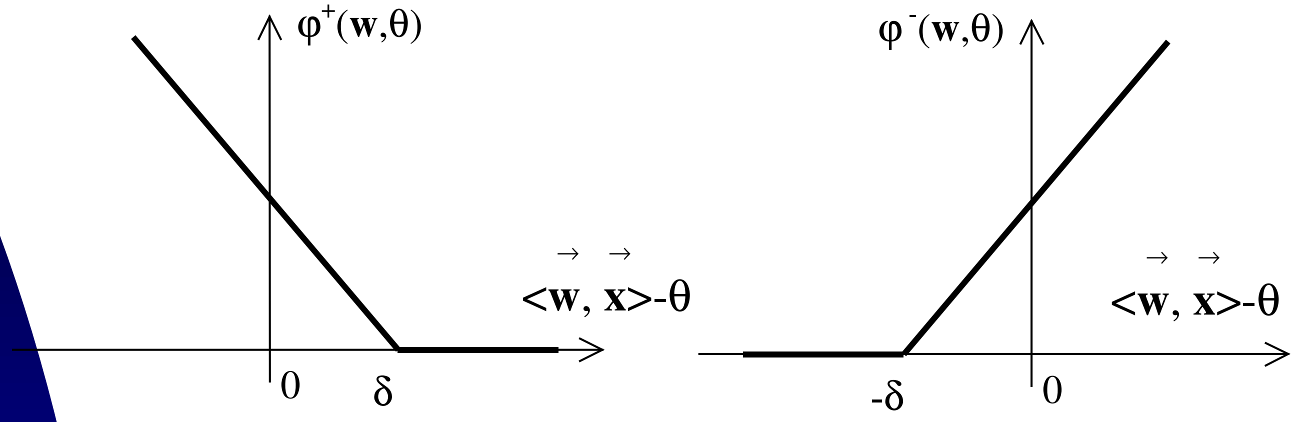
- The data sets G^+ and G^- are *linearly separable*, if and only if there exist such parameters $\mathbf{w}^*$ and θ^* that all elements $\mathbf{x}_j(k)$ of these sets are properly allocated:

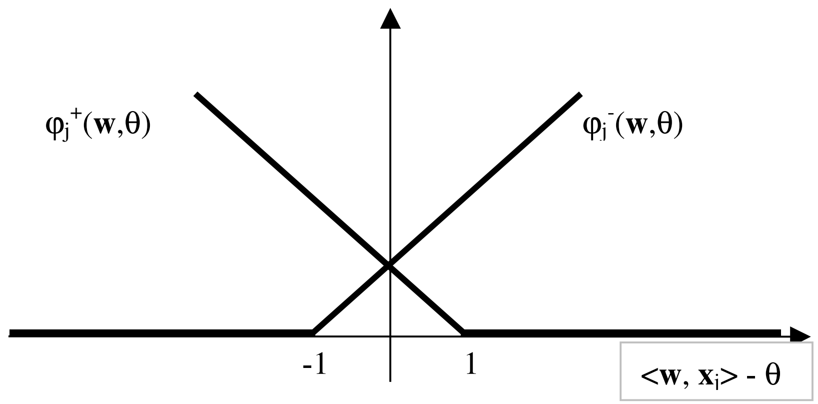
- $$(\exists \mathbf{w}^*, \theta^*)\ (\forall\ \mathbf{x}_j(k) \in G^+)\quad \langle \mathbf{w}^*, \mathbf{x}_j(k) \rangle > \theta^*$$
- *and* $$(\forall\ \mathbf{x}_j(k) \in G^-)\quad \langle \mathbf{w}^*, \mathbf{x}_j(k) \rangle < \theta^*$$

Perceptron penalty functions $\varphi_j^+(\mathbf{w},\theta)$ and $\varphi_j^-(\mathbf{w},\theta)$

- $(\forall\ \mathbf{x}_j(k) \in G^+)$
- $\varphi_j^+(w,\theta) = \begin{cases} \delta_j + \theta - \langle \mathbf{w},\mathbf{x}_j(k)\rangle & \textit{if}\ \ \langle \mathbf{w},\mathbf{x}_j(k)\rangle - \theta < \delta_j \\ 0 & \textit{if}\ \ \langle \mathbf{w},\mathbf{x}_j(k)\rangle - \theta \geq \delta_j \end{cases}$
- and $(\forall\ \mathbf{x}_j(k) \in G^-)$
- $\varphi_j^-(w,\theta) = \begin{cases} \delta_j - \theta + \langle \mathbf{w},\mathbf{x}_j(k)\rangle & \textit{if}\ \ \langle \mathbf{w},\mathbf{x}_j(k)\rangle - \theta > -\delta_j \\ 0 & \textit{if}\ \ \langle \mathbf{w}, \mathbf{x}_j(k)\rangle - \theta \leq -\delta_j \end{cases}$
- where δ_j - is the *margin* $(\delta_j \geq 0)$

Perceptron penalty functions





The penalty functions $\varphi_j^+(\mathbf{w},\theta)$ and $\varphi_j^-(\mathbf{w},\theta)$ with the margin $\delta_j = 1$

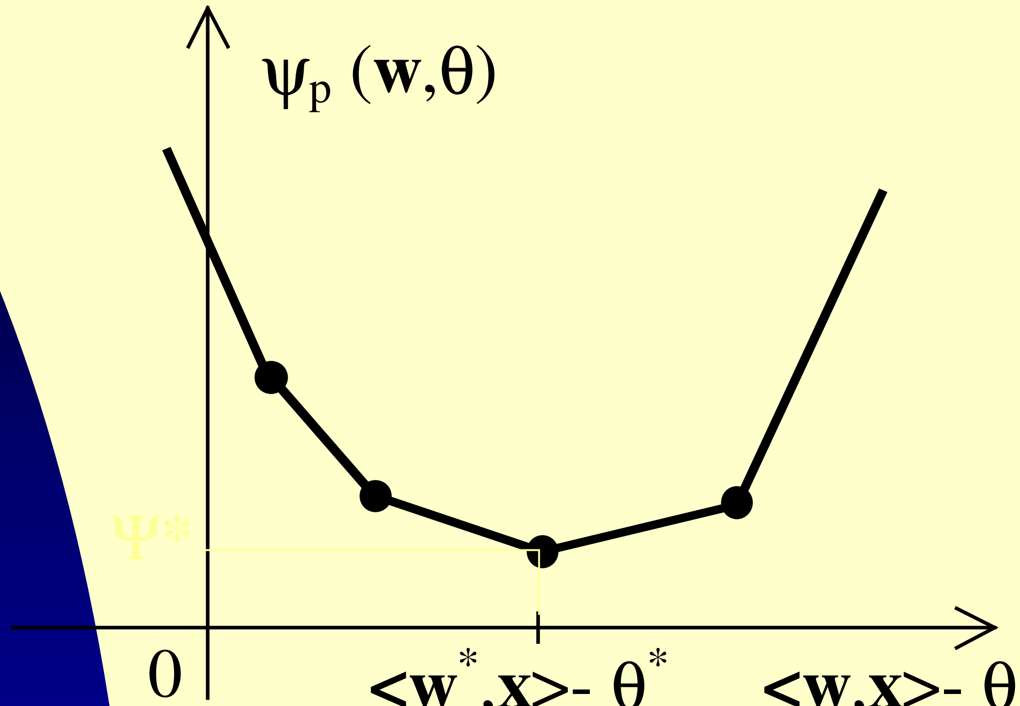
Perceptron criterion function $\Psi_p(\mathbf{w},\theta)$

$$\Psi_p(\mathbf{w},\theta) = \sum_{\mathbf{x}j \in G+} \alpha_j \varphi_j^{+}(\mathbf{w},\theta) + \sum_{\mathbf{x}j \in G-} \alpha_j \varphi_j^{-}(\mathbf{w},\theta) \qquad (7)$$

- where the nonnegative parameters α_j determine relative importance (*price*) of particular feature vectors $\mathbf{x}_j(k)$.

- $\Psi_p(\mathbf{w},\theta)$ is the convex and piecewise linear (*CPL*) function

Perceptron criterion function $\Psi_p(w,\theta)$



$\Psi_p(w,\theta)$ is the convex and piecewise linear (*CPL*) function

Perceptron criterion function $\Psi_p(\mathbf{w},\theta)$

- Minimisation task:
- $$\Psi_p{}^* = \Psi_p(\mathbf{w}^*,\theta^*) = \mathit{min}\, \Psi_p(\mathbf{w},\theta)$$
- The *basis exchange algorithms*, similar to *linear programming*, allow to find in an efficient manner the optimal parameters ($\mathbf{w}^*$, θ^*) and the minimal value $\Psi_p{}^*$ of the criterion function $\Psi_p(\mathbf{w},\theta)$, even in the case of large, multidimensional data sets G^+ and G^-.

Hepar - system's characteristics

The system *Hepar* comprises a clinical database and a shell of procedures that aim at the data analysis and the support of diagnosis.

The database of the system contains the results of medical findings of more than 800 patients from one of gastroenterological clinics. Each patient is described by about 200 symptoms and laboratory tests x_i. The patients from this database have been classified (labelled) by clinicians into about 25 liver diseases. Medical classification has been based mainly on the liver biopsy (invasive examination).

The support of diagnosis in the system is based on the comparison of a new patient (without the biopsy) with similar cases from the database. Graphical representation of the data on *diagnostic maps* is particularly important in the *Hepar* system.

Designing of diagnostic maps in the system *Hepar*

- *M*ap designing begins in the system with user's (medical doctor) definition of the diagnostic problem. Such definition is based on two types of declarations:

- *i. – choice of the differentiated (separated) classes* ω_k
- *ii.* - *choice of an initial set of diagnostic tests (features)*
- x_i *used for separation of the classes* ω_k

Designing of a diagnostic map for four groups of patients Γ_k

- Four groups of patients Γ_k (k =1,2,3,4) are defined by the user and are supposed to be allocated in four quarters of the map in the following manner:

- Γ_1 – *the upper-right quarter*
- Γ_2 – *the lower-right quarter*
- Γ_3 – *the lower-left quarter*
- Γ_4 – *the upper-left quarter*

Designing of a diagnostic map for four groups of patients Γ_k (cont.)

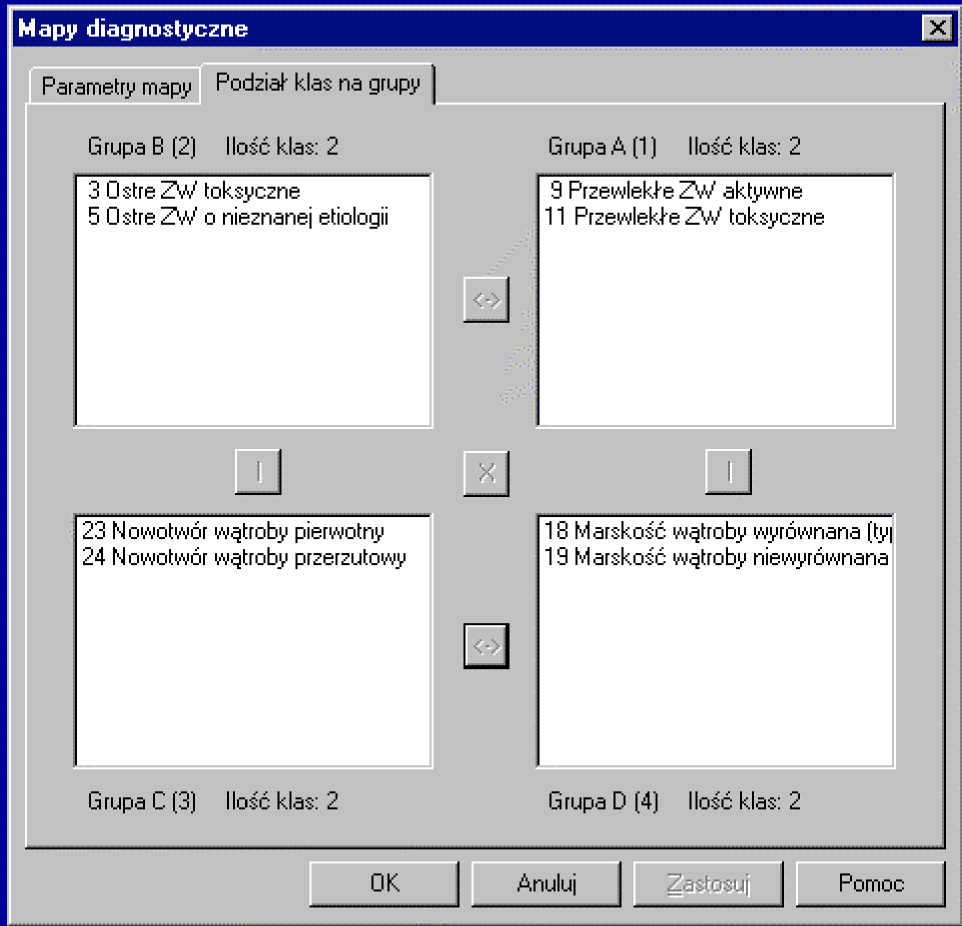
- The groups Γ_k are translated by the system into two pairs of the sets G_i^+ and G_i^- (i = 1, 2):

- $G_1^+ = \Gamma_1 \cup \Gamma_2$ ***and*** $G_1^- = \Gamma_3 \cup \Gamma_4$
- $G_2^+ = \Gamma_1 \cup \Gamma_4$ ***and*** $G_2^- = \Gamma_2 \cup \Gamma_3$

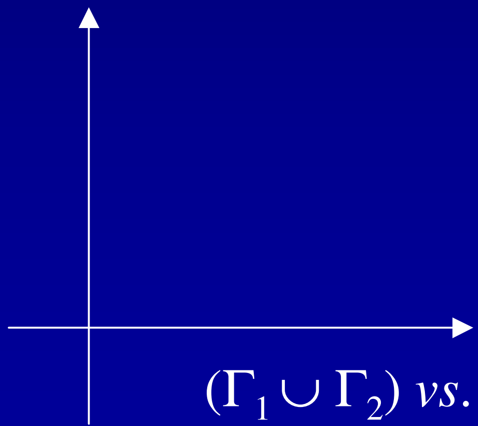
- The data sets G_1^+ and G_1^- are used in the definition of the criterion function $\Psi_1(\mathbf{w},\theta)$ linked to the first axis of the map $((\Gamma_1 \cup \Gamma_2)$ *vs.* $(\Gamma_3 \cup \Gamma_4))$. Similarly, G_2^+ and G_2^- are used in the function $\Psi_2(\mathbf{w},\theta)$ linked to the second axis of the map $((\Gamma_1 \cup \Gamma_4)$ *vs.* $(\Gamma_2 \cup \Gamma_3))$.

Designing of diagnostic maps in the system *Hepar*

- *Example*:



$(\Gamma_1 \cup \Gamma_4)$ *vs.* $(\Gamma_2 \cup \Gamma_3)$

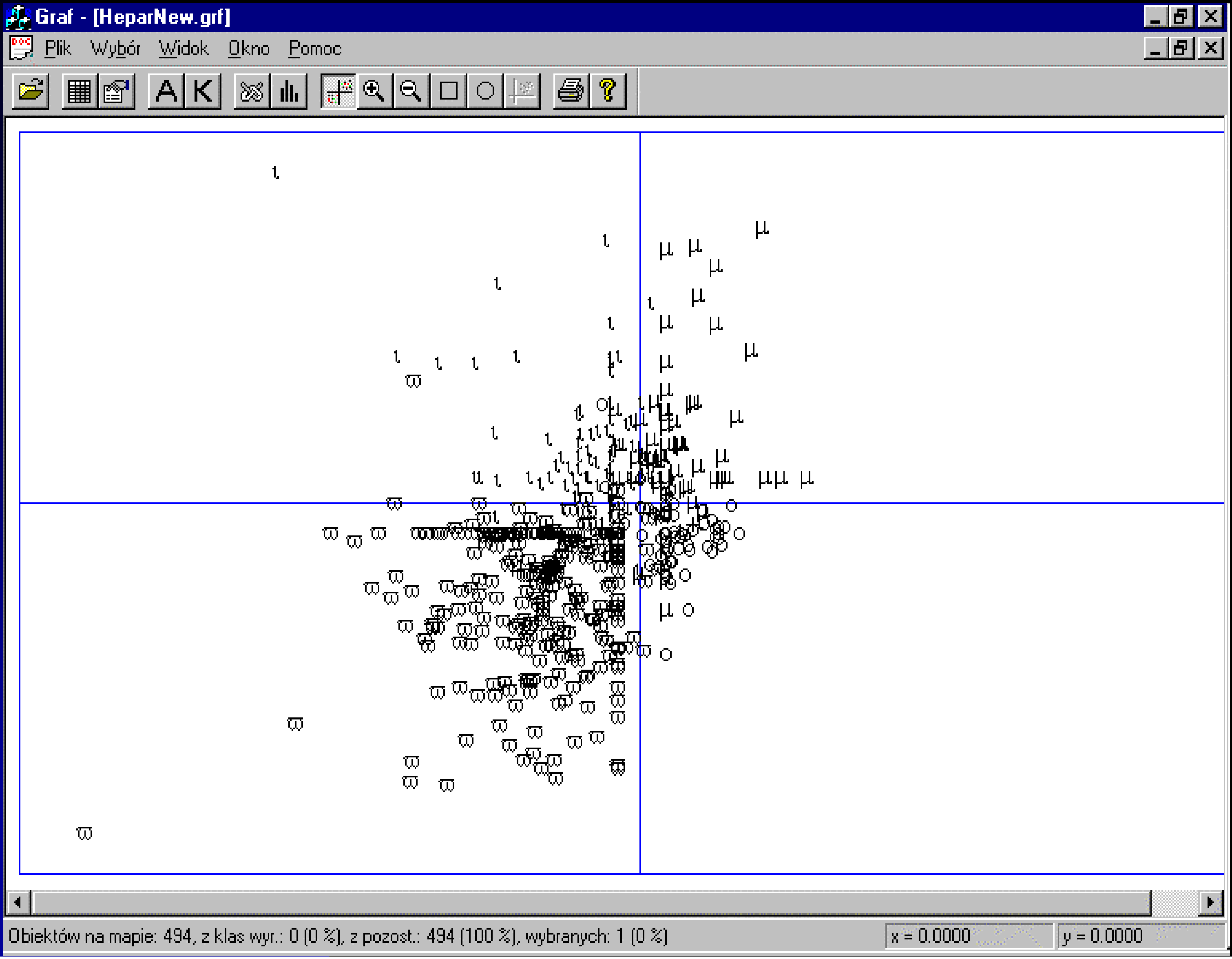


$(\Gamma_1 \cup \Gamma_2)$ *vs.* $(\Gamma_3 \cup \Gamma_4)$

Designing of a diagnostic map for four groups of patients Γ_k (cont.)

- Four learning sets C_k has been chosen to be a subject of differentiation by using of 106 symptoms and laboratory tests x_i:
- Γ_1 – Steatosis hepatis – 67 patients
- Γ_2 – Hiperbilirubinemia functionalis – 56 patients
- Γ_3 – Cirrhosis hepatis billiaris primaria – 272 patients
- Γ_4 – Hepatitis chronica activa – 91 patients

Eaxample of diagnostic map of the system *Hepar*



- - It is possible to improve the nearest neighbors decision rules (*CBR* cycle) through linear scaling transformations and dimensionality reduction.
- - Visualizing transformations of data sets can give not only a new insight into a structure of the diagnostic problem, but could result also in improving of decision rules correctness.

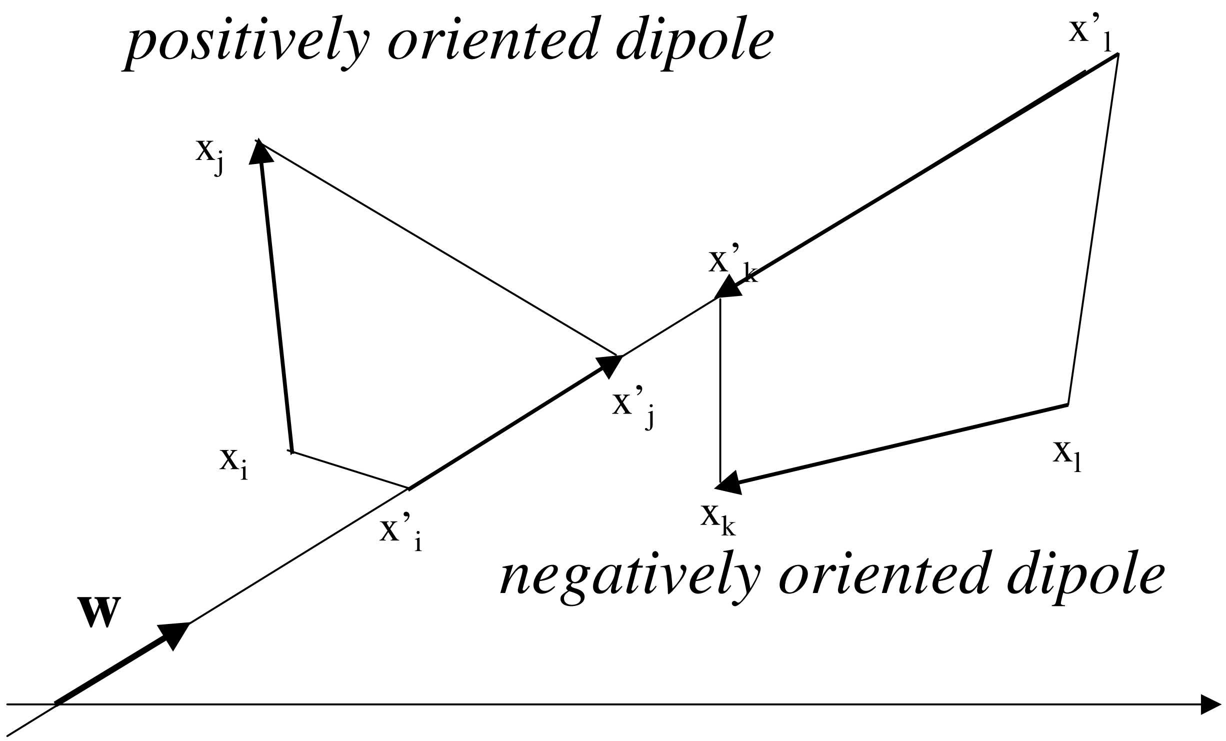
Designing the linear transformations based on the dipolar separabilty postulate (cont.)

The separability postulate can be reinforced by the *CPL* penalty functions $\varphi_{jj'}^{+}(\mathbf{w})$ and $\varphi_{jj'}^{-}(\mathbf{w})$ defined on the *differences* $\mathbf{r}_{jj'}$ of the feature vectors constituting particular dipoles $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$:

- $\mathbf{r}_{jj'} = \mathbf{x}_{j'}(k') - \mathbf{x}_j(k)$ (11)

- Similarly as the feature vectors $\mathbf{x}_j(k)$, each difference $\mathbf{r}_{jj'}$ could be allocated to the positive R^+ or/and to the negative set R^- , The set R^+ containes *positively oriented* dipoles and the set R^- *negatively* oriented dipoles $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k')\}$.

Geometrical representation of dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$, which are *positively* or *negatively* oriented



Geometrical representation of dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$, which are *positively* or *negatively* oriented