4

Induction of Similarity Measures for Case Based Reasoning Through Separable Data Transformations*

L. Bobrowski[1,2] and M. Topczewska[1]

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

Summary. Inducing similarity measures for the *case based reasoning* scheme through separable data transformations is considered in this chapter. Particular attention is paid to linear transformations of multidimensional data on visualising planes. Separable linear transformations are based both on solutions of eignevalue problems used in the *principal componet analysis* or in the *discriminant analysis* as well as on minimization of the convex and piecewise linear (CPL) criterion functions. The *perceptron* and the *differential* criterion functions belong among others to the CPL family. Such functions give possibility for flexible and efficient designing separable transformations of data sets.

4.1 Introduction

Decision support systems are often based on the case based reasoning (CBR) method [1]. An essential part of the CBR scheme is a search for such records in a database which are most similar to the case actually analysed [1]. Such a paradigm is also used in the nearest neighbours (K-NN) technique developed in the framework of the pattern recognition [2,3]. One of the central problems during implementation of the CBR or the K-NN scheme is the choice of a similarity measure or the distance function between the database records [4]. The quality of the decision support rules can be improved by adjusting the similarity measures or adequately tailoring the distance functions [5].

Here, we analyse possibilities of applying linear transformations of reference data sets for inducing similarity measures and diagnosis support rules from the learning sets. Particular attention is paid to linear transformations of multidimensional data on visualising planes. Designing linear transformation scheme that results from separability postulates is considered [5, 6]. The

*This work was partially supported by the grant W/WI/1/2005 from the Bialystok University of Technology, the KBN grant 3T11F01130 and by the grant 16/St/2007 from the Institute of Biocybernetics and Biomedical Engineering PAS.

L. Bobrowski and M. Topczewska: *Induction of Similarity Measures for Case Based Reasoning Through Separable Data Transformations*, Studies in Computational Intelligence (SCI) **73**, 127–148 (2008)
www.springerlink.com

separability postulates are reinforced through minimization of the convex and piecewise linear (CPL) criterion functions. The basis exchange algorithms, similar to linear programming, allow one to find a minimum of the CPL criterion functions efficiently, even in the case of large, multidimensional data sets [7].

4.2 Separabilty of Reference Sets

Let us assume that object descriptions stored in a database are represented as the so called feature vectors $\mathbf{x} = [\mathrm{x_1}, \ldots\ldots, \mathrm{x_n}]^{\mathrm{T}}$, or as points in the n-dimensional feature space $\boldsymbol{F}[n]$ [3]. The components $\mathrm{x_i}$ of vectors $\mathbf{x}$ are numerical results of various examinations of a given object. The feature vectors $\mathbf{x}$ can be of the mixed, qualitative–quantitative type because their components can be both real numbers ($\mathrm{x_i} \in \mathrm{R}$) as well as binary ones ($\mathrm{x_i} \in \{0, 1\}$).

We assume that a database contains descriptions of m objects $\mathbf{x_j}(k)$ ($j = 1, \ldots\ldots, m$) labelled in accordance with their class (*category*) $\boldsymbol{\omega}_{\mathrm{k}}(k = 1, \ldots, K')$. The labelling of the feature vectors should be done in accordance with an additional knowledge about particular decision support problem. For example, a clinical database contains the descriptions of m patients $\mathbf{x}_{\mathrm{j}}(k)(j = 1, \ldots\ldots, m)$ labelled in accordance with their clinical diagnosis $\boldsymbol{\omega}_{\mathrm{k}}$. The reference (learning) set $\mathrm{C_k}$ contains m_{k} labelled feature vectors $\mathbf{x}_{\mathrm{j}}(k)$ (*precedents*) related to the kth class $\boldsymbol{\omega}_{\mathrm{k}}$.

$$\mathrm{C_k} = \{\mathbf{x}_{\mathrm{j}}(k)\} \quad (j \in I_{\mathrm{k}}) \tag{4.1}$$

where I_{k} is the set of indices j of m_{k} feature vectors $\mathbf{x}_{\mathrm{j}}(k)$ belonging to the class $\boldsymbol{\omega}_{\mathrm{k}}$.

Definition 1. *The learning sets* $\mathrm{C_k}$ *(4.1) are separable in the feature space* $\boldsymbol{F}[n]$ *if they are disjoined in this space. It means that each of the feature vectors* $\mathbf{x}_{\mathrm{j}}$ *belongs to only one set* $\mathrm{C_k}$*:*

$$(\forall \boldsymbol{x}_j(k) \in C_k) \;\; \textit{and} \;\; (\forall \boldsymbol{x}_{j'}(k') \in C_{k'},\; k \neq k') \quad \boldsymbol{x}_{j'}(k') \neq \boldsymbol{x}_j(k) \tag{4.2}$$

In accordance with Definition 1, the feature vectors $\mathbf{x}_{\mathrm{j}}(k)$ *and* $\mathbf{x}_{\mathrm{j}'}(k')$ *from different reference sets* $\mathrm{C_k}$ *and* $\mathrm{C_{k'}}$ *cannot be equal.*

We are also considering separation of the sets $\mathrm{C_k}$ (4.1) by the hyperplanes $H(\mathbf{w}_{\mathrm{k}}, \boldsymbol{\theta}_{\mathrm{k}})$ in the feature space $\boldsymbol{F}[n]$

$$H(\boldsymbol{w}_k, \theta_k) = \left\{\boldsymbol{x} : \;\; \boldsymbol{w}_k^T \boldsymbol{x} = \theta_k\right\} \tag{4.3}$$

where $\mathbf{w}_{\mathrm{k}} = [\mathrm{w_{k1}}, \ldots, \mathrm{w_{kn}}]^{\mathrm{T}} \in \mathrm{R^n}$ is the weight vector, $\boldsymbol{\theta}_{\mathrm{k}} \in \mathrm{R^1}$ is the threshold, and $(\mathbf{w}_{\mathrm{k}})^{\mathrm{T}}\mathbf{x}$ is the inner product.

The feature vector $\mathbf{x}$ is situated on the *positive side* of the hyperplane $H(\mathbf{w}_{\mathrm{l}}, \boldsymbol{\theta}_{\mathrm{l}})$ if and only if $(\mathbf{w}_{\mathrm{k}})^{\mathrm{T}}\mathbf{x} > \boldsymbol{\theta}_{\mathrm{l}}$. Similarly, the vector $\mathbf{x}$ is situated on the *negative side* of $H(\mathbf{w}_1, \boldsymbol{\theta}_1)$ if and only if $(\mathbf{w}_{\mathrm{k}})^{\mathrm{T}}\mathbf{x} < \boldsymbol{\theta}_1$.

Definition 2. *The reference sets are linearly separable if each of the sets* C_k *(4.1) can be fully separated from the sum of the remaining sets* $C_{k'}$ *by some hyperplane* $H(\mathbf{w}_{k'}\theta_k)$ *(4.4):*

$$(\forall k \in \{1,\ldots,K'\}) \quad (\exists \boldsymbol{w}_k, \theta_k)\,(\forall \boldsymbol{x}_j(k) \in C_k) \;\; \boldsymbol{w}_k^T \boldsymbol{x}_j(k) > \theta_k$$
$$\textit{and} \quad (\forall \boldsymbol{x}_{j'}(k') \in C_{k'}) \;\; \boldsymbol{w}_k^T \boldsymbol{x}_{j'}(k') > \theta_k \qquad (4.4)$$

In accordance with the relation (4.4), all the vectors $\mathbf{x}_j(k)$ *belonging to the learning set* C_k *are situated on the positive side of the hyperplane* $H(\mathbf{w}_k, \theta_k)$ *(4.2) and all the feature vectors* $\mathbf{x}_{j'}(k')$ *from the remaining sets* $C_{k'}$ *are situated on the negative side of this hyperplane.*

It can be proved that the sets C_k *(4.1) are linearly separable if all the feature vectors* $\mathbf{x}_j(k)$ *are linearly independent (sufficient condition).*

4.3 Distance Functions Induced by Linear Transformations of the Feature Space $F[n]$

The nearest neighbours decision support rules are based on the distances $\delta(\mathbf{x}_0, \mathbf{x}_j(k))$ between the feature vector $\mathbf{x}_0$ of a new object and the labelled vectors $\mathbf{x}_j(k)$ from the reference sets C_k (4.1) [2]. Let us assume for a moment that m labelled feature vectors $\mathbf{x}_j(k)$ (4.1) are *ranked* $\{\mathbf{x}_{j(1)}, \mathbf{x}_{j(2)}, \ldots, \mathbf{x}_{j(n)}\}$ in respect to the distances $\delta(\mathbf{x}_0, \mathbf{x}_j(k))$ between the vectors $\mathbf{x}_0$ and $\mathbf{x}_j(k)$:

$$(\forall i \in \{1,\ldots,m-1) \quad \delta(\mathbf{x}_0, \mathbf{x}_{j(i)}) \;\leq\; \delta(\mathbf{x}_0, \mathbf{x}_{j(i+1)}) \qquad (4.5)$$

Let us define the *reference ball* $B_x(\mathbf{x}_0, K)$ which is centred in $\mathbf{x}_0$ and contains K first vectors $\mathbf{x}_{j(i)}(k)$:

$$B_x(\mathbf{x}_0, K) = \{\mathbf{x}_j(k) : \; \delta(\mathbf{x}_0, \mathbf{x}_{j(i)}) \;\leq\; \delta(\mathbf{x}_0, \mathbf{x}_{j(K)})\} \qquad (4.6)$$

In accordance with the K-nearest neighbours (K-*NN*) classification rule, the object $\mathbf{x}_0$ is allocated into this class $\boldsymbol{\omega}_k$ $(k = 1, \ldots.., K')$ where most of the labelled feature vectors $\mathbf{x}_j(k)$ from the ball $B_x(\mathbf{x}_0, K)$ (10) belong [2]:

$$\boldsymbol{if}\ (\forall l \in \{1,\ldots,K'\})\ n_k \;\geq\; n_l\ \boldsymbol{then}\ \mathbf{x}_0 \in \boldsymbol{\omega}_k \qquad (4.7)$$

where n_k is the number of the vectors $\mathbf{x}_j(k)$ from the set C_k (4.1) contained in the ball $B_x(\mathbf{x}_0, K)$ (6).

The decision rule similar to (4.11) is applied also in the case based reasoning scheme. It is assumed in this case that the reference ball $B_x(\mathbf{x}_0, K)$ contains such vectors $\mathbf{x}_{j(i)}(k)$ which are most *similar* to the vector $\mathbf{x}_0$.

The Euclidean distance $\delta_E(\mathbf{x}_0, \mathbf{x}_{j(i)})$ between the feature vectors $\mathbf{x}_0$ and $\mathbf{x}_{j(i)}$ is commonly used in the case based reasoning or in the nearest neighbours classification rule (8):

$$\delta_E{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)}) = (\mathbf{x}_0 - \mathbf{x}_{j(i)})^T (\mathbf{x}_0 - \mathbf{x}_{j(i)}) \qquad (4.8)$$

A quality of the decision rule (4.11) based on the Euclidean distance $\delta_E(\mathbf{x}_0,\ \mathbf{x}_{j(i)})$ can be improved in some cases through modification of the distance function through transformations of the feature space $\boldsymbol{F}[n]$.

The Mahalanobis distance function $\delta_M(\mathbf{x}_0, \mathbf{x}_j)$ in the feature space $\boldsymbol{X}$ is defined on the basis of the covariance matrix Σ [4]

$$\delta_M{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)}) = (\mathbf{x}_0 - \mathbf{x}_{j(i)})^T \Sigma^{-1} (\mathbf{x}_0 - \mathbf{x}_{j(i)}) \tag{4.9}$$

The Mahalanobis distance function $\delta_M{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)})$ takes into account the linear dependencies in the pairs of the features x_k and x_1. When the covariance matrix Σ is equal to the unit matrix $\boldsymbol{I}_n$, then the Mahalanobis distance $\delta_M{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)})$ is reduced to the Euclidean distance $\delta_M{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)}))$ (8).

Let us consider the linear transformations of the feature vectors $\mathbf{x}_j(k)$. Such transformations can be represented in the matrix form given below:

$$\mathbf{y}_j(k) = \boldsymbol{W}^T \mathbf{x}_j(k) \quad (j = 1, \ldots.., m) \tag{4.10}$$

where $\boldsymbol{W} = [\mathbf{w}_1,\ \mathbf{w}_2, \ldots\ldots,\ \mathbf{w}_{n'}]$ is the matrix of dimension $(n \times n')$ with $1 \le n' \le n$.

The relation (4.10) allows one to generate the transformed learning sets C'_k, where

$$C'_k = \{\mathbf{y}_j(k)\} \quad (j \in I_k) \tag{4.11}$$

Let us define the *induced distance function* $\delta_I(\mathbf{x}_0,\ \mathbf{x}_{j(i)})$ between the feature vectors $\mathbf{x}_0$ and $\mathbf{x}_{j(i)}$ as the Euclidean distance function $\delta_E(\mathbf{y}_0,\ \mathbf{y}_{j(i)})$ (4.8) between adequate points $\mathbf{y}_0$ and $\mathbf{y}_{j(i)}$ transformed in accordance with (4.10).

$$\begin{aligned} \delta_I{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)}) &= \delta_E{}^2(\mathbf{y}_0, \mathbf{y}_{j(i)}) = (\mathbf{y}_0 - \mathbf{y}_{j(i)})^T (\mathbf{y}_0 - \mathbf{y}_{j(i)}) \\ &= (\mathbf{x}_0 - \mathbf{x}_j(k))^T \boldsymbol{W}\boldsymbol{W}^T (\mathbf{x}_0 - \mathbf{x}_j(k)) \end{aligned} \tag{4.12}$$

The *induced ball* $B_I(\mathbf{x}_0, K)$ can be defined by using the distance function $\delta_I(\mathbf{x}_0, \mathbf{x}_{j(i)})$ (4.12).

$$\begin{aligned} B_I(\mathbf{x}_0, K) &= \{\mathbf{x}_j(k) : \delta_E{}^2(\mathbf{y}_0, \mathbf{y}_{j(i)}) \le \delta_E{}^2(\mathbf{y}_0, \mathbf{y}_{j(K)})\} \\ &= \{\mathbf{x}_j(k) : \delta_I{}^2(\mathbf{x}_0, \mathbf{x}_{j(i)}) \le \delta_I{}^2(\mathbf{x}_0, \mathbf{x}_{j(K)})\} \end{aligned} \tag{4.13}$$

where points $\mathbf{x}_{j(i)}$ are *ranked* $\{\mathbf{x}_{j(1)},\ \mathbf{x}_{j(2)}, \ldots\ldots..,\ \mathbf{x}_{j(n)}\}$(4.5) in accordance with the induced distance function $\delta_I(\mathbf{x}_0,\ \mathbf{x}_{j(i)})$ (4.12).

The induced ball $B_y(\mathbf{x}_0, K)$ contains such K feature vectors $\mathbf{x}_j(k)$ which are the most similar to $\mathbf{x}_0$ in accordance with the distance function $\delta_I(\mathbf{y}_0,\ \mathbf{y}_{j(i)})$ (4.11).

The case based reasoning (*CBR*) or the nearest neighbours decision rules (4.7) can be based on the induced ball $B_I(\mathbf{x}_0, K)$ (4.13):

If most of the labelled vectors $\mathbf{x}_j(k)$ from the induced ball $B_I(\mathbf{x}_0, K)$ belongs to the class ω_k, **then** the object represented by $\mathbf{x}_0$ should be assigned to this class. (4.14)

The performance of the above decision (*classification*) rule can be optimised through a special choice of the vectors $\mathbf{w}_i$ $(i = 1, \ldots.., n')$ in the transformation (4.10). A basic measure of the classification rule performance is the *error rate* – the fraction of new objects that are assigned to the wrong category [3].

4.4 Whitening of Reference Sets

An important role in classification is played by such linear transformations (4.9), which reduce correlation of the learning sets C_k (4.1) [2]. Such transformations can be built on the basis of the eigenvectors $\mathbf{k}_i$ and the eigenvalues λ_i of the covariance matrix Σ. Let us take into consideration the covariance matrix Σ_k estimated on the set C_k (4.1)

$$\Sigma_k = \sum_{j \in I_k} (x_j(k) - \mu_k)\,(x_j(k) - \mu_k)^T / (m_k - 1) \tag{4.15}$$

where μ_k is the mean vector in the set C_k

$$\mu_k = \Sigma \underset{j \in I_k}{\mathbf{x}_j(k)/m_k} \tag{4.16}$$

The eigenvalue problem with the covariance matrix Σ_k is formulated as the search for the eigenvectors $\mathbf{k}_i$ and the eigenvalues λ_i of the covariance matrix Σ_k. The eigenvectors $\mathbf{k}_i$ and the eigenvalues λ_i fulfil the below equation

$$\Sigma_k \ \mathbf{k}_i = \lambda_i \ \mathbf{k}_i \tag{4.17}$$

with an additional condition of the unit length

$$\mathbf{k}_i{}^T \mathbf{k}_i = 1 \tag{4.18}$$

The eigenvectors $\mathbf{k}_i$ and $\mathbf{k}_k$ corresponding to different eigenvalues λ_i and λ_k $(\lambda_i \neq \lambda_k)$ are orthogonal

$$\mathbf{k}_i{}^T \mathbf{k}_k = 0 \tag{4.19}$$

Let us assume that the linear transformations (4.10) is defined by n' $(1 \leq n' \leq n)$ orthogonal eigenvectors $\mathbf{k}_i$ with positive eigenvalues λ_i $(\lambda_i > 0)$. Typically, the eigenvectors $\mathbf{k}_i$ and the greatest eigenvalues λ_i are taken into

consideration. We are considering the linear transformation (4.10) with the columns of the matrix $\boldsymbol{W}$ formed by the vectors $\mathbf{k}_i/(\lambda_i)^{1/2}$

$$\boldsymbol{W}_k = [\mathbf{k}_1/(\lambda_1)^{1/2}, \ldots\ldots\ldots, \mathbf{k}_{n'}/(\lambda_{n'})^{1/2}] \tag{4.20}$$

The transformed vectors $\mathbf{y}_j(k)$ (4.9) form the set C'_k (4.11) with the mean vectors $\mu_{k'}$ (4.16). The correlation matrix ${\Sigma_k}'$ (15) are defined on the transformed vectors $\mathbf{y}_j(k)$ (9) from one set C'_k (4.11)

$$\begin{aligned} {\Sigma_k}' &= \Sigma(\mathbf{y}_j(k) - {\mu_k}')(\mathbf{y}_j(k) - {\mu_k}')^T/(m_k - 1) \\ &= \boldsymbol{W}^T \sum_{\substack{j \in I_k \\ j \in I_k}} (\mathbf{x}_j(k) - \mu_k)(\mathbf{x}_j(k) - \mu_k)^T \boldsymbol{W}/(m_k - 1) \\ &= \boldsymbol{W}^T \Sigma_k \boldsymbol{W} = \boldsymbol{I}_{n' \times n'} \end{aligned} \tag{4.21}$$

where $\boldsymbol{I}_{n' \times n'}$ is the unit matrix of the dimension $(n' \times n')$.

It could be seen that the decision rule (4.13) with the Euclidean distance $\delta_E(\mathbf{y}_0, \mathbf{y}_{j(i)})$ (4.8) in the transformed space is equivalent to the decision rule (4.7) with the Mahalanobis distance functions ${\delta_M}^2(\mathbf{x}_0, \mathbf{x}_{j(i)})$ (4.9) in the feature space $\boldsymbol{F}[n]$, where the points $\mathbf{y}_0$ and $\mathbf{y}_{j(i)}$ are obtained through the transformation (4.10) with (4.20) of the points $\mathbf{x}_0$ and $\mathbf{x}_{j(i)}$, adequately.

In accordance with the equation (4.20), the transformation (4.10) decorrelates the set C'_k (4.11). The classification rule (4.14) based on the ball $B_I(\mathbf{x}_0, K)$ (4.13) which is induced by the transformation (4.10) gives the possibility to decrease the error rate [5]. Results of some experiments which support this statement are described in a farther part of the presented chapter. In these experiments the induced ball $B_I(\mathbf{x}_0, K)$ (13) has been defined on the basis of the Euclidean distance function ${\delta_E}^2(\mathbf{y}_0, \mathbf{y}_{j(i)})$ (4.8) in the transformed space. Generally, the decision rule (4.7) with the Euclidean distance ${\delta_E}^2(\mathbf{y}_0, \mathbf{y}_{j(i)})$ (4.8) can be matched in the best manner to data sets C'_k (4.11) with the unit correlation matrix ${\Sigma_k}'$.

Example 1. The numerical experiment has been performed on two-dimensional data sets C_k and C'_k (points on the plane). Data were generated from normal distributions with different covariance matrices and had different mean vectors. They belonged to three overlapping classes. The correlation coefficients were accordingly $\rho_1 = -0.9$, $\rho_2 = 0.2$ and $\rho_3 = -0.6$. There were 30 objects in every class. To check the differences between classification quality using whitening process the original data had been transformed. First transformation was the whitening based on the transformation matrix built using covariance matrix of the first class, analogically – on the second and on the third class covariance matrix. The last transformation has been performed using transformation matrix built using pooled estimate of the common covariance matrix (Fig. 4.1).

The results of classification errors for *K-NN* rule for the number of neighbours from $K = 1$ to $K = 10$ are shown in Table 4.1.

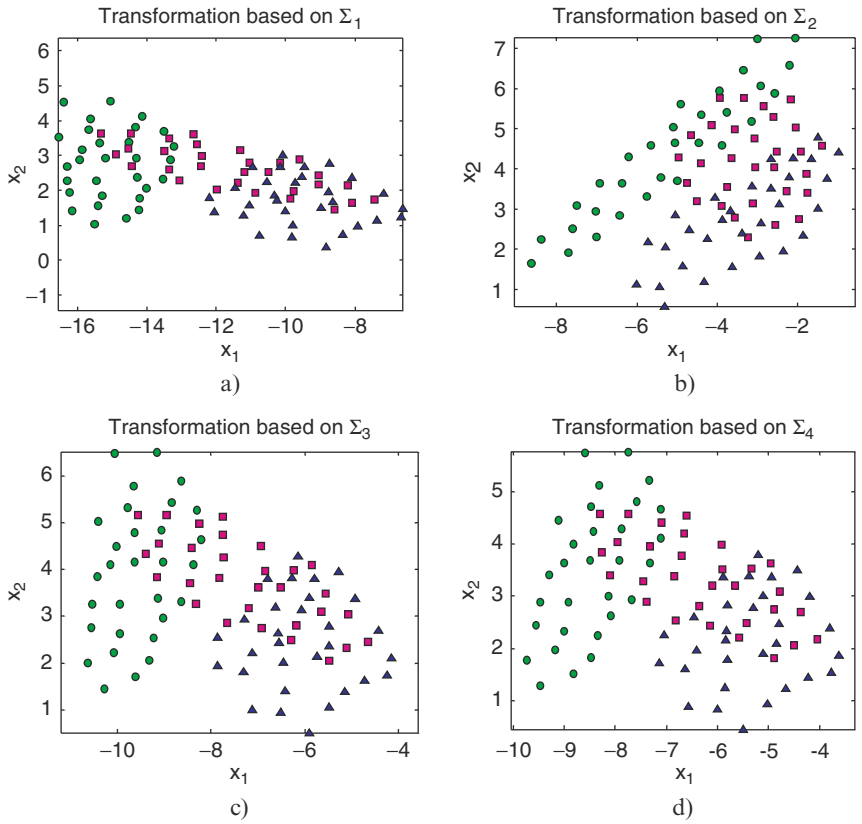

Fig. 4.1. Plots of transformed data (a) based on Σ_1, (b) based on Σ_2, (c) based on Σ_3 and (d) based on Σ_W

The mean value of the classification error for original data using *K-NN* rule was 43%. Adapting decorrelations we have achieved 33% as a mean of error_Σ_1, 39% (mean of error_Σ_2), 35% (mean of error_Σ_3) and 35% (mean of error_Σ_W). We can observe in the above results that the decorrelation of the learning sets C_k can improve the *K-NN* rule based on the Euclidean distance $\delta_E(\mathbf{x}_0, \mathbf{x}_j)$. The decorrelation of the learning sets C_k has entailed including the Mahalonobis distance $\delta_M(\mathbf{x}_0, \mathbf{x}_j)$ from these sets. With such interpretation, we can claim that the replacement of the Euclidean distance $\delta_E(\mathbf{x}_0, \mathbf{x}_j)$ by Mahalonobis distance $\delta_M(\mathbf{x}_0, \mathbf{x}_j)$ can lead to the improvement of the *K-NN* or the CBR rule. In the case of more than two classes using transformations based on single class C_k is preferred, because of the effect of conjoin different covariance matrices into one pooled estimate of the common covariance matrix. In our example the best classification quality we achieved for transformed data with transformation matrix built using covariance matrix for the second class Σ_2 (77%).

Table 4.1. Comparison of the classification error for *K-NN* classifiers ($K = 1, 2, \ldots, 10$) for correlated learning sets C_k (error_nd) and decorrelated sets C'_k (error_Σ_1 – decorrelation based on the covariance matrix of the C_1 set, error_Σ_2 – decorrelation based on the covariance matrix of the C_2 set, error_Σ_3 – decorrelation based on the covariance matrix of the C_3 set, error_Σ_W – decorrelation based on the pooled estimate of the common covariance matrix)

		Decorrelation			
K	Error_nd	Error_Σ_1	Error_Σ_2	Error_Σ_3	Error_Σ_W
1	0.4111	0.4667	0.5667	0.5222	0.5222
2	0.2444	0.2556	0.3444	0.3	0.3111
3	0.4	0.4	0.5333	0.4667	0.4778
4	0.3667	0.2	0.3667	0.2667	0.2667
5	0.5	0.3889	0.4444	0.4222	0.4222
6	0.3889	0.3111	0.3	0.2556	0.2
7	0.4444	0.3778	0.3889	0.3778	0.3444
8	0.4556	0.2778	0.2556	0.3111	0.2889
9	0.5111	0.4222	0.4	0.3444	0.3333
10	0.5333	0.2889	0.3111	0.2444	0.3111

4.5 Perceptron Criterion Functions (CPL)

The *perceptron* criterion function $\Phi(\mathbf{w}, \theta)$ originated from neural networks theory [3,9]. $\Psi(\mathbf{w}, \theta)$ is the convex and piecewise linear (CPL) criterion function. The designing transformation (4.10) can be based on the minimisation of the perceptron criterion function [6].

It is convenient to define the perceptron criterion function $\Phi(\mathbf{w}, \theta)$ by using the positive G^+ and the negative G^- sets of the feature vectors $\mathbf{x}_j$ (1).

$$G^+ = \{\mathbf{x}_j\} \quad (j \in J^+) \quad \textit{and} \quad G^- = \{\mathbf{x}_j\} \quad (j \in J^-) \tag{4.22}$$

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

$$\varphi_j{}^+(\mathbf{w}, \theta) = \begin{array}{lll} \ldots\ldots\ldots & & \\ 1 - \mathbf{w}^T\mathbf{x}_j + \theta & \textbf{if} & \mathbf{w}^T\mathbf{x}_j - \theta \leq 1 \\ 0 & \textbf{if} & \mathbf{w}^T\mathbf{x}_j - \theta > 1 \end{array} \tag{4.23}$$

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

$$\upsilon_j{}^-(\mathbf{w}, \theta) = \begin{array}{lll} 1 + \mathbf{w}^T\mathbf{x}_j - \theta & \textbf{if} & \mathbf{w}^T\mathbf{x}_j - \theta \geq -1 \\ 0 & \textbf{if} & \mathbf{w}^T\mathbf{x}_j - \theta < -1 \end{array} \tag{4.24}$$

where $\mathbf{w} = [w_1, \ldots, w_n]^T \in R^n$ is the weight vector and $\theta_k \in R^1$ is the threshold.

Both the penalty functions $\upsilon_j{}^+(\mathbf{w},\theta)$ and $\upsilon_j{}^-(\mathbf{w},\theta)$ are convex and piecewise linear. The penalty function $\upsilon_j{}^+(\mathbf{w},\theta)$ is aimed at placing the vector $\mathbf{x}_j(\mathbf{x}_j \in G^+)$ on the positive side of the hyperplane $H(\mathbf{w},\theta)$ (4.3). Similarly, the function $\upsilon_j{}^-(\mathbf{w},\theta)$ should insert the vector $\mathbf{x}_j(\mathbf{x}_j \in G^-)$ on the negative side of this hyperplane.

The perceptron criterion function $\Phi(\mathbf{w},\theta)$ is determined by the sets G^+ and G^- is the weighted sum of the penalty functions $\upsilon_j{}^+(\mathbf{w},\ \theta)$ and $\upsilon_j{}^-(\mathbf{w},\ \theta)$

$$\Phi(\mathbf{w},\theta) = \underset{j \in J^+}{\Sigma} \alpha_j{}^+\varphi_j{}^+(\mathbf{w},\theta) + \underset{j \in J^-}{\Sigma} \alpha_j{}^-\varphi_j{}^-(\mathbf{w},\theta) \tag{4.25}$$

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

$$\Phi^* = \Phi(\mathbf{w}^*,\theta^*) = \min_{\mathbf{w},\theta} \Phi(\mathbf{w},\theta) \geq 0 \tag{4.26}$$

The basis exchange algorithms which are similar to the linear programming allow one to find the minimum of the criterion function $\Phi(\mathbf{w},\theta)$ efficiently even in the case of large, multidimensional data sets G^+ and G^- (4.22) [7].

It has been proved that the minimum value Φ^* of the perceptron criterion function $\Phi(\mathbf{w},\theta)$ (4.25) is equal to zero $(\Phi^* = 0)$ if and only if the positive G^+ and the negative G^- sets (4.22) are linearly separable (4.4). In this case, all elements $\mathbf{x}_j$ of the set G^+ (4.22) are located on the positive side of the hyperplane $H(\mathbf{w}^*,\theta^*)$ (4.3) and all elements $\mathbf{x}_j$ of the set G^- are located on the negative side:

$$\begin{aligned} &(\forall \mathbf{x}_j \in G^+)(\mathbf{w}^*)^T\mathbf{x}_j > \theta_1^* \\ \textbf{\textit{and}}\ &(\forall \mathbf{x}_{j'} \in G^-)(\mathbf{w}^*)^T\mathbf{x}_{j'} < \theta_1^* \end{aligned} \tag{4.27}$$

If the sets G^+ and G^- (4.22) are not linearly separable (4.4), then the above relation is fulfilled not by all but by a majority of the elements $\mathbf{x}_j$ of these sets.

Minimization of the function $\Phi(\mathbf{w},\theta)$ (4.25) allows one to find optimal parameters $(\mathbf{w}^*,\theta^*)$ which can define the hyperplane $H(\mathbf{w}^*,\theta^*)$ (4.3), which separates relatively well two sets G^+ and G^- (4.22). The vector $\mathbf{w}_1{}^*$ can be used also as one of the columns of the transformation matrix $\boldsymbol{W}$ (4.10).

4.6 Four-Fields Diagnostic Maps of the System *Hepar*

The computer system *Hepar* aggregates the clinical database with tools for data exploration and diagnosis support [8]. The database of the system *Hepar* contains hepathological data. An essential part of the system is data visualisation module. For the purpose of data visualisation there are used linear transformations from multidimensional feature space $\boldsymbol{F}[n]$ on a plane. Such transformations allow for inducing the distance function $\delta_I{}^2(\mathbf{x}_0,\mathbf{x}_{j(i)})$ (4.12)

based both on the Euclidean distance ${\delta_{\mathrm{E}}}^2(\mathbf{y}_0, \mathbf{y}_{\mathrm{j(i)}})$ (4.8) as well as on a subjective measures of similarity.

The parameters $\mathbf{w}^*$ and $\boldsymbol{\theta}^*$ determining minimum (4.26) of the criterion function $\Psi(\mathbf{w}, \boldsymbol{\theta})$ (4.25) can be also used in definition of the affine transformation of the feature vectors $\mathbf{x}$ on a line (4.9):

$$y = (\mathbf{w}^*)^{\mathrm{T}}\mathbf{x} - \boldsymbol{\theta}^* \tag{4.28}$$

where $\mathbf{w} = [\mathrm{w}_1, \ldots\ldots, \mathrm{w}_\mathrm{n}]^{\mathrm{T}}$ is the parameter vector which determines direction of the line.

Such transformations have been applied in the system *Hepar* for definition of the visualising planes. This system allows for designing pairs of special visualizing transformations (4.28), which result in the so called *diagnostic maps*. Two linearly independent transformations (4.28) give possibility to produce such a visualizing plane (*diagnostic map*), which relatively well separates four groups of patients. The diagnostic maps are used for inducing the similarity measure between feature vector of a new patient $\mathbf{x}_0$ and the vectors $\mathbf{x}_\mathrm{j}(k)$ from the reference sets $\mathrm{C_k}$ (4.1).

The affine transformation of the feature vectors $\mathbf{x}_\mathrm{j}$ (4.1) on a plane can be represented in a below manner

$$\mathbf{y}_\mathrm{j} = [\mathrm{y_{j1}}, \mathrm{y_{j2}}]^{\mathrm{T}} = [(\mathbf{w}_\mathrm{l}^*)^{\mathrm{T}}\mathbf{x}_\mathrm{j} - {\boldsymbol{\theta}_1}^*, (\mathbf{w}_2{}^*)^{\mathrm{T}}\mathbf{x}_\mathrm{j} - {\boldsymbol{\theta}_2}^*]^{\mathrm{T}} \tag{4.29}$$

where $\mathbf{w}_\mathrm{i}^* = [\mathrm{w_{i1}}, \ldots\ldots, \mathrm{w_{in}}]^{\mathrm{T}}$ $(i = 1, 2)$ are the parameter vectors that span a plane.

The *scatterplots* or, in other words, the *maps* of data can be generated as a result of visualisation of the transformed points $\mathbf{y}_\mathrm{j}(k)$. If the vectors $\mathbf{w}_\mathrm{i}$ are orthogonal $((\mathbf{w}_1{}^*)^{\mathrm{T}}\ \mathbf{w}_2 = 0)$ and have the unit length $((\mathbf{w}_1{}^*)^{\mathrm{T}}\ \mathbf{w}_1{}^* = (\mathbf{w}_2{}^*)^{\mathrm{T}}\ \mathbf{w}_2{}^* = 0)$ then the transformations (4.2) describes the *projection* of the feature vectors $\mathbf{x}_\mathrm{j}(k)$ on the visualizing plane $P(\mathbf{w}_1{}^*,\ \mathbf{w}_2{}^*; \boldsymbol{\theta}^*)$

$$P(\mathbf{w}_1, \mathbf{w}_2; \boldsymbol{\theta}) = \{\mathbf{x} : \mathbf{x} = \alpha_1\mathbf{w}_1 + \alpha_2\mathbf{w}_2 + \boldsymbol{\theta},\ \text{where } \alpha_i \in R^1\} \tag{4.30}$$

where $\boldsymbol{\theta}_\mathbf{i} = [\theta_1, \theta_2]^{\mathrm{T}}$

Example 2. Let us consider this example in order to explain the basic principles of the diagnostic map designing in the framework of the system *Hepar*. We have taken into consideration four learning sets $\mathrm{C_k}$ (4.1) extracted from the *Hepar* database [8]

$$\begin{array}{ll} C_9\text{–Hepatitis chronica activa} & \text{–91 patients} \\ C_{13}\text{–Steatosis hepatis} & \text{–67 patients} \\ C_{15}\text{–Hiperbilirubinemia functionalis} & \text{–56 patients} \\ C_{22}\text{–Cirrhosis hepatis billiaris primaria} & \text{–272 patients} \end{array} \tag{4.31}$$

Patients from these sets $\mathrm{C_k}$ have been described by the feature vectors $\mathbf{x}_\mathrm{j}(k)$ of dimensionality n equal to 106. The components $\mathrm{x_i}$ of the vectors $\mathbf{x}_\mathrm{j}(k)$

were numerical results of various diagnostic examinations of a given patient. Numerical results of both laboratory tests ($x_i \in R$) as well as patients symptoms ($x_i \in \{0, 1\}$) have been taken as features x_i.

The maps can reflect an actual diagnostic hypothesis of a medical doctor (an user). The user declares which classes ω_k should be located into particular quarters of the map and which features (tests) x_i are to be used in the visualizing transformation or, in other words, used for hypothesis examination. The above map (Fig. 4.2) resulted from the affine transformation (4.29) of the 106 – dimensional feature vectors $\mathbf{x}_j(k)$ on a visualizing plane.

The affine transformation (4.29) of the feature vectors $\mathbf{x}_j$ on a visualising plane is determined by two pairs of parameters $(\mathbf{w}_1{}^*,\ \theta_1{}^*)$ and $(\mathbf{w}_2{}^*,\ \theta_2{}^*)$. These parameters have been induced from the sets (4.30) through minimisation of two perceptron criterion functions $\Phi_1(\mathbf{w}, \theta)$ and $\Phi_2(\mathbf{w},\ \theta)$ (4.25). Each function $\Phi_k(\mathbf{w},\ \theta)$ was defined by their own pair of the sets $G_k{}^+$ and $G_k{}^-$ (22), where

$$G_1{}^+ = C_{13} \cup C_{15} \quad \textit{and} \quad G_1{}^- = C_9 \cup C_{22} \tag{4.32}$$

$$G_2{}^+ = C_9 \cup C_{13} \quad \textit{and} \quad G_2{}^- = C_{15} \cup C_{22} \tag{4.33}$$

If each pair of the sets $G_k{}^+$ and $G_k{}^-$ $(k = 1,\ 2)$ is linearly separable (4.4), then the transformation (4.29) based on the above sets assures the exact placements of the learning sets C_k (4.30) in an adequate quarter of the diagnostic map (Fig. 4.2).

The transformation (4.29) defines the coordinates $\mathbf{y}_j(k) = [y_{j1}(k),\ y_{j2}(k)]$ on the map of particular feature vectors $\mathbf{x}_j(k)$. The vector of $\mathbf{x}_0$ of a new patient can be located on the map as $\mathbf{y}_0$ by using the transformation (4.29). As a result, the system can be used in the diagnosis support in accordance with the CNR or the *K-NN* schemes (4.7), which are based on the Euclidean distances $\delta_E(\mathbf{y}_0, \mathbf{y}_j[k])$ (4.8) between the transformed vectors $\mathbf{y}_0$ and $\mathbf{y}_j[k]$. It has been demonstrated experimentally that despite the significant reduction of the problem dimensionality (from $n = 106$ to $n' = 2$), the replacement of the distances $\delta_E(\mathbf{x}_0, \mathbf{x}_j[k])$ by the distances $\delta_E(\mathbf{y}_0, \mathbf{y}_j[k])$ induced through a diagnostic map gives possibility to reduce the error rate of the classification rules [8, 9].

4.7 Fisher Linear Discriminant and Principal Components

Let us consider further linear transformations (4.10) of data sets C_k (4.1) from n-dimensional feature space $\boldsymbol{F}[n]$ onto line. Such problem is analysed in the *discriminant analysis*. Discriminant analysis seeks direction $\mathbf{w}$ that are efficient in separation on a line of two data sets C_1 and C_2 (4.1).

$$y = (\mathbf{w})^T \mathbf{x} \tag{4.34}$$

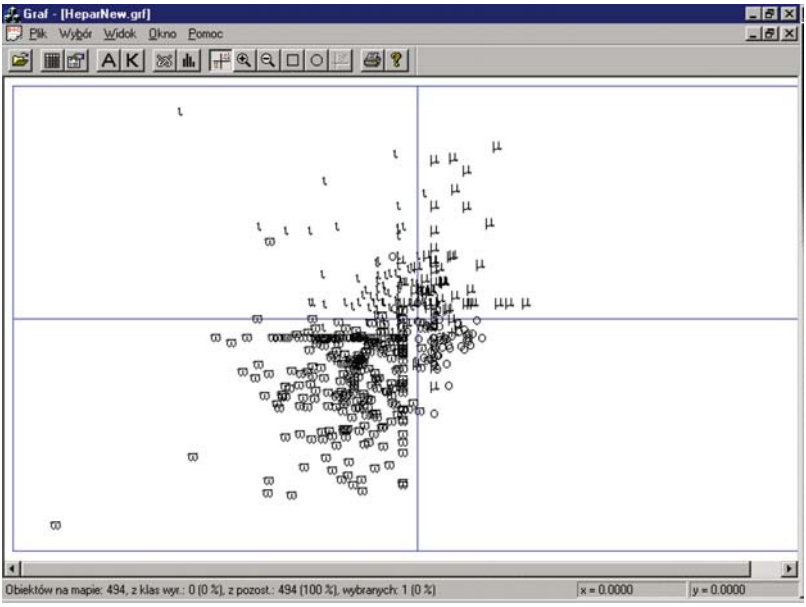

Fig. 4.2. The diagnostic map with the following structure: C_9, *the upper-left quarter*; C_{13}, *the upper-right quarter*; C_{15}, *the lower-right quarter*; C_{22}, *the lower-left quarter*

The direction vectors $\mathbf{w}$ which are used in the discriminant analysis have the unit length

$$\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1 \tag{4.35}$$

The linear transformation (4.34) with an additional condition (4.33) describes the projection $y_j = (\mathbf{w})^{\mathrm{T}}\ \mathbf{x}_j$ of the corresponding vectors $\mathbf{x}_j$ onto a line in the direction of $\mathbf{w}$.

A fundamental role in the *discriminant analysis* is played by the Fisher's criterion function $J(\mathbf{w})$.

$$J(\mathbf{w}) = |\mu_1(\mathbf{w}) - \mu_2(\mathbf{w})|/(s_1(\mathbf{w})^2 + s_2(\mathbf{w})^2) \tag{4.36}$$

where $|\mu_1(\mathbf{w}) - \mu_2(\mathbf{w})|$ is the distance between the projected mean vectors μ_1 and μ_2 (4.16)

$$|\mu_1(\mathbf{w}) - \mu_2(\mathbf{w})| = |\mathbf{w}^{\mathrm{T}}(\mu_1 - \mu_2)| \tag{4.37}$$

and $s_k(\mathbf{w})^2$ $(k = 1,\ 2)$ is the *within-class scatter* (a measure of variance) of the projected points y_j from the set C_k.

$$s_k(\mathbf{w})^2 = \sum_{j \in I_k} (y_j(k) - \mu_k(\mathbf{w}))^2 \tag{4.38}$$

The below optimization problem is based on the Fisher's criterion function $J(\mathbf{w})$.

$$J(\mathbf{w}^*) = \max_{\mathrm{w}} J(\mathbf{w}) \tag{4.39}$$

In accordance with the Fisher's criterion, the vector $\mathbf{w}^*$ that constitutes maximum of the function $J(\mathbf{w})$ (4.30) determines the best discriminant line. The optimal vector $\mathbf{w}^*$ determines large distance between the projected means μ_1 and μ_2 (4.31) relatively to some measure of the variance of the projected points $\mathrm{y_j}$.

The criterion function $J(\mathbf{w})$ can be represented in a matrix form. Let us define for this purpose the scatter matrices $\boldsymbol{S}_\mathrm{k}(k = 1, 2)$ and $\boldsymbol{S}_\mathrm{W}$

$$S_\mathrm{k} = \sum_{\mathrm{j} \in \mathrm{I_k}} (\mathbf{x_j}(k) - \mu_\mathrm{k})(\mathbf{x_j}(k) - \mu_\mathrm{k})^\mathrm{T} \tag{4.40}$$

and

$$S_\mathrm{W} = S_1 + S_2 \tag{4.41}$$

$\boldsymbol{S}_\mathrm{W}$ is called the *within-class scatter matrix*. The scatter $\mathrm{s_k}(\mathbf{w})^2$ (4.38) can be expressed as

$$\mathrm{s_k}(\mathbf{w})^2 = \sum_{j \in \mathrm{I_k}} (\mathbf{w}^\mathrm{T}\mathbf{x_j}(k) - \mathbf{w}^\mathrm{T}\mu_\mathrm{k})^2 = \Sigma \mathbf{w}^\mathrm{T}(\mathbf{x_j}(k) - \mu_\mathrm{k})(\mathbf{x_j}(k) - \mu_\mathrm{k})^T\mathbf{w} = \mathbf{w}^\mathrm{T} S_\mathrm{k}\mathbf{w} \tag{4.42}$$

thus

$$\mathrm{s_1}(\mathbf{w})^2 + \mathrm{s_2}(\mathbf{w})^2 = \mathbf{w}^\mathrm{T} S_\mathrm{W}\mathbf{w} \tag{4.43}$$

Similarly,

$$(\mu_1(\mathbf{w}) - \mu_2(\mathbf{w}))^2 = (\mathbf{w}^\mathrm{T}\mu_1 - \mathbf{w}^\mathrm{T}\mu_2)^2 = \mathbf{w}^\mathrm{T}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathrm{T}\mathbf{w} = \mathbf{w}^\mathrm{T} S_B \mathbf{w} \tag{4.44}$$

where

$$S_\mathrm{B} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathrm{T} \tag{4.45}$$

$\mathbf{S}_\mathrm{B}$ is called as the between-class scatter matrix.

The criterion function $J(\mathbf{w})$ (4.36) can be written as

$$J(\mathbf{w}) = \mathbf{w}^\mathrm{T} S_\mathrm{B}\mathbf{w} / \mathbf{w}^\mathrm{T} S_\mathrm{W}\mathbf{w} \tag{4.46}$$

The vector $\mathbf{w}^*$ that maximizes $J(\mathbf{w})$ must satisfy a generalized eigenvalue problem for some constant λ

$$S_\mathrm{B}\mathbf{w} = \lambda S_W \mathbf{w} \tag{4.47}$$

The vector $\mathbf{w}_\mathrm{F}$ that maximizes $J(\mathbf{w})$ is known as

$$\mathbf{w}_\mathrm{F} = {S_\mathrm{W}}^{-1}(\mu_1 - \mu_2) \tag{4.48}$$

Fisher's linear discriminant $\mathrm{y} = (\mathbf{w}_\mathrm{F})^\mathrm{T}\mathbf{x}$ (4.34), which is determined by the optimal vector $\mathbf{w}_\mathrm{F}$, yields the maximum ratio of the between-class scatter matrix to the within-class scatter on the projecting line.

In the case of the probabilistic normal model, when the conditional densities $f(\mathbf{x}/\omega_1)$ and $f(\mathbf{x}/\omega_2)$ are multivariate normal distributions $N(\mu_1,\ \Sigma)$ and $N(\mu_2,\ \Sigma)$ with the same covariance matrix Σ, then the optimal (*Bayesian*) decision boundary is the hyperplane $H(\mathbf{w}_B, \theta_B)$ (4.3), where

$$(\mathbf{w}_B)^T\mathbf{x} = \theta_B \tag{4.49}$$

and $\mathbf{w}_B$ is determined by an equation similar to (4.48)

$$\mathbf{w}_B = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \tag{4.50}$$

In this case, the optimal decision rule has the following form

$$\begin{array}{lll} \textit{if} & \mathbf{w}_B{}^T\mathbf{x} > \theta_B & \textit{then } \mathbf{x} \text{ should be allocated into the class } \omega_1 \\ \textit{if} & \mathbf{w}_B{}^T\mathbf{x} < \theta_B & \textit{then } \mathbf{x} \text{ should be allocated into the class } \omega_2 \end{array} \tag{4.51}$$

The above considerations have been related to discrimination between only two classes ($K' = 2$). When the number of classes c is greater than 2 ($K' > 2$), then the generalization of Fisher's linear discrimination involves $K' - 1$ linear discriminant functions [3]. In this case, it is designed projection (4.10) from n-dimensional space to a $(K' - 1)$-dimensional space.

Discriminant analysis seeks a projection that best separates data in a last squares sense. In contrast, principal component analysis (PCA) or *Karhunen-Loeve transform* seeks a projection that best represents data in a last squares sense. PCA deals with dimensionality reduction through such linear transformations (4.10) from n-dimensional space to a n' - dimensional space which preserve variability in data as much as possible.

Principal component analysis is based on n' linear transformations $y_i = (\mathbf{k}_i)^T\mathbf{x}$ that are defined by the eigenvectors $\mathbf{k}_i = [k_{i1},\ k_{i2}, \ldots\ldots,\ k_{in}]^T$ of the covariance matrix Σ_k (4.15).

$$y_i = (\mathbf{k}_i)^T\mathbf{x} = k_{i1}x_1 + k_{i2}\ x_2 + \ldots\ldots\ldots + k_{in}x_n \tag{4.52}$$

The covariance matrix Σ_k (4.15) of dimensionality $n \times n$ can have up to n normalised (4.18) eigenvectors $\mathbf{k}_i$ with positive eigenvalues λ_i. The eigenvectors $\mathbf{k}_i$ are ranked in accordance with the eigenvalues λ_i.

$$\mathbf{k}_1,\ \mathbf{k}_2, \ldots\ldots\ldots\ldots, \mathbf{k}_n \tag{4.53}$$

where

$$\lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots\ldots \geq \lambda_n > 0$$

The first principal component $y_1 = (\mathbf{k}_1)^T\ \mathbf{x}$ is defined by the eigenvector $\mathbf{k}_1$ with the largest eigenvalue λ_1. The second principal component $y_1 = (\mathbf{k}_1)^T\ \mathbf{x}$ is defined by the second eigenvector $\mathbf{k}_2$ and so on. Principal components are defined by n' eigenvectors $\mathbf{k}_i$ ($1 \leq n' \leq n$) with the largest eigenvalues λ_i. Often, there are just a few large eigenvalues λ_i and it can be assumed that remaining

$n - n'$ dimensions contain noise. Considerable dimensionality reduction can be achieved in such case through linear transformation of data.

The variance $\sigma_i{}^2$ of the ith principal component $y_i = (\mathbf{k}_i)^T\mathbf{x}$ can be estimated in the following manner (4.15 and 4.38)

$$\sigma_i{}^2 = \sum_{j\in I_k}(y_j(k) - \mu_k)^2/(m_k - 1) = \sum_{j\in I_k}((\mathbf{k}_i)^T(\mathbf{y}_j(k) - \mu_k)^2/(m_k - 1) \quad (4.54)$$
$$= \sum_{j\in I_k}((\mathbf{k}_i)^T(\mathbf{y}_j(k) - \mu_k)(\mathbf{y}_j(k) - \mu_k)^T\mathbf{k}_i/(m_k - 1) = (\mathbf{k}_i)^T\Sigma_k\mathbf{k}_i = \lambda_i$$

As it results from the above relation, the first principal component $y_1 = (\mathbf{k}_1)^T\ \mathbf{x}$ has the largest variance $\sigma_1{}^2$, and (4.53)

$$\sigma_1 \geq \sigma_2 \geq \ldots\ldots\ldots \geq \sigma_n > 0 \quad (4.55)$$

Example 3. Two eigenvectors $\mathbf{k}_i$ and $\mathbf{k}_{i'}$ with eigenvalues λ_i and $\lambda_{i'}$ of the kth covariance matrix Σ_k (4.15) can be used in the below visualising transformation (4.20 and 4.29), the data map determined by the kth set C_k (4.1).

$$\mathbf{y} = [y_1, y_2]^T = [(\mathbf{k}_i/(\lambda_i)^{1/2})^T(\mathbf{x} - \mu_k), (\mathbf{k}_{i'}/(\lambda_{i'})^{1/2})^T(\mathbf{x} - \mu_k)]^T \quad (4.56)$$

where μ_k is the mean vector (4.15) of the set C_k (4.1).

In accordance with the relation (4.55), all feature vectors $\mathbf{x}_j(k)$ (4.1) are transformed into the points $\mathbf{y}_j(k)$ on the visualising plane $P_k(\mathbf{k}_1, \mathbf{k}_2; \mathbf{0})$ (4.30) determined by the kth set C_k (4.1).

The mean value μ'_k (4.16) of the transformed points $\mathbf{y}_j(k)$ from the set C_k (4.1) is equal to zero.

$$\mu\prime_k = \sum_{j\in I_k}\mathbf{y}_j(k)/m_k = [0, 0]^T = \mathbf{0} \quad (4.57)$$

The covariance matrix Σ'_k (4.15) of the transformed points $\mathbf{y}_j(k)$ from the set C_k (4.1) is equal the unit matrix $\boldsymbol{I}_{2\times 2}$.

$$\Sigma'_k = \sum_{j\in I_k}(\mathbf{y}_j(k) - \mu'_k)(\mathbf{y}_j(k) - \mu'_k)^T/(m_k - 1)$$
$$= \sum_{j\in I_k}\mathbf{y}_j(k)\mathbf{y}_j(k)^T/(m_k - 1) = (W'_k)^T\Sigma_k W'_k = I_{2x2} \quad (4.58)$$

where Σ_k is the covariance matrix (4.15) and the matrix $\boldsymbol{W}'_k$ has the following form (4.20)

$$W'_k = [\mathbf{k}_i/(\lambda_i)^{1/2}, \mathbf{k}_{i'}/(\lambda_{i'})^{1/2}] \quad (4.59)$$

As it results from the relation (4.50) the transformed features y_1 and y_2 (where $\mathbf{y} = [y_1, y_2]^T$ (4.48)) are uncorrelated. Each of these features y_i has the mean value equal to zero and the variance equal to one in the set C_k (4.1).

In accordance with the considerations given in the *Example* 1 it should be profitable to use the visualizing transformation (4.48) and the diagnostic map for an inducing of similarity measure for the CBR or *K-NN* decision rules (4.7) with the Euclidean distance ${\delta_{\mathrm{E}}}^2(\mathbf{y}_0, \mathbf{y}_{\mathrm{j(i)}})$ (4.8).

For this purpose, the decisionic rule (4.7) can be modified in the following manner.

$$\textit{if} \quad (\forall l \in \{1, \ldots, K'\}) n_{\mathrm{k}}(k) \quad \geq \quad n_l(k) \quad \textit{then } \mathbf{x}_0 \quad \in \quad \omega_{\mathrm{k}} \tag{4.60}$$

where $n_l(k)$ is the number of the vectors $\mathbf{y}_{\mathrm{j}}(l)$ from the set C_l (4.1) contained in the ball $B_{\mathrm{k}}(\mathbf{y}_0, K)$, and

$$B_{\mathrm{k}}(\mathbf{y}_0, K) = \{\mathbf{y}_{\mathrm{j}} : \delta(\mathbf{y}_0, \mathbf{y}_{\mathrm{j(i)}}) \quad \leq \delta(\mathbf{y}_0, \mathbf{y}_{\mathrm{j(K)}})\} \tag{4.61}$$

where points $\mathbf{y}_{\mathrm{j(i)}}$ are *ranked* $\{\mathbf{y}_{\mathrm{j(1)}}, \ \mathbf{y}_{\mathrm{j(2)}}, \ldots\ldots\ldots, \ \mathbf{y}_{\mathrm{j(n)}}\}$ (4.5) in accordance with the Euclidean distance function $\delta_{\mathrm{E}}(\mathbf{y}_0, \mathbf{y}_{\mathrm{j(i)}})$ (4.8) on the plane P_{k} $(\mathbf{k_1}, \mathbf{k}_2; \mathbf{0})$ (4.30) determined by the kth set C_{k} (4.1).

The visualizing plane P_{k} $(\mathbf{k_1}, \mathbf{k}_2; \mathbf{0})$ (4.30) design in the above manner can be called as the *one-field diagnostic map*. Each of the maps P_{k} $(\mathbf{k_1}, \mathbf{k}_2; \mathbf{0})$ is centered on one of reference sets C_{k} (4.1).

4.8 Dipolar Separability Postulates

The linear transformations (4.10) can be defined on a variety of principles. Let us use for this purpose the concept of the mixed and clear dipoles formed by the feature vectors $\mathbf{x}_{\mathrm{j}}(k)$ (4.1) [6].

Definition 3. *A pair of the feature vectors* $(\mathbf{x}_{\mathrm{j}}(k), \mathbf{x}_{\mathrm{j}'}(k'))$ $(\mathbf{x}_{\mathrm{j}}(k) \neq \mathbf{x}_{\mathrm{j}'}(k'),$ $j' > j)$ *constitutes a mixed dipole if and only if the vectors* $\mathbf{x}_{\mathrm{j}}(k)$ *and* $\mathbf{x}_{\mathrm{j}'}(k')$ *belong to different classes* ω_{k} $(k \neq k')$. *Similarly, a pair of different feature vectors from the same class* ω_{k} *constitutes a clear dipole* $(\mathbf{x}_{\mathrm{j}}(k), \mathbf{x}_{\mathrm{j}'}(k))$.

The dipoles $\{\mathbf{x}_{\mathrm{j}}(k), \ \mathbf{x}_{\mathrm{j}'}(k')\}$ of the length $\delta_{\mathrm{x}}(j, j')$ are transformed by (4.10) into the dipoles $\{\mathbf{y}_{\mathrm{j}}(k), \ \mathbf{y}_{\mathrm{j}'}(k')\}$ – the pairs of the points $\mathbf{y}_{\mathrm{j}}(k)$ and $\mathbf{y}_{\mathrm{j}'}(k)$ situated in the Euclidean distance $\delta_{\mathrm{y}}(j, j')$, where

$$\begin{aligned} {\delta_{\mathrm{x}}}^2(j, j') = \ & (\mathbf{x}_{\mathrm{j}}(k) - \mathbf{x}_{\mathrm{j}'}(k'))^{\mathrm{T}}(\mathbf{x}_{\mathrm{j}}(k) - \mathbf{x}_{\mathrm{j}'}(k') & (4.62) \\ {\delta_{\mathrm{y}}}^2(j, j') = \ & (\mathbf{y}_{\mathrm{j}}(k) - \mathbf{y}_{\mathrm{j}'}(k'))^{\mathrm{T}}(\mathbf{y}_{\mathrm{j}}(k) - \mathbf{y}_{\mathrm{j}'}(k')) & (4.63) \\ = \ & (\mathbf{x}_{\mathrm{j}}(k) - \mathbf{x}_{\mathrm{j}'}(k'))^{\mathrm{T}} W W^{\mathrm{T}} (\mathbf{x}_{\mathrm{j}}(k) - \mathbf{x}_{\mathrm{j}'}(k')) \end{aligned}$$

We are interested in designing such transformations (4.10) which fulfil the following *separability inequalities*:

$$(\forall (j, j') \in I_{\mathrm{c}}) \quad {\delta_{\mathrm{y}}}^2(j, j') \leq \quad {\rho_{\mathrm{c}}}^2(j, j') \tag{4.64}$$

$$(\forall (j, j') \in I_{\mathrm{m}}) \quad {\delta_{\mathrm{y}}}^2(j, j') \geq \quad {\rho_{\mathrm{m}}}^2(j, j') \tag{4.65}$$

where I_c and I_m are the so called *control sets* (the sets of indices (j, j') of selected clear and mixed dipoles adequately, $\rho_c(j, j')$ and $\rho_m(j, j')$ are non-negative parameters (*margins*).

Separability postulate. *The linear transformation* (4.10) *should shorten the clear dipoles* $\{\mathbf{x}_j(k),\ \mathbf{x}_{j'}(k)\}$ *from the control set* I_c *to the length* ${\delta_y}^2(j, j')$ *less than* $\rho_c(j, j')$ (4.64) *and lengthen the mixed dipoles* $\{\mathbf{x}_j(k),\ \mathbf{x}_{j'}(k')\}$ *from the control set* I_m *to the length* ${\delta_y}^2(j, j')$ *more than* $\rho_m(j, j')$ (4.65).

The above separability postulate is aimed at designing such linear transformations (4.10) which enhance differences between categories ω_k. This postulate can be treated as an alternative which is complementary to the Fisher's criterion (4.39) used in discriminant analysis [3].

In the case of a linear transformation on the ith line $y = (\mathbf{w}_i)^T\mathbf{x}$, the separability inequalities (4.64) and (4.65) can be represented as the following sets of inequalities ($i = 1, 2, \ldots\ldots, n'$, with $1 \le n' \le n$):

$$(\forall(j, j') \in I_{ci}) - \rho_{ci}(j, j') < (\mathbf{w}_i)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k)) < \rho_{ci}(j, j') \tag{4.66}$$

$$(\forall(j, j') \in {I_{mi}}^+)\ (\mathbf{w}_i)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k')) > \rho_{mi}(j, j') \tag{4.67}$$

$$(\forall(j, j') \in {I_{mi}}^-)\ (\mathbf{w}_i)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k')) < -\rho_{mi}(j, j') \tag{4.68}$$

where ${I_{mi}}^+$ and ${I_{mi}}^-$ are disjoined subsets of the control set I_{mi} (${I_{mi}}^+ \cap {I_{mi}}^- = \emptyset$ and ${I_{mi}}^+ \cup {I_{mi}}^- = I_{mi}$) of the mixed dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$, $\rho_{c_i}(j, j')$ and $\rho_{m_i}(j, j')$ are the clear and the mixed margins defined on the ith line.

Remark 1: If two orthonormal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ (${\mathbf{w}_i}^T\mathbf{w}_i = 1$, ${\mathbf{w}_i}^T\mathbf{w}_k = 0$) fulfil the inequalities (separability postulate, (4.67), and (4.68)), then the inequalities (4.64) and (4.65) with the below parameters ${\rho_c}^2(j, j')$ and ${\rho_m}^2(j, j')$ are also fulfilled

$$(\forall(j, j') \in I_c)\ \ {\rho_c}^2(j, j') = {\rho_{c1}}^2(j, j') + {\rho_{c2}}^2(j, j') \tag{4.69}$$

$$(\forall(j, j') \in I_m)\ \ {\rho_m}^2(j, j') = {\rho_{m1}}^2(j, j') + {\rho_{m2}}^2(j, j') \tag{4.70}$$

4.9 Reinforcement of the Separability Postulates Through the Differential Criterion Function

The *differential* criterion function $\Psi(\mathbf{w})$ similar to the perceptron function $\Phi(\mathbf{w}, \theta)$ (25) can be used for the purpose of finding such vector of parameters $\mathbf{w}_i$, which fulfil in a best manner (fully or partly) the inequalities (4.64–4.67) [6, 9]. The criterion functions $\Psi(\mathbf{w})$ is a positive combination of the penalty functions ${\pi_{jj'}}^+(\mathbf{w})$, ${\pi_{jj'}}^-(\mathbf{w})$ and ${\pi_{jj'}}^0(\mathbf{w})$ defined on the *differential vectors* $\mathbf{r}_{jj'}$:

$$(\forall(j, j') \in I_c \cup I_m)\mathbf{r}_{jj'} = \mathbf{x}_{j'} - \mathbf{x}_j \tag{4.71}$$

where $\mathbf{x}_j(k) \ne \mathbf{x}_{j'}(k')$ and $j' > j$.

The *CPL* penalty functions ${\pi_{jj'}}^+(\mathbf{w})$, ${\pi_{jj'}}^-(\mathbf{w})$ and ${\pi_{jj'}}^0(\mathbf{w})$ is defined in a similar manner to ${\upsilon_j}^+(\mathbf{w},\ \theta)$ (4.23) and ${\upsilon_j}^-(\mathbf{w},\ \theta)$ (4.24)

$$
\begin{aligned}
&(\forall (j,j') \in I_{\mathrm{m}}{}^{+}) && && && (4.72)\\
&\qquad \rho_{\mathrm{m}}(j,j') - \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \leq \quad \rho_{\mathrm{m}}(j,j')\\
&\psi_{\mathrm{jj}'}{}^{+}(\mathbf{w}) = \\
&\qquad 0 && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} > && \rho_{\mathrm{m}}(j,j')
\end{aligned}
$$

where $\rho_{\mathrm{m}}(j,j') = \rho_{\mathrm{m_i}}(j,j')$ (4.67). The penalty functions $\pi_{\mathrm{jj}'}{}^{+}(\mathbf{w})$ are aimed at reinforcement the inequalities (4.67).

$$
\begin{aligned}
&(\forall (j,j') \in \quad I_{\mathrm{m}}{}^{-}) && && && (4.73)\\
&\qquad \rho_{\mathrm{mi}}(j,j') + \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \geq -\rho_{\mathrm{m}}(j,j')\\
&\psi_{\mathrm{jj}'}{}^{-}(\mathbf{w}) = \\
&\qquad 0 && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && < -\rho_{\mathrm{m}}(j,j')
\end{aligned}
$$

The penalty functions $\pi_{\mathrm{jj}'}{}^{-}(\mathbf{w})$ are aimed at reinforcement the inequalities (4.68).

$$
\begin{aligned}
&(\forall (j,j') \in \quad I_{\mathrm{c}}) && && && (4.74)\\
&\qquad -\rho_{\mathrm{c}}(j,j') - \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \geq -\rho_{\mathrm{c}}(j,j')\\
&\psi_{jj'}{}^{0}(\mathbf{w}) = 0 && \textit{if} && -\rho_{\mathrm{c}}(j,j') < \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && < \quad \rho_{\mathrm{c}}(j,j')\\
&\qquad -\rho_{\mathrm{c}}(j,j') + \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \textit{if} && \mathbf{w}^{\mathrm{T}}\mathbf{r}_{\mathrm{jj}'} && \geq \quad \rho_{\mathrm{c}}(j,j')
\end{aligned}
$$

where $\rho_{\mathrm{c}}(j,j') = \rho_{\mathrm{c_i}}(j,j')$ (separability postulate). The penalty functions $\pi_{\mathrm{jj}'}{}^{0}(\mathbf{w})$ are aimed at reinforcement the inequalities (separability postulate).

The criterion function $\Psi(\mathbf{w})$ is the weighted sum of the above penalty functions

$$
\begin{aligned}
\Psi(\mathrm{w}) = \; & \Sigma\gamma_{\mathrm{jj}'}\psi_{\mathrm{jj}'}{}^{+}(\mathbf{w}) + \Sigma\gamma_{\mathrm{jj}'}\psi_{\mathrm{jj}'}{}^{-}(\mathbf{w}) + \Sigma\gamma_{\mathrm{jj}'}\psi_{\mathrm{jj}'}{}^{0}(\mathbf{w})\\
& (j,j')\in \mathrm{I}_{\mathrm{m}}{}^{+} \quad (j,j')\in I_{\mathrm{m}}{}^{-} \quad (j,j') \in I_{\mathrm{c}} \qquad (4.75)
\end{aligned}
$$

where $\gamma_{\mathrm{jj}'}$ ($\gamma_{\mathrm{jj}'} > 0$) are positive parameters (*prices*) related to particular dipoles $(\mathbf{x}_{\mathrm{j}}(k), \mathbf{x}_{\mathrm{j}'}(k'))$.

The criterion function $\Psi(\mathbf{w})$ belongs to the family of the convex and piecewise linear (CPL) criterion functions.

The function $\Psi(\mathbf{w})$ (4.72) can be specified as the criterion function $\Psi_{\mathrm{i}}(\mathbf{w}_{\mathrm{i}})$ linked to the ith axis ($i = 1, \ldots, n'$) of the transformed space. The specification of the criterion function $\Psi_{\mathrm{i}}(\mathbf{w}_{\mathrm{i}})$ to the ith axis is done through an adequate choice of the function parameters. The sets of dipoles I_{ci} (separability postulate), $I_{\mathrm{mi}}{}^{+}$ (4.67) and $I_{\mathrm{mi}}{}^{-}$ (4.68) and the sets of margins $\rho_{\mathrm{c_i}}(j,j')$ (4.64) and $\rho_{\mathrm{m_i}}(j,j')$ ((4.65), separability postulate) can be specified in a different manner for particular axis.

Minimization of the function $\Psi_{\mathrm{i}}(\mathbf{w})$ allows one to find the parameters vector $\mathbf{w}_{\mathrm{i}}{}^{*}$, which defines (4.6) the ith column of the transformation matrix $\boldsymbol{W}$ (4.10) or the ith axis of the ($i = 1, \ldots, n'$) of the transformed space.

$$
\Psi_{\mathrm{i}}{}^{*} = \Psi_{i}(\mathrm{w}_{\mathrm{i}}{}^{*}) = \min_{\mathrm{w}} \Psi_{\mathrm{i}}(\mathrm{w}) \geq 0 \qquad (4.76)
$$

The transformation (4.10) defined by the optimal vectors $\mathbf{w_i}^*$ (4.76) will be called as the *dipolar* one. The basis exchange algorithms allow to find the minimal value $\Psi_i{}^*$ of the criterion function $\Psi_i(\mathbf{w})$ in an efficient manner [5]. It can be proved that the minimal value $\Psi_i{}^*$ is equal to zero ($\Psi_i{}^* = 0$) if and only if all the inequalities (separability postulate, (4.67), and (4.68)) can be fulfilled on some line $y = (\mathbf{w})^T\mathbf{x}$. In this case, all the inequalities (separability postulate, (4.67), and (4.68)) are fulfilled on the optimal line $y = (\mathbf{w_i}^*)^T\mathbf{x}$.

Example 4. Let us examine such dipolar transformation (4.10) of the feature vectors $\mathbf{x}_j(k)$ on the visualising plane ($n' = 2$), which fulfil both the inequalities (4.64) and (4.65) or the separability postulate. There is a structural difference between the separability inequalities (4.64) and (4.65). All the inequalities (4.64) should be realised by both the axes $\mathbf{w}_1{}^*$ and $\mathbf{w}_2{}^*$ of the visualising plane. Realisation of each inequality (4.65) by only one axis $\mathbf{w}_1{}^*$ or $\mathbf{w}_2{}^*$ is sufficient for fulfilling of the separability postulate. In other words, if the length $\delta_y(j, j')$ (4.64) of the mixed dipole $(\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$ along one axis $\mathbf{w_i}^*$ is greater than $\rho_m(j, j')$, then the length of this dipole on the plane is also greater then $\rho_m(j, j')$. The length $\delta_y(j, j')$ of the mixed dipole $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$ along the ith axis is greater then $\rho_m(j, j')$ (4.65) if and only if one of the inequalities (4.67) or (4.68) is fulfilled by the optimal vector $\mathbf{w_i}^*$. In a consequence, the indices (j, j') of such mixed dipoles which are sufficiently long on the first axis $\mathbf{w}_1{}^*$ cannot be considered on the second axis $\mathbf{w}_2{}^*$. In a result, the indices (j, j') from the set I_m could be divided along two axis of the visualising plane. Such division reduces the sets I_{m1} and I_{m2} of mixed dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k')\}$ considered on particular axis and, in result, increases chance for fulfilling all the inequalities (separability postulate, (4.67), and (4.68)) on the optimal line $y = (\mathbf{w_i}^*)^T\mathbf{x}$.

To realise the inequality (4.64) for the clear dipole $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k)\}((j, j') \in I_c)$, both the axes $\mathbf{w}_1{}^*$ and $\mathbf{w}_2{}^*$ of the visualising plane should produce small enough lengths $|((\mathbf{w_i}^*)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k))|$ (separability postulate). If the vectors $\mathbf{w}_1{}^*$ and $\mathbf{w}_2{}^*$ are orthogonal and the first vector $\mathbf{w}_1{}^*$ produces the length $|(\mathbf{w}_1{}^*)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k))| \le \rho_c{}^2(j, j')$, then the second vector $\mathbf{w}_2{}^*$ should fulfill the below condition (4.64)

$$|(\mathrm{w}_2{}^*)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k))| \le \rho_c{}^2(j, j') - |(\mathbf{w}_1{}^*)^T(\mathbf{x}'_j(k) - \mathbf{x}_j(k))| \qquad (4.77)$$

To fulfill the separability postulate, the second vector $\mathbf{w}_2{}^*$ should produce such length $|(\mathbf{w}_2{}^*)^T(\mathbf{x}_{j'}(k) - \mathbf{x}_j(k))|$, which is small enough in accordance with (4.77).

The linear transformations (4.10) based on the dipolar model can be used in designing diagnostic maps. Such maps give possibility to enhance clusters of points $\mathbf{y}_j(k)$ on the visualizing plane. The number L of clusters (fields) on the diagnostic map can be equal or greater than the number K of the classes ω_k. The number L of clusters on the diagnostic map can be equal to the number of classes K, if all the feature vectors $\mathbf{x}_j(k)$ from each set C_k (4.1) are used in

producing clear dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k)\}$. It means that the set I_{ck} (separability postulate) contains all clear dipoles based on the set C_k (4.1).

$$I_{ck} = \{(j, j') : (j' > j) \wedge (\mathbf{x}_j(k) \in C_k) \wedge (\mathbf{x}_{j'}(k) \in C_k)\} \tag{4.78}$$

and

$$I_c = I_{c1} \cup \ldots\ldots\ldots \cup \quad I_{ck} \tag{4.79}$$

In this case, the dipolar transformation similarly as the Fishers one is aimed at focussing all the transformed points $\mathbf{x}_j(k)$ from each set C_k (class ω_k) into one cluster on the map under condition of preserving the classes ω_k separability.

In some cases, given set C_k has its own internal structure and it could be profitable to enhance such structure by dividing this set into more than one cluster. For this purpose, the set I_{ck} (4.78) can be modified in a below manner:

$$I_{ck} = \{(j, j') : (j' > j) \wedge (\mathbf{x}_j(k) \in C_k) \wedge (\mathbf{x}_{j'}(k) \in C_k) \wedge (\delta_x(j, j') \leq \rho_0)\} \tag{4.80}$$

where $\delta_x(j, j')$ is the dipol length (4.62) and ρ_0 is a 'small' parameter.

As it results from the relation (4.80), the set I_{ck} contains the indices (j, j') of only such clear dipoles $\{\mathbf{x}_j(k), \mathbf{x}_{j'}(k)\}$ which are 'short'.

4.10 CPL Criterion Functions with Feature Costs

Data sets C_k (4.1) used in decision support systems are often multidimensional. Many features (attributes) x_i are used for description of particular objects $\mathbf{x}_j(k)$. A large part of these features x_i can be unimportant or redundant in decision support rules. Such features should be removed in accordance with one of feature selection procedures.

The feature selection procedure can be based on the CPL criterion functions with feature costs. Let us introduce for this purpose the modified perceptron criterion function $\Phi(\mathbf{w}, \theta)$ (4.25) and the modified differential criterion function $\Psi(\mathbf{w})$ (4.75). The modified perceptron function $\Phi_\lambda{}'(\mathbf{w}, \theta)$ can have the following form:

$$\begin{aligned} \Phi_\lambda\prime(\mathbf{w}, \theta) &= \Phi(\mathbf{w}, \theta) + \lambda \Sigma \gamma_i \phi_i(\mathbf{w}, \theta) \\ &\qquad\qquad i \in \{0, 1, \ldots, n\} \\ = \Sigma \alpha_j \varphi_j{}^+(\mathbf{w}, \theta) &+ \Sigma \alpha_j \varphi_j{}^-(\mathbf{w}, \theta) + \lambda(\Sigma \gamma_i |w_i| + \gamma_0 \\ j \in J^+ \quad j \in J^- &\qquad i \in \{1, \ldots, n\} \end{aligned} \tag{4.81}$$

where $\alpha_j \geq 0$, $\lambda \geq 0$, $\gamma_i > 0$, $\mathbf{w} = [w_1, \ldots, w_n]^T$ is the weight vector, the function $\Phi(\mathbf{w}, \theta)$ is defined by the formula (4.25), and the cost functions $\phi_i(\mathbf{w}, \theta)$ are equal to modulus $|w_i|$ of particular weights w_i.

Minimization of the CPL criterion function $\Phi_\lambda{}'(\mathbf{w}, \theta)$ (4.81) allows to find the optimal parameters $\mathbf{w}^*$ and θ^*.

$$\Phi_{\lambda}^{*} = \Phi_{\lambda}'(w^{*}, \theta^{*}) = \min_{w,\theta} \Phi_{\lambda}'(w, \theta) \quad \geq 0 \tag{4.82}$$

Minimum of the function $\Phi_{\lambda}{}'(\mathbf{w}, \theta)$ (4.81) can be found by using the basis exchange alghorithm.

It can be shown that in the case of linearly separable (4.4) sets G^{+} and G^{-} (4.22), the minimal value $\Phi_{\lambda}{}^{*}$ of the criterion function $\Phi_{\lambda}{}'(\mathbf{w}, \theta)$ (4.81) with sufficient small value of the parameter $\lambda (0 < \lambda < \lambda_{g})$ is equal to

$$\Phi_{\lambda}^{*} = \Phi_{\lambda}'(\mathbf{w}^{*}, \theta^{*}) = \lambda(\Sigma \quad \gamma_{i}|w_{i}^{*}| + \gamma_{0}|\theta|) \quad > 0 \tag{4.83}$$

and (4.25)

$$\Phi(w^{*}, \theta^{*}) = 0 \tag{4.84}$$

The optimal parameters $\mathbf{w}^{*}$ and θ^{*} give a balance between an *increasing tendency* resulting from the penalty functions $\upsilon_{j}{}^{+}(\mathbf{w}, \theta)$ (4.23) and $\upsilon_{j}{}^{-}(\mathbf{w}, \theta)$ (4.24) and *decreasing tendency* resulting from the cost functions $\phi_{i}(\mathbf{w}, \theta)$ (4.81). An influence of the cost functions $\phi_{i}(\mathbf{w}, \theta)$ (4.81) decreases with the value of the parameter λ.

The feature selection rules can be based on the optimal parameters $\mathbf{w}^{*} = [w_{1}{}^{*}, \ldots, w_{n}{}^{*}]^{T}$ (4.82):

$$\textit{if} \quad w_{i}^{*} = 0 \quad \textit{then} \quad \textit{theith feature } x_{i} \textit{ can be neglected}, \tag{4.85}$$

or/

$$\textit{if} \quad |w_{i}{}^{*}| \quad < \quad \varepsilon \quad \textit{then} \quad \textit{theithfeature } x_{i} \textit{ can be neglected}, \tag{4.86}$$

where ϖ is a small parameter.

The differential criterion function $\Psi(\mathbf{w})$ (4.75) can be also modified by adding the cost functions $\phi_{i}(\mathbf{w}, \theta)$ in a manner similar to (4.81). The feature selection rules similar to (4.85) and (4.86) can be based on the modified differential function $\Psi(\mathbf{w})$ (4.75).

4.11 Concluding Remarks

Similarity measures for the *case based reasoning* scheme of decision support can be induced through separable data transformations. In particular, linear transformations of data sets corresponding to particular categories allow to reduce dimensionality of the data sets under the condition of preserving the categories separability. Separable linear transformations can be designed both through solutions of eignevalue problems used in the *principal componet analysis* or in the *discriminant analysis* [4] as well as through minimization of the convex and piecewise linear (CPL) criterion functions [9]. The *perceptron* and the *differential* criterion functions belong, among others, to the CPL family.

Functions from the CPL family give possibility for flexible modelling and solving many problems of exploratory data analysis [9]. In particular, the feature selection problem can be solved through minimization of the CPL criterion functions. The basis exchange algorithms, which are similar to the linear programming, allow one to find the minimum of the CPL criterion functions efficiently even in the case of large, multidimensional data sets [7].

References

1. P. Perner, *Data Mining on Multimedia Data*, Springer, Berlin 2002
2. K. Fukunaga: *Statistical Pattern Recognition*, Academic Press, Inc., San Diego, 1990
3. O.R. Duda, P.E. Hart: *Pattern Classification*, Sec.. Edition. J. Wiley, New York, 2001
4. R.A. Johnson, D.W. Wichern: *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New York, 1991
5. Bobrowski L., Topczewska M., Improving the *K-NN* classification with the Euclidean distance through linear data transformations, pp. 22–32 in: *ICDM* 2004, Eds. P. Perner et al., Lipsk, Germany, Springer Verlag 2004, *Springer Lecture Notes in Artificial Intelligence* 3275
6. L. Bobrowski, M. Topczewska: "Linear visualising transformations and convex, piecewise linear criterion functions", *Bioc. and Biom. Eng.*, pp. 69–78, Vol. 22, Nr.1, 2002
7. L. Bobrowski: "Design of piecewise linear classifiers from formal neurons by some basis exchange, *Pattern Recognition*, 24(9), pp. 863–870, 1991
8. L. Bobrowski, H. Wasyluk, Diagnosis supporting rules of the *Hepar* system, pp. 1309–1313 in: *MEDINFO* 2001, Eds: V. L. Petel, R. Rogers, R. Haux, IOS Press, Amsterdam 2001
9. Bobrowski L.: *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych* (*Data mining based on convex and piecewise linear (CPL) criterion functions)* (in Polish), Technical University Białystok, 2005