Ranked Modeling of Liver Diseases Sequence

Leon Bobrowski[1,2], **Tomasz Łukaszuk**[1], **Hanna Wasyluk**[3]

[1]Białystok Technical University, Faculty of Computer Science, Poland,
[2]Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland,
[3]Medical Center of Postgraduate Education, Warsaw, Poland

Summary: Ranked model in the form of linear transformation of multivariate feature vectors on a line can reflect a causal order between liver diseases. A priori medical knowledge about order between liver diseases and clinical data sets has been used in the definition of the convex and piecewise linear (CPL) criterion function.The linear ranked transformations have been designed here through minimization of such CPL criterion functions.

Keywords: sequential patterns, ranked linear transformations, convex and piecewise linear (CPL) criterion functions, linear separability of data sets, sequence of liver diseases

1. Introduction

Discovering regularities in multivariate data sets or databases is one of the main goals of exploratory data analysis and pattern recognition methods [1], [2]. Discovering trends in temporal databases is a particularly interesting problem with many important applications.

An adequate standardization is usually needed before applying data analysis tools. In a standardized clinical data representation, patients are represented in the form of feature vectors with the same number of numerical components (features) or as points in a multidimensional feature space. A particular pattern in clinical data is represented as a distinct set of feature vectors or as a special cloud of points in a feature space.

The regression analysis plays a prominent role in data exploration [3]. The regression model may describe a dependence of one feature on a selected set of other features. The ranked regression method can also serve a similar purpose [4], [5], [6]. The ranked models are particularly useful when values of the dependent feature cannot be measured precisely or directly and additional information about feature vectors is available only in the form of ranked relations within selected pairs of these vectors. Such ranked relations can be treated as a priori knowledge about linear sequential patterns hidden in data. In this context, inducing the linear ranked model from the ranked pairs can be treated as a pattern recognition problem. The induced ranked model can also be used for prognosis or decision support purposes.

The method of inducing linear ranked models from a set of feature vectors and ranked relations within selected pairs of these vectors was proposed in the previous papers [4], [5]. This method is based on the minimization of convex and piecewise-linear (*CPL*) criterion functions. Properties of this approach in the context of modeling a causal sequence of liver diseases are analysed in the presented paper. Feature vectors from hepathological database of the system Hepar and additional medical knowledge in the form of a causal sequence of liver diseases were used in designing ranked linear transformation [7].

2. Feature vectors and oriented dipoles

We are taking into consideration a data set *C* built from *m* feature vectors $\mathbf{x}_j=[x_{j1},...x_{jn}]^T$ which were numbered in a fixed manner

$$C = \{\mathbf{x}_j\} \; (j = 1,\ldots,m) \qquad (1)$$

The vectors $\mathbf{x}_j$ belong to the *n*-dimensional feature space *F*[*n*] ($x_j \in F[n]$). The component (feature) x_{ji} of the vector $\mathbf{x}_j$ is a numerical result of the *i*-th examination (*i*=1,...,*n*) of a given patient or event O_j (*j*=1,...,*m*). The feature vectors $\mathbf{x}_j$ are of a mixed type if they represent different types of diagnostic measurements ($x_i \in \{0,1\}$) or ($x_i \in R$)).

Let the symbol "≺" means the relation "*follows*" which is fulfilled within ranked pairs {$\mathbf{x}_j$, $\mathbf{X}_j$'} (*j*<*j*') of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_j$' with the indices (*j*, *j*') from some set :

$$(\forall (j,j') \in J) \quad \mathrm{x}_j \prec \mathrm{x}_{j'} \Leftrightarrow (\mathrm{x}_{j'} \;\; \mathit{follows} \;\; \mathrm{x}_j)$$
$$\mathit{or} \quad \mathrm{x}_{j'} \prec \mathrm{x}_{j'} \Leftrightarrow (\mathrm{x}_j \;\; \mathit{follows} \;\; \mathrm{x}_{j'}) \qquad (2)$$

The relation $\mathbf{x}_j \prec \mathbf{x}_j$' between the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_j$' means that the vector $\mathbf{x}_j$' follows the vector $\mathbf{x}_j$ in some sequence. This relation should be determined on the basis of some additional information about set of pairs (not necessary all) of the vectors $\mathbf{x}_j$. For example, medical doctors can compare two patients with the same disease and decide that one of them is in a more serious condition. As another example, it can be stated that one of two students is more talented than another one.

In the paper we analyse the problem of designing such transformations of the feature vectors $\mathbf{x}_j$ on the (ranked) line $y=\mathbf{w}^T\mathbf{x}$ which preserve the relation "≺" (2) as precisely as possible

$$y_j = y_j(\mathbf{w}) = \mathbf{w}^T \mathbf{x}_j \,, \qquad (3)$$

where $\mathbf{w}=[w_1,...,w_n]^T$ is the weight vector.

The family of the relations (2) defines the sequential pattern *S*($\mathbf{x}$) of the vectors x_j in the feature space *F*[*n*] ($\mathbf{x}_j \in F[n]$).

Definition 1: The sequential pattern *S*($\mathbf{x}$) is linear in the feature space *F*[*n*] if and only if there exists such *n*-dimensional weight vector $\mathbf{w}$ ($\mathbf{w} \in R^n$) that the below implication holds:

$$\begin{array}{ll}(\forall (j,j') \in J) & \mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow \mathbf{w}^T\mathbf{x}_j < \mathbf{w}^T\mathbf{x}_{j'} \\ \quad and & \mathbf{x}_{j'} \prec \mathbf{x}_j \Rightarrow \mathbf{w}^T\mathbf{x}_{j'} < \mathbf{w}^T\mathbf{x}_j\end{array} \tag{4}$$

where J is a set of indices (j, j') of ranked pairs of vectors $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$.

Definition 2: The ranked pair $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ constitutes the *positively oriented dipole* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $((j,j' \in J^+)$, if and only if $\mathbf{x}_j \prec \mathbf{x}_{j'}$.

$$(\forall (j,j') \in J^+) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \tag{5}$$

Definition 3: The ranked pair $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ constitutes the *negatively oriented dipole* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $((j,j' \in J^-)$, if and only if $\mathbf{x}_{j'} \prec \mathbf{x}_j$.

$$(\forall (j,j') \in J^-) \quad \mathbf{x}_{j'} \prec \mathbf{x}_j \tag{6}$$

Definition 4: The line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked in accordance with the dipoles $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ orientations if and only if

$$(\forall (j,j') \in J^+) \quad y_j(\mathrm{w}) < y_{j'}(\mathrm{w}) \quad and \quad (\forall (j,j') \in J^-) \quad y_j(\mathrm{w}) > y_{j'}(\mathrm{w}) \tag{7}$$

where J^+ and J^- are sets of indices (j, j') of the positively and negatively oriented dipoles $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$, and $J^+ \cup J^- = J$, $J^+ \cap J^- = \emptyset$.

3. Designing ranked models through minimization of a CPL criterion function

Let us introduce the positive set R^+ and the negative set R^- of the differential vectors $\mathbf{r}_{jj'}=(\mathbf{x}_{j'}-\mathbf{x}_j)$ on the basis of the sets of indices J^+ (6) and J^- (7).

$$\begin{array}{l}R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^+\} \\ R^- = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^-\}\end{array} \tag{8}$$

We examine the possibility of separating the sets R^+ and R^- by the hyperplane $H(\mathbf{w})$ passing through the origin 0 of the feature space.

$$H(\mathbf{w}) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\} \tag{9}$$

Definition 5: The sets R^+ and R^- (8) are separable by some hyperplane $H(\mathbf{w})$ (9) if and only if the below inequalities hold

$$(\exists \mathrm{w}) \quad \begin{array}{ll}(\forall (j,j') \in J^+) & \mathbf{w}^T\mathbf{r}_{jj'} > 0 \\ (\forall (j,j') \in J^-) & \mathbf{w}^T\mathbf{r}_{jj'} < 0\end{array} \tag{10}$$

If all the above inequalities are fulfilled for some vector $\mathbf{w}$, then the hyperplane $H(\mathbf{w})$ (9) separates the sets R^+ and R^- (8).

Lemma 1: The linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked (7) in accordance with dipoles' $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ orientations if and only if the hyperplane $H(\mathbf{w})$ (9) separates (10) the sets R^+ and R^- (8).

Proof: If the hyperplane $H(\mathbf{w})$ (9) separates the sets R^+ and R^- (8), then the ranked relations (7) hold on the line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3). On the other hand, the fulfilling of the ranked relations (7) guarantees that the inequalities (10) hold.

Designing a separating hyperplane $H(\mathbf{w})$ (9) could be carried out through the minimization of the convex and piecewise linear (*CPL*) criterion function similar to the perceptron criterion function $\Phi(\mathbf{w})$ [2]. For this purpose let us introduce the positive $\varphi_{jj'}^+(\mathbf{w})$ and the negative $\varphi_{jj'}^-(\mathbf{w})$ penalty functions:

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathrm{w}) = \begin{array}{lll}1+\mathbf{w}^T\mathbf{r}_{jj'} & if & \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if & \mathbf{w}^T\mathbf{r}_{jj'} \leq -1\end{array} \tag{11}$$

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathrm{w}) = \begin{array}{lll}1+\mathbf{w}^T\mathbf{r}_{jj'} & if & \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if & \mathbf{w}^T\mathbf{r}_{jj'} \leq -1\end{array} \tag{12}$$

The criterion function $\Phi(w)$ is the weighted sum of the penalty functions $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$:

$$\Phi(\mathbf{w}) = \sum_{(j,j') \in J^+} \gamma_{jj'} \varphi_{jj'}^+(\mathbf{w}) + \sum_{(j,j') \in J^-} \gamma_{jj'} \varphi_{jj'}^-(\mathbf{w}) \tag{13}$$

where $\gamma_{jj'}$ $(\gamma_{jj'}>0)$ is a positive parameter (*price*) related to the dipole $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$.

$\Phi(\mathbf{w})$ (13) is the convex and piecewise linear (*CPL*) function as the sum of the penalty functions $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$ of the same kind. The basis exchange algorithms, similar to the linear programming, allow to find the minimum of such function efficiently, even in the case of large, multidimensional data sets R^+ and R^- [7]:

$$\Phi^* = \Phi(\mathbf{w}^*) = \min_{\mathbf{w}} \Phi(\mathbf{w}) \geq 0 \tag{14}$$

The optimal parameter vector $\mathbf{w}^*$ and the minimal value Φ^* of the criterion function $\Phi(\mathbf{w})$ (13) can be applied to a variety of data ranking problems. In particular, the vector $\mathbf{w}^*$ defining the best ranked line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) can be found this way.

The minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) can be used to measure the linearity of the sequential patterns $S(\mathbf{x})$ (Def. 1) in a given feature space $F[n]$.

Lemma 2: The minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) is equal to zero if and only if the sequential pattern $S(\mathbf{x})$ (Def. 1) is linear.

Proof: If there exists such a vector $\mathbf{w}^*$ that the ranking of the points $y_j(\mathbf{w}^*)$ on the line (3) is fully consistent (7) with the relations "≺", then the sets R^+ and R^- (8) can be separated (10) by the hyperplane $H(\mathbf{w}^*)$ (9). In this case, the minimal value Φ^* of the perceptron criterion function $\Phi(\mathbf{w})$ (13) is equal to zero, as it results from the pattern recognition theory [1]. On the other hand, if the minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) is equal to zero in the point $\mathbf{w}^*$, then the values $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$ of all the penalty functions (11) and (12) have to be equal to zero. This means that the sets R^+ and R^- (8) can be separated (10) by the hyperplane $H(\mathbf{w}^*)$ (9). As a result, the ranking of the points $y_j(\mathbf{w}^*)$ on the line (3) is fully consistent (7) with the relations (4) and (5).

4. Causal sequence of learning sets

Let us assume that a clinical database contains descriptions of m patients $O_j(k)$ $(j=1,...m)$ labeled in accordance with their clinical diagnosis $\omega_k(k=1,...K)$. Each patient $O_j(k)$ is represented by n-dimensional feature vector $\mathbf{x}_j(k)$. The feature vector $\mathbf{x}_j(k)$ represents the j-th patient $O_j(k)$ linked to the k-th disease ω_k. The learning set C_k contains m_k labeled feature vectors $\mathbf{x}_j(k)$ that are linked to the k-th disease (class) ω_k.

$$C_k = \{x_j(k)\} \quad (j \in I_k) \tag{15}$$

where I_k is the set of indices j of m_k feature vectors $\mathbf{x}_j(k)$ labeled to the class ω_k.

We assume that the learning set C_k have been formed in a learning causal sequence:

$$C_1 \rightarrow C_2 \rightarrow \ldots \rightarrow C_{K-1} \rightarrow C_K \tag{16}$$

where symbol "$C_{k-1} \rightarrow C_k$" means that "disease ω_k appears after ω_{k-1}" or "disease ω_{k-1} is a cause of ω_k". The consistent *indexing* of the sets C_k and the diseases ω_k has been used in the sequence (16). This means that:

$$(\forall k, k' \in \{1, \ldots, K\}) \quad (k < k') \Rightarrow (C_k \rightarrow C_{k'}) \tag{17}$$

The causal relation "$C_k \rightarrow C_{k'}$" (17) between learning sets C_k and $C_{k'}$ can be used for determining the causal ranked relation "≺" (2) between feature vectors $\mathbf{x}_j(k)$ ($\mathbf{x}_j(k) \in C_k$) and $x_{j'}(k')$ ($x_{j'}(k) \in C_{k'}$) (15):

$$(\forall k, k' \in \{1, \ldots, K\}) \quad (C_k \rightarrow C_{k'}) \Rightarrow ((\forall \mathbf{x}_j(k) \in C_k) \; and (\forall \mathbf{x}_{j'}(k') \in C_{k'})) \; \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k') \tag{18}$$

or

$$(\forall k, k' \in \{1, \ldots, K\}) \quad (k < k') \Rightarrow (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \tag{19}$$

Let us remark that there is no ranked relation "≺" (2) between feature vectors $\mathbf{x}_j(k)$ and $x_{j'}(k)$ from the same set .

We can assume that the indices j of the feature vectors $\mathbf{x}_j(k)$ are consistent with the learning sets C_k (15). This means that the set C_1 contains m_1 first feature vectors $\mathbf{x}_j(k)$, the set C_2 contains m_2 next vectors $\mathbf{x}_j(k)$, and so on. As a consequence, the following relation of consistent indexing holds:

$$(\forall \mathbf{x}_j(k) \in C_k) and (\forall \mathbf{x}_{j'}(k') \in C_{k'}) \quad (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \Rightarrow (j < j') \tag{20}$$

Lemma 3: In the case of consistent indexing (20), the linear transformation $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$ (3) is completely ranked (7) if and only if the set R^+ (8) of the differential vectors $\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{X}_j)$ is situated on the positive side of some hyperplane $H(\mathbf{w})$ (9):

$$(\exists \mathbf{w})(\forall \mathbf{r}_{jj'} \in R^+) \quad \mathbf{w}^T \mathbf{r}_{jj'} > 0 \tag{21}$$

Proof: The relations (19) and (20) guarantee that all dipoles {$\mathbf{x}_j$, $\mathbf{x}_{j'}$} are positively oriented (*Def. 2*) and the negative set R^- (8) is empty (R^-=ø). As a result, the vector **w** defines such hyperplane $H(\mathbf{w})$ which linearly separates (10) the sets R^+ and R^-(8). This means that the assumptions of the *Lemma 1* are fulfilled.

Let us assume that the set R^+ (8) contains <u>all</u> positively oriented dipoles {$\mathbf{x}_j(1)$, $\mathbf{x}_{j'}(2)$}($j<j'$) (5), that can be generated from two learning sets C_1 and C_2 (15) in accordance with the relation $C_1 \rightarrow C_2$ (16) and consistent indexing (20),

$$R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'}(2) - \mathbf{x}_j(1)) : \mathbf{x}_j(1) \in C_1 \quad and \quad \mathbf{x}_{j'}(2) \in C_2\} \tag{22}$$

The set R+ (22) is complete if it contains all positively oriented dipoles {$\mathbf{x}_j(1)$, $\mathbf{x}_{j'}(2)$}(5) that can be created from two learning sets C_1 and C_2 (15) with the relation $C_1 \rightarrow C_2$ (16).

Definition 6: Two learning sets C_1 and C_2 (15) are linearly separable if and only if the below inequalities hold

$$(\exists \mathbf{w}, \theta) \quad \begin{array}{ll} (\forall \mathbf{x}_{j'} \in C_2) & \mathbf{w}^T \mathbf{x}_{j'} > \theta \\ (\forall \mathbf{x}_j \in C_1) & \mathbf{w}^T \mathbf{x}_{j'} < \theta \end{array} \tag{23}$$

where θ (θ ∈ R^1) is a *threshold*.

The above parameters ($\mathbf{w}$,θ) define the hyperplane $H(\mathbf{w},\theta)$ in the feature space, where:

$$H(\mathbf{w}, \theta) = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} = \theta\} \tag{24}$$

It is possible to relate the linear separability of two learning sets C_1 and C_2 (15) with the complete ranking of the linear transformation $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$ (3).

Theorem 1: The linear transformation $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$ (3) is completely ranked (21) in accordance with the complete set R^+ (22) if and only if there exists such threshold θ' that the hyperplane $H(\mathbf{w},\theta')$ (24) separates two learning sets C_1 and C_2 (15).

Proof: If the line $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$ (3) is fully ranked (7), then

$$(\exists \mathbf{w})(\forall \mathbf{x}_j \in C_1) and (\forall \mathbf{x}_{j'} \in C_2) \quad \mathbf{w}^T \mathbf{x}_{j'} > \mathbf{w}^T \mathbf{x}_j \tag{25}$$

Let us define the positive $\theta^+(\mathbf{w})$ and the negative $\theta^-(\mathbf{w})$ thresholds on the line $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$:

$$\theta^+(\mathbf{w}) = \min\{\mathbf{w}^T \mathbf{x}_{j'} : \mathbf{x}_{j'} \in C_2\} \tag{26}$$

$$\theta^-(\mathbf{w}) = \max\{\mathbf{w}^T \mathbf{x}_j : \mathbf{x}_j \in C_1\} \tag{27}$$

The below inequality results from the relation (25)

$$\theta^+(\mathbf{w}) > \theta^-(\mathbf{w}) \tag{28}$$

The threshold θ' can be defined as follows

$$\theta' = \frac{\theta^+(\mathbf{w}) + \theta^-(\mathbf{w})}{2} \tag{29}$$

It can be directly verified that the hyperplane $H(\mathbf{w},\theta')$ (24) separates the learning sets C_1 and C_2 (15).

On the other hand, the hyperplane $H(\mathbf{w},\theta')$ (24) that separates sets C_1 and C_2 (15), determines a linear transformation $y(\mathbf{w}) = \mathbf{w}^T\mathbf{x}$ (3) which is completely ranked (21). As it results from the definition of the linear separabilty (23), each element $\mathbf{r}_{jj'} = \mathbf{x}_{j'}(2) - \mathbf{x}_j(1)$ of the set R^+(22) fulfills the relation (21).

5. Causal sequence of liver diseases

The database of the system *Hepar* contains descriptions of patients with variety of chronic liver diseases ω_k (k=1,...K) [7]. The feature vectors **x** in the database of *Hepar* are the mixed, qualitative-quantitative type. They contain both symptoms and signs ($x_i \in \{0,1\}$) as well as the numerical results of laboratory tests ($x_i \in R$). About 200 different features x_i describe one patient case in this system.

For the purpose of these computations, each patient has been described by the feature vector $\mathbf{x}_j(k)$ composed of 62 features x_i chosen as a standard by medical doctors. The following K=7 groups of patients C_k (15) have been extracted from the *Hepar* database:

The data sets C_k (30) have been formed as the causal learning sequence (16) in accordance with medical knowledge. The ranked relation "<" (2) between feature vectors $\mathbf{x}_j(k)$ $(\mathbf{x}_j(k) \in C_k)$ and $x_j(k')$ $(x_j(k) \in C_k)$ (15) has been defined (18) on the basis of the causal sequence (16). This ranked relation allowed to define both oriented dipoles $\{x_j, x_j\}$(5) (6) as well as the positive set R^+ and the negative set R^- (8) of the differential vectors $\mathbf{r}_{jj'}(\mathbf{x}_{j'}\text{-}\mathbf{x}_j)$. The sets R^+ and R^- have been used in the definition of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13). The optimal parameter vector $\mathbf{w}^*$ (14) constituting the minimum of the function $\Phi(\mathbf{w})$ (13) defines the ranked linear model (3) that can be used for prognosis purposes:

$$y_j = y_j(\mathbf{w}^*) = (\mathbf{w}^*)^T \mathbf{x}_j = w_1^* x_{j1} + \ldots + w_n^* x_{jn} \tag{31}$$

The solution of the feature selection problem allows to determine the most important features x_i influencing the future of a given patient x_0 and to neglect the unimportant features x_i. The feature selection problem can also be based on the minimization of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13) [7].

The linear model (31) fulfills the ranked relation (4) for a great part of feature vectors $\mathbf{x}_j$:

$$\mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \tag{32}$$

As a result, the causal sequence (16) of the learning sets C_k (30) is preserved in a great part by the ranked model (31). In accordance with the equation (31), each learning set C_k (30) is transformed in the set C_k' of the points $\mathbf{x}_j(k)$ on the ranked line:

$$C_k' = \{y_j(k)\} \quad (j \in I_k) \tag{33}$$

C_1. *Non hepatitis patients*	- 16 patients	
C_2. *Hepatitis acuta*	- 8 patients	
C_3. *Hepatitis persistens*	- 44 patients	
C_4. *Hepatitis chronica activa*	- 95 patients	
C_5. *Cirrhosis hepatitis compensata*	- 38 patients	(30)
C_6. *Cirrhosis decompensata*	- 60 patients	
C_7. *Carcinoma hepatis*	- 11 patients	

	Total: 272 patients	

The sets C_k' can be characterized by mean values μ_k and variances σ_k^2, where

$$\mu_k = \frac{\sum_j y_j(k)}{m_k} \quad (j \in I_k) \tag{34}$$

and

$$\sigma_k^2 = \frac{\sum_j (y_j(k) - \mu_k)}{m_k} \quad (j \in I_k) \tag{35}$$

The results of computations based on the model (31) of data sets C_k (30) are summarized in the below Table 1:

We can remark that the ranked implications (32) do not depend on the linear scaling (36) of the model. This means that

$$(\forall \alpha > 0)(\forall \beta) \quad (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \Rightarrow$$
$$\alpha (\mathbf{w}^*)^T \mathbf{x}_j + \beta < \alpha (\mathbf{w}^*)^T \mathbf{x}_{j'} + \beta \tag{37}$$

The parameters α and β have been fixed through minimization of the sum $Q(\alpha,\beta)$ of the differences $|k\text{-}\alpha(\mathbf{w}^*)^T\mathbf{x}_j(k)+\beta|$ for all the sets C_k (30) and all the feature vector $\mathbf{x}_j(k)$.

Table 1. The mean values μ_k and variances σ_k^2 of the sets C_k' (33).

Data sets C_k' (33)	Number of patients m_k	Mean value μ_k	Variance σ_k^2 (σ_k)
C_1'	16	-1.02	0.46 (0.68)
C_2'	8	-0.58	0.57 (0.76)
C_3'	44	0.12	1.1 (1.05)
C_4'	95	0.89	1.46 (1.21)
C_5'	38	2.11	2 (1.41)
C_6'	60	3.02	2.2 (1.48)
C_7'	11	3.78	0.62 (0.79)

Let us consider an additional linear scaling $y' = \alpha y + \beta$ of the model $y = (\mathbf{w}^*)^T\mathbf{x}$ (31) in order to improve the interpretability of its prognostic applications.

$$y_j'(k) = \alpha (\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \tag{36}$$

where α and β are the scaling parameters.

$$Q(\alpha, \beta) = \sum_{k=1,\ldots,K} \sum_{j \in I_k} \left| k - \alpha (\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \right| \tag{38}$$

where I_k is the set of indices j of the feature vectors $\mathbf{x}_j(k)$ from the set C_k (30).

Let us remark that $\mathcal{Q}(\alpha,\beta)$ is the convex and piecewise linear (*CPL*) function. The basis exchange algorithms also allow to find efficiently the parameters α^* and β^* constituting the minimum of the function $\mathcal{Q}(\alpha,\beta)$. Some results of the scaled model evaluation are shown in the Table 2 and on the Fig. 1.

The linear ranked model $y=\alpha^*(\mathbf{w}^*)^T\mathbf{x}+\beta^*$ can be used in the diagnosis support of a new patient $\mathbf{x}_0$. The location of the point $y_0=\alpha^*(\mathbf{w}^*)^T\mathbf{x}_0+\beta^*$ on the ranked line (30) constitutes a valuable characteristic of the patient and his perspectives. In the case the scaled model (Fig. 1), we can expect that the point y_0 representing a new patient $\mathbf{x}_0$ with the k-th disease ω_k will be situated near the index k.

6. Concluding remarks

The linear ranked models can be induced from data sets C_k (15) on the basis of additional medical knowledge in the form of the causal sequence (16) of diseases ω_k (k=1,...,K). The ranked relation "≺" (2) between feature vectors $\mathbf{x}_j(k)$ from different learning sets C_k and C_k' has been defined (18) on the basis of the causal sequence (16). This ranked relation allowed to define both the oriented dipoles $\{x_j, x_{j'}\}$ (5) (6) as well as the positive set R^+ and the negative set R^- (8) of the differential vectors .

The sets R^+ and R^- (8) have been used in the definition of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13). The optimal parameter vector $\mathbf{w}^*$ (14), which is the minimum point of the function $\Phi(\mathbf{w})$ (13) defines the ranked linear model (31) that can be used for the purpose of prognosis. The prognostic model (31) can be improved through linear scaling (36). An example of the *CPL* criterion function (α,β) for choosing the scaling parameters α and β is provided by the equation (38).

The feature selection problem allows to determine the most important features $\mathbf{x}_i$ influencing significantly the future of a given patient and to neglect unimportant features. The feature selection problem can be solved through the minimization of a modified *CPL* criterion function $\Phi(\mathbf{w})$ (13) [6], [7].

Table 2. The mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (33) obtained from the ranked model (31) after scaling (36) with the optimal parameters α and β*.*

Data sets C_k' (33)	Number of patients m_k	Mean value μ_k'	Variance σ_k^2 (σ_k)
C_1'	16	1.41	0.64 (0.8)
C_2'	8	1.93	0.79 (0.89)
C_3'	44	2.75	1.51 (1.23)
C_4'	95	3.65	1.99 (1.41)
C_5'	38	5.08	2.74 (1.65)
C_6'	60	6.14	3.02 (1.74)
C_7'	11	7.03	0.85 (0.92)

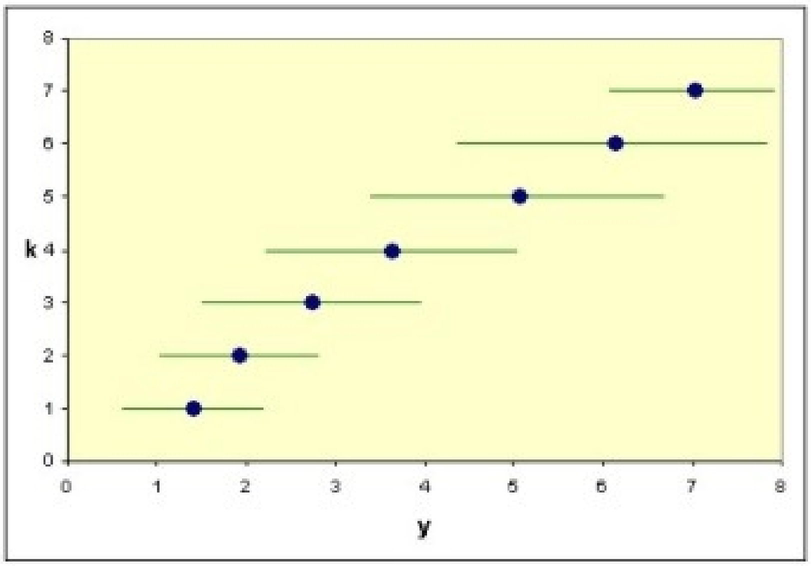

Figure 1. Graphical presentation the mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (33) obtained from the ranked model (31) after scaling (36) with the optimal parameters α and β*.*

The ranked model of liver diseases (31) could be applied in screening procedures in the search for potentially ill patients eligible for further investigations and therapy. The ranked model (31) can also be specified for risk prognosis for individual patients.

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

References

[1] Duda O. R., Hart P. E., Stork D. G.: Pattern Classification, J. Wiley, New York, 2001.

[2] Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press 1972.

[3] Johnson R. A., Wichern D. W.: Applied Multivariate Statistical Analysis, Prentice-Hall Inc., Englewood Cliffs, New York, 1991.

[4] Bobrowski L., Łukaszuk T.: Ranked Linear Modeling in Survival Analysis, pp. 61-67 in: Lecture Notes of the ICB Seminars: Statistics and Clinical Practice, ed. by L. Bobrowski, J. Doroszewski, N. Victor, IBIB PAN, Warsaw, 2005.

[5] Bobrowski L.: Ranked Modelling with Feature Selection Based on the CPL Criterion Functions, in: Machine Learning and Data Mining in Pattern Recognition, eds. P. Perner et al., Lecture Notes in Computer Science vol. 3587, Springer Verlag, Berlin, 2005.

[6] 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), Białystok Technical University, 2005.

[7] Bobrowski L., Wasyluk H.: Diagnosis Support rules of the Hepar system, pp. 1309-1313 in: MEDINFO 2001, eds: V. L. Petel, R. Rogers, R. Haux, IOS Press, Amsterdam, 2001.

[8] Bobrowski L.: Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, Pattern Recognition, 24(9), pp. 863-870, 1991.