Ranked Modeling of Causal Sequences of Diseases for the Purpose of Early Diagnosis*

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

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

Abstract. Medical knowledge in the form of causal sequences of diseases can be used in designing ranked models. Considered ranked models are based on linear transformation of multivariate feature vectors on a line that preserves in a best possible way a causal order between diseases. Clinical data sets from particular diseases supplied with a causal order within pairs of these diseases may be used in the definition of the convex and piecewise linear (*CPL*) criterion function. The linear ranked transformations can be designed through minimization of the *CPL* criterion functions.

Keywords: causal order between diseases, ranked linear transformations, convex and piecewise linear (*CPL*) criterion functions, linear separability of data sets, causal sequence of liver diseases.

1 Introduction

Medical knowledge in the form of a causal sequence of particular diseases ω_k (k = 1,…., K) could be available in some cases. The causal sequence of liver diseases is a good example of such a situation [1]. Clinical databases allow to form *reference* (*learning*) sets C_k for such diseases that are linked in the causal sequence. Patients $O_j(k)$ allocated by medical doctors to the disease ω_k are represented in the form of feature vectors $\mathbf{x}_j(k)$ with the same number n of numerical components (features) or as points in n-dimensional feature space. Each learning sets C_k should be formed by a sufficiently large number of the feature vectors $\mathbf{x}_j(k)$ allocated to the disease ω_k.

Methods of exploratory data analysis or pattern recognition give the possibility of discovering regularities in multivariate data sets or in large databases [2], [3]. Enhancing trends in temporal databases is a particularly interesting problem with many important applications. The regression analysis methods plays a prominent role in data exploration and can be used for trends enhancing and modeling [4].

* 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.

E. Kącki, M. Rudnicki, J. Stempczyńska (Eds.): Computers in Medical Activity, AISC 65, pp. 23–31.
springerlink.com

The regression models describe a dependence of one feature on a selected set of other features and can be used for the purpose of prognosis. The ranked regression models can also serve a similar purpose [5], [6], [7]. The ranked regression 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 [5] [6]. This method was based on the minimization of convex and piecewise-linear (*CPL*) criterion functions. Properties of this approach in the context of modeling of causal sequences of diseases are analyzed in the presented paper. These considerations are illustrated on the example of 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 [1].

2 Ranked Relations between Feature Vectors Based on a Causal Sequence of Diseases

In some cases medical knowledge allows to form a *causal sequence* among regarded diseases ω_k ($k = 1,\ldots, K$):

$$\omega_1 \rightarrow \omega_2 \rightarrow \ldots\ldots \rightarrow \omega_K \tag{1}$$

The symbol "$\omega_i \rightarrow \omega_{i+1}$"in the above sequence means that the given patient disease ω_{i+1} resulted from the disease ω_i, or ω_{i+1} is a consequence of the disease ω_i ($i = 1,\ldots, K$-1). In other words, the disease ω_{i+1} is more *advanced* than ω_i in the process of the disease development in given patient.

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 ω_k ($k = 1,\ldots, K$). Each patient $O_j(k)$ is represented by n -dimensional feature vector $\mathbf{x}_j(k) = \mathbf{x}_j = [x_{j1},.....,x_{jn}]^T$. The vectors $\mathbf{x}_j$ belong to the *feature space* $F[n]$ ($\mathbf{x}_j \in F[n]$). The component x_{ji} ($i = 1,\ldots.., n$) of the vector $\mathbf{x}_j$ is a numerical result of the i-th examination (*feature*) x_i of a given patient O_j ($j = 1,.....,m$). The feature vectors $\mathbf{x}_j$ can be of a mixed type, and represent different types of features (measurements) x_i of a given patient O_j (for example: $x_{ji} \in \{0,1\}$ or $x_{ji} \in R$).

The labeled feature vector $\mathbf{x}_j(k)$ represents such a patient $O_j(k)$, that has been assigned by medical doctors to the k -th disease ω_k. The learning set C_k contains m_k labeled feature vectors $\mathbf{x}_j(k)$:

$$C_k = \{ \mathbf{x}_j(k): j \in J_k \} \tag{2}$$

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

The causal sequence (1) determines the below sequence of the learning sets C_k (2):

$$C_1 \rightarrow C_2 \rightarrow \ldots\ldots \rightarrow C_K \tag{3}$$

The causal sequence (1) allows also to determine the below *ranked relation* "$\prec$" between the feature vectors $\mathbf{x}_j(k)$ ($\mathbf{x}_j(k) \in C_k$) representing patients $O_j(k)$ assigned to particular diseases ω_k:

$$\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k') \Leftrightarrow \mathbf{x}_{j'}(k') \text{ is a more risky than } \mathbf{x}_j(k) \tag{4}$$

We can remark that the feature vectors $\mathbf{x}_{j'}(k')$ from the learning set $C_{k'}$ (1) are *more risky than* the vectors $\mathbf{x}_j(k)$ from the learning set C_k if and only if the disease ω_k appears earlier than $\omega_{k'}$ in the causal sequence (1).

$$\begin{aligned}&(\forall k, k' \in \{1,\ldots, K\})\ (\forall \mathbf{x}_j(k) \in C_k) \text{ and } (\forall \mathbf{x}_{j'}(k') \in C_{k'})\\ &\qquad \text{if } \omega_k \rightarrow \omega_{k'} \text{ then } \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')\end{aligned} \tag{5}$$

3 Ranked Linear Transformations

Let us consider a linear transformation $y = \mathbf{w}^T\mathbf{x}$ of n-dimensional feature vectors $\mathbf{x}_j$ ($\mathbf{x}_j \in R^n$) on the points y_j of the line R^1 ($y_j \in R^1$):

$$(\forall j \in \{1,\ldots\ldots,m\}) \quad y_j = \mathbf{w}^T\mathbf{x}_j \tag{6}$$

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

We are considering a linear transformation (*ranked line*) $y = (\mathbf{w}^*)^T\mathbf{x}$ (6) which preserves the relation "$\prec$" (5) as precisely as possible. It means that the below implication is fulfilled in the ideal case:

$$\begin{aligned}&(\exists \mathbf{w}^* \in R^n)\ (\forall k, k' \in \{1,\ldots, K\})\ (\forall \mathbf{x}_j(k) \in C_k) \text{ and } (\forall \mathbf{x}_{j'}(k') \in C_{k'})\\ &\qquad \text{if } \omega_k \rightarrow \omega_{k'} \text{ then } (\mathbf{w}^*)^T\mathbf{x}_j(k) < (\mathbf{w}^*)^T\mathbf{x}_{j'}(k')\end{aligned} \tag{7}$$

The problem of there being an optimal weight vector $\mathbf{w}^*$ which assures the implication (7) can be linked to the linear separability of the positive set C^+ and the negative set C^- of the differential vectors $\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j)$ [5]:

$$\begin{aligned}C^+ &= \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j): j < j' \text{ and } \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')\}\\ C^- &= \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j): j < j' \text{ and } \mathbf{x}_{j'}(k') \prec \mathbf{x}_j(k)\}\end{aligned} \tag{8}$$

Definition 1: The positive set C^+ and the negative set C^- of are linearly separable if and only if there exists such a weight vector $\mathbf{w}'$ that the below inequalities hold

$$\begin{aligned}&(\exists \mathbf{w}')\ (\forall \mathbf{r}_{jj'} \in C^+)\ (\mathbf{w}')^T\mathbf{r}_{jj'} > 0\\ &\qquad (\forall \mathbf{r}_{jj'} \in C^-)\ (\mathbf{w}')^T\mathbf{r}_{jj'} < 0\end{aligned} \tag{9}$$

The weight vector $\mathbf{w}'$ defines the hyperplane $H(\mathbf{w}')$ in the feature space:

$$H(\mathbf{w}') = \{\mathbf{x}: (\mathbf{w}')^T\mathbf{x} = 0\} \tag{10}$$

The hyperplane $H(\mathbf{w}')$ passes through the point $\mathbf{0}$ in in the feature space. If the inequalities (9) hold, then the hyperplane $H(\mathbf{w}')$ separates the sets C^+ and C^- (9). This means that all elements $\mathbf{r}_{jj'}$ of the set C^+ are located on the positive side of the hyperplane $H(\mathbf{w}')$ and all elements $\mathbf{r}_{jj'}$ of the set C^- are located on the negative side of this hyperplane.

The below *Lemma* can be proved [5]:

Lemma 1: The linear transformation $y = (\mathbf{w}')^T\mathbf{x}$ (6) preserves all the ranked relations "$\prec$" (5) if and only if the hyperplane $H(\mathbf{w}')$ separates the sets C^+ and C^- (9).

The line $y = (\mathbf{w}')^T\mathbf{x}$ (6) is *completely ranked* if and only if the hyperplane $H(\mathbf{w}')$ separates the sets C^+ and C^- (9).

4 Designing Ranked Transformations through Minimization of the CPL Criterion Function

It is known that the minimisation of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ similar to the perceptron criterion function allows to find such a hyperplane $H(\mathbf{w}^*)$ (10) which separates two sets in the best possible manner [1]. In order to define an adequate *CPL* criterion function $\Phi(\mathbf{w})$ let us introduce the positive $\varphi_{jj'}{}^+(\mathbf{w})$ and the negative $\varphi_{jj'}{}^-(\mathbf{w})$ penalty functions [7]:

$$(\forall \mathbf{r}_{jj'} \in C^+)$$

$$\varphi_{jj'}{}^+(\mathbf{w}) = \begin{cases} 1 - \mathbf{w}^T\mathbf{r}_{jj'} & \textit{if } \mathbf{w}^T\mathbf{r}_{jj'} \le 1 \\ 0 & \textit{if } \mathbf{w}^T\mathbf{r}_{jj'} > 1 \end{cases} \tag{11}$$

and

$$(\forall \mathbf{r}_{jj'} \in C^-)$$

$$\varphi_{jj'}{}^-(\mathbf{w}) = \begin{cases} 1 + \mathbf{w}^T\mathbf{r}_{jj'} & \textit{if } \mathbf{w}^T\mathbf{r}_{jj'} \ge -1 \\ 0 & \textit{if } \mathbf{w}^T\mathbf{r}_{jj'} < -1 \end{cases} \tag{12}$$

The criterion function $\Phi(\mathbf{w})$ is the 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 differential vectors $\mathbf{r}_{jj'} = \mathbf{x}_{j'} - \mathbf{x}_j$ from the positive set C^+ ($(j,j')\in J_p^+$) or from the negative set C^- (8) ($(j,j')\in J_p^-$).

$\Phi(w)$ (13) is the convex and piecewise linear (*CPL*) criterion function as the sum of such type of penalty functions as $\phi_{jj'}^{+}(w)$ and $\phi_{jj'}^{-}(w)$ The basis exchange algorithms, like linear programming, allow one to find the minimum of such a function efficiently, even in the case of large multidimensional data sets C^{+} and C^{-} (8) [9]:

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

The below *Lemma* can be proved [5]:

Lemma 2: The minimal value $\Phi(\mathbf{w}^{*})$ (14) of the criterion function $\Phi(\mathbf{w})$ is equal to zero if and only if the line $y = (\mathbf{w}^{*})^{T}\mathbf{x}$ (6) preserves all the implications "$\prec$" (7).

5 Example: Causal Sequence of Liver Diseases

The database of the system *Hepar* contains descriptions of patients with variety of chronic liver diseases ω_k ($k = 1,..., K$) [8]. The feature vectors $\mathbf{x}_j(k)$ in the database of *Hepar* are of 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 case of a patient 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:

C_1. *Non hepatitis patients*	- 16 patients
C_2. *Hepatitis acuta*	- 8 patients
C_3. *Hepatitis persistens*	- 44 patients
C_4. *Hepatitis chronica active*	- 95 patients
C_5. *Cirrhosis hepatitis compensate*	- 38 patients
C_6. *Cirrhosis decompensate*	- 60 patients
C_7. *Carcinoma hepatis*	- 11 patients

	Total: 272 patients

(15)

In accordance with medical knowledge, the learning sets C_k (15) have been formed as the causal sequence (3) with $K = 7$. The ranked relation "$\prec$" (5) between feature vectors $\mathbf{x}_j(k)$ ($\mathbf{x}_j(k) \in C_k$) and $\mathbf{x}_{j'}(k')$ ($\mathbf{x}_{j'}(k') \in C_{k'}$) has been defined on this basis. This ranked relation allowed to define the the positive set C^{+} (8) and the negative set C^{-} of the differential vectors $\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j)$. The sets C^{+} and C^{-} 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 $y = (\mathbf{w}^*)^T\mathbf{x}$ (6) that can be used among others for prognosis purposes:

$$y_j(k) = (\mathbf{w}^*)^T\mathbf{x}_j(k) \tag{16}$$

The feature selection allows to determine the most important features x_i influencing the future of a given patient $\mathbf{x}_0$ and to neglect the unimportant features x_i. The feature selection problem can be also based on the minimization of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13) [6].

The causal sequence (3) of diseases ω_k (the learning sets C_k (15)) is preserved in a great part by the ranked model (16). Elements $\mathbf{x}_j(k)$ of each learning set C_k (15) are transformed in accordance with the equation (16) on the points $y_j(k)$ of the ranked line. In result, the sets C_k' of the numbers $y_j(k)$ (16) can be defined (2):

$$C_k' = \{y_j(k): j \in J_k\} \tag{17}$$

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

$$\mu_k = \frac{\sum_j y_j(k)}{m_k} \quad (j \in J_k) \tag{18}$$

and

$$\sigma_k^2 = \frac{\sum_j (y_j(k) - \mu_k)^2}{m_k} \quad (j \in J_k) \tag{19}$$

The results of computations based on the model (17) of data sets C_k (15) are summarized in the below Table 1:

Table 1. The mean values μ_k and variances σ_k^2 of the sets C_k' (17)

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 $y = (\mathbf{w}^*)^T \mathbf{x}$ (16) in order to improve the interpretability of its prognostic applications.

$$y'_j(k) = \alpha\ (\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \tag{20}$$

where α and β are the scaling parameters.

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

$$(\forall \alpha > 0)(\forall \beta) \quad (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \Rightarrow \alpha(\mathbf{w}^*)^T \mathbf{x}_j + \beta < \alpha(\mathbf{w}^*)^T \mathbf{x}_{j'} + \beta \tag{21}$$

The parameters α and β have been fixed through minimization of the sum $Q(\alpha, \beta)$ of the differences $\left| k - \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \right|$ for all the sets C_k (15) and all the feature vector $\mathbf{x}_j(k)$.

$$Q(\alpha, \beta) = \sum_{k=1,..,K} \sum_{j \in I_k} \left| k - \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \right| \tag{22}$$

where I_k is the set of indices j of the feature vectors $\mathbf{x}_j(k)$ from the set C_k (15).

Let us remark that $Q(\alpha, \beta)$ 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 $Q(\alpha, \beta)$. Some results of the scaled model evaluation are shown in the Table 2 and on the Fig. 1.

Table 2. The mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (17) obtained from the ranked model (16) after scaling (20) with the optimal parameters α^* and β^*

Data sets C_k' (33)	Number of patients m_k	Mean value μ_k'	Variance $\sigma_k'^2$ (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)

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 (16) constitutes a valuable characteristic of the patient $\mathbf{x}_0$ 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.

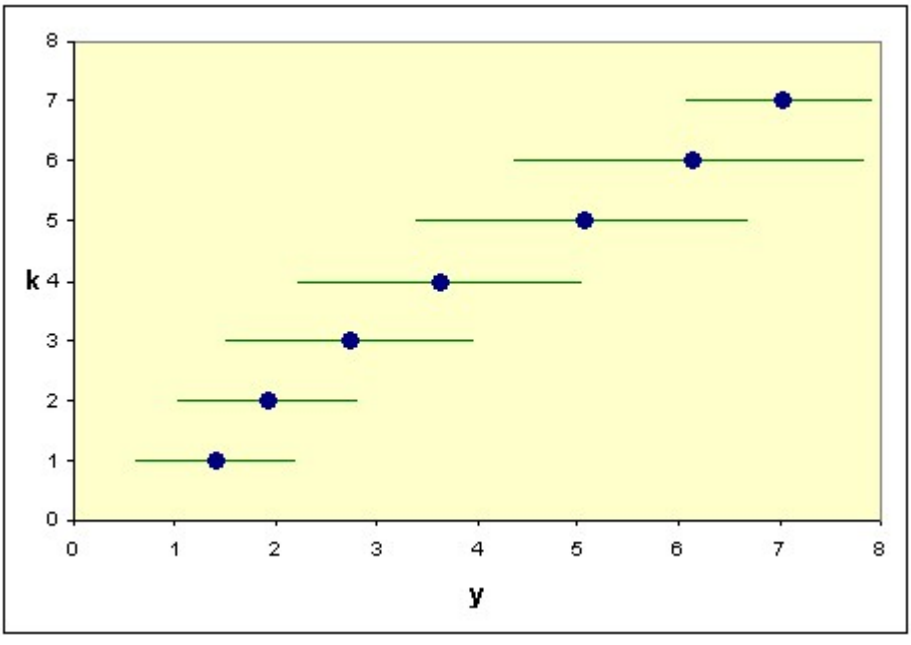

Fig. 1. Graphical presentation the mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (17) obtained from the ranked model (18) after scaling (20) with the optimal parameters α^* and β^*

6 Concluding Remarks

The medical knowledge in the form of a causal sequence of diseases ω_k that are represented by learning sets C_k of labeled feature vectors $\mathbf{x}_j(k)$ ($k = 1,\ldots, K$) allows for designing ranked models. Such models can be based on linear transformations $y_j(k) = (\mathbf{w}^*)^T\mathbf{x}_j(k)$ (16) of n-dimensional feature vectors $\mathbf{x}_j(k)$ on the points $y_j(k)$ of the ranked line which preserves in the best possible manner the order relations "$\prec$" (5) between vectors $\mathbf{x}_j(k)$. The ranked model can be used for the purpose of prognosis, in particular for risk prognosis of new patients. An initial part of the ranked line contains mainly the points $y_j(k)$ related to the first diseases ω_k in the causal sequence () and the final part of the line contains mainly the points $y_j(k)$ representing patients $O_j(k)$ from the last, typically the most dangerous diseases in the sequence (1). The location $y_0 = (\mathbf{w}^*)^T\mathbf{x}_0(k)$ of a new patient O_0 on the ranked line can be used for the purpose of initial diagnosis or risk evaluation for this patient. The screening procedures for the search of potentially ill patients eligible for further investigations and therapy could be based on the above ranked model.

The feature selection problem allows to determine the most important features 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].

Bibliography

1. Bobrowski, L., Łukaszuk, T., Wasyluk, H.: Ranked modeling of liver diseases sequence. European Journal of Biomedical Informatics (2007) (to be published)
2. Duda, O.R., Hart, P.E., Stork, D.G.: Pattern Classification. J. Wiley, New York (2001)
3. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London (1972)
4. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice-Hall Inc., Englewood Cliffs (1991)
5. Bobrowski, L.: Ranked modelling with feature selection based on the CPL criterion functions. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), vol. 3587, pp. 218–227. Springer, Heidelberg (2005)
6. Bobrowski, L., Łukaszuk, T.: Ranked linear modeling in survival analysis. In: Bobrowski, L., Doroszewski, J., Victor, N. (eds.) Lecture Notes of the ICB Seminars: Statistics and Clinical Practice, pp. 61–67. IBIB PAN, Warsaw (2005)
7. Bobrowski, L.: Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych (Data mining based on convex and piecewise linear (CPL) criterion functions). Białystok Technical University (2005) (in Polish)
8. Bobrowski, L., Wasyluk, H.: Diagnosis support rules of the Hepar system. In: Petel, V.L., Rogers, R., Haux, R. (eds.) MEDINFO 2001, pp. 1309–1313. IOS Press, Amsterdam (2001)
9. Bobrowski, L.: Design of piecewise linear classifiers from formal neurons by some basis exchange technique. Pattern Recognition 24(9), 863–870 (1991)