ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2009
Informatyka – Zeszyt 4

Małgorzata Krętowska[1]

PROGNOSTIC ABILITIES OF DIPOLES BASED ENSEMBLES – COMPARATIVE ANALYSIS

Abstract: In the paper, comparative analysis of ensembles of dipolar neural networks and regression trees was conducted. The techniques are based on the dipolar criterion function. Appropriate formation of dipoles (pairs of feature vectors) allows using them for analysis of censored survival data. As the result the methods return aggregated Kaplan-Meier survival function. The results, obtained by neural networks and regression trees based ensembles, are compared by using Brier score and direct and indirect measures of predictive accuracy.

Keywords: survival analysis, dipolar criterion, regression trees ensembles, neural networks ensembles

1. Introduction

In real data problems, the question about the future behavior of a given patient arises. Such situation is very common in survival data, in which the failure time is under investigation. In medical data, the failure is often defined as death or disease relapse and time is measured from the initial event, e.g. surgery. Analyzing the survival data requires taking into account censored observations, for which the exact failure time is unknown. The follow-up time for such patients gives us only the information, that the failure did not occur before.

Besides statistical techniques (the most common Cox's proportional hazards model [2]), which require some conditions to fulfill, other non-statistical methods are developed. Artificial neural networks and regression trees belong to the most popular ones. Recently, also methods concerning the use of ensemble of regression trees in prognosis of survival time appear. Their application allows receiving the tool unaffected by small changes in dataset, what is particularly important in discovering the risk factors. Hothorn *et al.* [5] proposes boosting survival trees to create aggregated survival function. Krętowska [11] developed the approach by using the dipolar regression trees or dipolar neural networks [12] instead of the structure proposed in

[1] Faculty of Computer Science, Białystok Technical University, Białystok

[5]. The technique proposed by Ridgeway [15] allows minimizing the partial likelihood function (boosting Cox's proportional hazard model). The Hothorn *et al.* [6] developed two approaches for censored data: random forest and gradient boosting.

The paper is organized as follows. In Section 2. the introduction to survival data as well as Kaplan-Meier survival function are presented. Section 3. introduces the idea of dipoles and in Section 4. two dipoles based structures: neural networks and regression trees are described. In Section 5. the algorithm of building the ensemble of predictors is presented. Experimental results are presented in Section 7. They were carried out on the base of two real datasets. The first one contains the feature vectors describing the patients with primary biliary cirrhosis of the liver [3], the other includes the information from the Veteran's Administration lung cancer study [7]. Section 8. summarizes the results.

2. Kaplan-Meier survival function

We have learning sample $L = (\mathbf{x}_i, t_i, \delta_i)$, $i = 1, 2, ..., n$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i - survival time and δ_i - failure indicator, which is equal to 0 for censored cases and 1 for uncensored ones.

The distribution of random variable T, which represents the true survival time, may be described by the marginal probability of survival up to time $t > 0$ ($S(t) = P(T > t)$). The estimation of survival function $S(t)$ may be done by using the Kaplan-Meier product limit estimator [8], which is calculated on the base of learning sample L and is denoted by $\hat{S}(t)$:

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(\frac{m_j - d_j}{m_j} \right) \tag{1}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered survival times from the learning sample L, in which the event of interest occurred, d_j is the number of events at time $t_{(j)}$ and m_j is the number of patients at risk at $t_{(j)}$ (i.e., the number of patients who are alive at $t_{(j)}$ or experience the event of interest at $t_{(j)}$).

The 'patients specific' survival probability function is given by $S(t|\mathbf{x}) = P(T > t|\mathbf{X} = \mathbf{x})$. The conditional survival probability function for the new patient with covariates vector $\mathbf{x}_{new}$ is denoted by $\hat{S}(t|\mathbf{x}_{new})$.

3. Dipoles

The methodology used during the learning process of artificial neural network and induction of regression tree bases on the concept of dipole [1]. The dipole is a pair of

different covariate vectors $(\mathbf{x}_i, \mathbf{x}_j)$ from the learning set. Mixed and pure dipoles are distinguished. Mixed dipoles are formed between objects that should be separated, while pure ones between objects that are similar from the point of view of the analyzed criterion. The aim is to find such a hyper-plane $H(\mathbf{w}, \theta)$ that divides possibly high number of mixed dipoles and possibly low number of pure ones. It is done by minimization of the dipolar criterion function.

Two types of piece-wise linear and convex penalty functions $\varphi_j^+(\mathbf{v})$ and $\varphi_j^-(\mathbf{v})$ are considered:

$$\varphi_j^+(\mathbf{v}) = \begin{cases} \delta_j - < \mathbf{v}, \mathbf{y}_j > \text{ if } < \mathbf{v}, \mathbf{y}_j > \leq \delta_j \\ 0 \qquad\qquad\qquad\quad \text{if } < \mathbf{v}, \mathbf{y}_j > > \delta_j \end{cases} \tag{2}$$

$$\varphi_j^-(\mathbf{v}) = \begin{cases} \delta_j + < \mathbf{v}, \mathbf{y}_j > \text{ if } < \mathbf{v}, \mathbf{y}_j > \geq -\delta_j \\ 0 \qquad\qquad\qquad\quad \text{if } < \mathbf{v}, \mathbf{y}_j > < -\delta_j \end{cases} \tag{3}$$

where δ_j is a margin ($\delta_j = 1$), $\mathbf{y}_j = [1, x_1, \ldots, x_N]^T$ is an augmented covariate vector and $\mathbf{v} = [-\theta, w_1, \ldots, w_N]^T$ is an augmented weight vector. Each mixed dipole $(\mathbf{y}_i, \mathbf{y}_j)$, which should be divided, is associated with function $\varphi_{ij}^m(\mathbf{v})$ being a sum of two functions with opposite signs ($\varphi_{ij}^m(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_i^-(\mathbf{v})$ or $\varphi_{ij}^m(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_i^+(\mathbf{v})$). For pure dipoles that should remain undivided we associate function: $\varphi_{ij}^p(\mathbf{v})$ ($\varphi_{ij}^p(\mathbf{v}) = \varphi_j^+(\mathbf{v}) + \varphi_i^+(\mathbf{v})$ or $\varphi_{ij}^c(\mathbf{v}) = \varphi_j^-(\mathbf{v}) + \varphi_i^-(\mathbf{v})$). A dipolar criterion function is a sum of the penalty functions associated with each dipole:

$$\Psi_d(\mathbf{v}) = \sum_{(j,i) \in I_p} \alpha_{ij} \varphi_{ij}^p(\mathbf{v}) + \sum_{(j,i) \in I_m} \alpha_{ij} \varphi_{ij}^m(\mathbf{v}) \tag{4}$$

where α_{ij} determines relative importance (price) of the dipole $(\mathbf{y}_i, \mathbf{y}_j)$, I_p and I_m are the sets of pure and mixed dipoles, respectively.

The rules of dipoles formations depend on the purpose of our research. Assuming that the analysis aims at dividing the feature space into such areas, which would include the patients with similar survival times, pure dipoles are created between pairs of feature vectors, for which the difference of failure times is small, mixed dipoles - between pairs with distant failure times. Taking into account censored cases the following rules of dipole construction can be formulated:

1. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the pure dipole, if
   - $\sigma_i = \sigma_j = 1$ and $|t_i - t_j| < \eta$
2. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   - $\sigma_i = \sigma_j = 1$ and $|t_i - t_j| > \zeta$
   - $(\sigma_i = 0, \sigma_j = 1$ and $t_i - t_j > \zeta)$ or $(\sigma_i = 1, \sigma_j = 0$ and $t_j - t_i > \zeta)$

Parameters η and ζ are equal to quartiles of absolute values of differences between uncensored survival times. The parameter η is fixed as 0.2 quartile and ζ - 0.6.

As the result of minimization of the dipolar criterion function we receive the values of parameters $\mathbf{v}$ of the hyper-plane. Depending on the set of dipoles, parameters $\mathbf{v}$ describe the neuron in artificial neural network or internal node in regression tree.

4. Individual prognostic structures

Two prognostic structures are considered in the paper: dipolar neural network [10] and regression tree [9]. The basic element of the structures (that is binary neurons and internal nodes) are characterized by the hyper-plane with parameters $\mathbf{v}$:

$$z = f(\mathbf{y}, \mathbf{v}) = \begin{cases} 1 & \text{if } \mathbf{v}^T\mathbf{y} \geq 0 \\ 0 & \text{if } \mathbf{v}^T\mathbf{y} < 0 \end{cases} \tag{5}$$

From the geometrical point of view an element divides a feature space into two subspaces by using hyperplane $H(\mathbf{v}) = \{\mathbf{y} : \mathbf{v}^T\mathbf{y} = 0\}$. If the vector $\mathbf{y}$ is situated on the positive side of the hyper-plane, the element returns 1 ($z = 1$).

Neural network

A dipolar neural network model, considered in the paper, consists of two layers: input and output layer. The neurons weight values are obtained by sequential minimization of the dipolar criterion functions. The function is built from all the pure dipoles and those mixed dipoles which were not divided by previous neurons. The learning phase is finished when all the mixed dipoles are divided. The other, optimization phase, aims at distinguishing and enlargement of prototypes (i.e. active fields which contain the largest number of feature vectors $\mathbf{x}$) and at reduction of redundant neurons [10].

The output layer of R binary neurons divided the N-dimensional feature space into disjoint regions - *active fields* (AF). Each region is represented by R-dimensional output vector: $\mathbf{z} = [z_1, z_2, \ldots, z_R]^T$, where $z_i \in \{0,1\}$. As the result, the set of active fields $SAF = \{AF^1; AF^2; \ldots, AFi^k\}$ is received. Each active field AF^j contains the subset L^j of observations from the learning sample L.

Regression tree

Hierarchical and sequential structure of a regression tree recursively partitions the feature space. The tree consists of terminal nodes (leaves) and internal (non-terminal) nodes. An internal node contains a split (5), which tests the value of an expression

of the covariates. Each distinct outcome (0 or 1) of the test generates one child node, which means that all non-terminal nodes have two child nodes. A terminal node generates no descendant. The function in a given node is designed on the base on those feature vectors that have reached the node. The induction of survival tree is stopped if one of the following conditions is fulfilled: 1) all the mixed dipoles are divided; 2) the set that reach the node consists of less than 5 uncensored cases.

Each terminal node represents one region in the N-dimensional feature space. Similarly to the neural network results, the leave j contains the subset L^j of observations from the learning sample L.

5. Ensembles of predictors

Let assume, that we have a set of k dipolar predictors (dipolar neural networks or dipolar regression trees): DP_i, $i = 1,2,\ldots,k$. The set is called ensemble when each of k predictors is generated on base of k learning samples $(L_1, L_2, \ldots, L_k)$ drawn with replacement from a given sample L. As the result of each dipolar predictor DP_i, the set $SL_i = \{L_i^1; L_i^2; \ldots, L_i^{k_i}\}$ of observations from learning sample L_i. Having a new covariate vector $\mathbf{x}_{new}$, each DP_i, $i = 1,2,\ldots,k$ returns the subset of observations $L_i(\mathbf{x}_{new})$ which is connected with region (or active field in case of neural networks), to which the new vector belongs. Having k sets $L_i(\mathbf{x}_{new})$, aggregated sample $L_A(\mathbf{x}_{new})$ is built [5]:

$$L_A(\mathbf{x}_{new}) = \{L_1(\mathbf{x}_{new}); L_2(\mathbf{x}_{new}); \ldots; L_k(\mathbf{x}_{new})\}$$

The aggregated conditional Kaplan-Meier survival function, calculated on the base of set $L_A(\mathbf{x}_{new})$ can be referred to as $\hat{S}_A(t|\mathbf{x}_{new})$.

The algorithm for receiving the aggregated survival function is as follows:

1. Draw k bootstrap samples $(L_1, L_2, \ldots, L_k)$ of size n with replacement from L
2. Induction of dipolar predictor DP_i based on each bootstrap sample L_i, $i = 1,2,\ldots,k$
3. Build aggregated sample $L_A(\mathbf{x}_{new}) = \{L_1(\mathbf{x}_{new}); L_2(\mathbf{x}_{new}), \ldots, L_k(\mathbf{x}_{new})\}$
4. Compute the Kaplan-Meier aggregated survival function for a new observation $\mathbf{x}_{new}$: $\hat{S}_A(t|\mathbf{x}_{new})$.

6. Measures of predictive accuracy

Beside the problems concerning the use of censored data in the process of building the prediction tool, the question how to evaluate the prediction ability of received models appears. The lack of exact failure times for a part of data causes that the

classical measures based on difference between empirical and theoretical values can not be used. Instead of them, other, censoring oriented, measures are proposed.

One of them is the Brier score introduced by Graf *at al.* [4]. The Brier score as a function of time is defined by

$$BS(t) = \frac{1}{n}\sum_{i=1}^{N}(\hat{S}(t|\mathbf{x}_i)^2 I(t_i \leq t \wedge \sigma_i = 1)\hat{G}(t_i)^{-1} + (1-\hat{S}(t|\mathbf{x}_i))^2 I(t_i > t)\hat{G}(t)^{-1}) \tag{6}$$

where $\hat{G}(t)$ denotes the Kaplan-Meier estimator of the censoring distribution. It is calculated on the base of observations $(t_i, 1-\delta_i)$. $I(condition)$ is equal to 1 if the condition is fulfilled, 0 otherwise. The BS equal to 0 means the best prediction.

The Brier score belongs to direct estimators of prediction ability, because it uses the information explicitly from the data. Another direct approach is proposed by Schemper and Henderson [14]. The predictive accuracy (without covariates), expressed by absolute predictive error (*APE*), at each distinct failure time $t_{(j)}$ is defined as:

$$\hat{M}(t_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}\left[I(t_i > t_{(j)})(1-\hat{S}(t_{(j)})) + \delta_i I(t_i \leq t_{(j)})\hat{S}(t_{(j)}) + (1-\delta_i)I(t_i \leq t_{(j)})\left\{(1-\hat{S}(t_{(j)}))\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)} + \hat{S}(t_{(j)})(1-\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)})\right\}\right] \tag{7}$$

The measure with covariates ($\hat{M}(t_{(j)}|\mathbf{x})$) is obtained by replacing $\hat{S}(t_{(j)})$ by $\hat{S}(t_{(j)}|\mathbf{x})$ and $\hat{S}(t_i)$ by $\hat{S}(t_i|\mathbf{x})$. To receive overall estimators of *APE* with ($\hat{D}_x$) and without covariates ($\hat{D}$) the weighed averages of estimators over failure times are calculated:

$$\hat{D} = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \hat{M}(t_{(j)}) \tag{8}$$

$$\hat{D}_{\mathbf{x}} = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \hat{M}(t_{(j)}|\mathbf{x}) \tag{9}$$

where $w = \sum_j \hat{G}(t_{(j)})^{-1} d_j$, d_j is the number of events at time $t_{(j)}$ and $\hat{G}(t)$ denotes the Kaplan-Meier estimator of the censoring distribution (see equation 6).

The indirect estimation of predictive accuracy was proposed by Schemper [13]. In the approach the estimates (without $\tilde{M}(t_{(j)})$ and with covariates $\tilde{M}(t_{(j)}|\mathbf{x})$) are defined by

$$\tilde{M}(t_{(j)}) = 2\hat{S}(t_{(j)})(1-\hat{S}(t_{(j)})) \tag{10}$$

$$\tilde{M}(t_{(j)}|\mathbf{x}) = 2n^{-1}\sum_i \hat{S}(t_{(j)}|\mathbf{x}_i)(1-\hat{S}(t_{(j)})|\mathbf{x}_i) \tag{11}$$

The overall estimators of predictive accuracy with ($\tilde{D}_{S,\mathbf{x}}$) and without ($\tilde{D}_S$) covariates are calculated similarly to the estimators $\hat{D}_{\mathbf{x}}$ and $\hat{D}$. The only change is replacing $\hat{M}(t_{(j)})$ and $\hat{M}(t_{(j)}|\mathbf{x})$ by $\tilde{M}(t_{(j)})$ and $\tilde{M}(t_{(j)}|\mathbf{x})$ respectively.

Based on the above overall estimators of absolute predictive error, explained variation can be defined as:

$$\tilde{V}_S = \frac{\tilde{D}_S - \tilde{D}_{S,\mathbf{x}}}{\tilde{D}_S} \tag{12}$$

and

$$\hat{V} = \frac{\hat{D} - \hat{D}_{\mathbf{x}}}{\hat{D}} \tag{13}$$

7. Experimental results

All the experiments were performed using the ensemble of 200 dipolar predictors *DP*. The measures of predictive accuracy were calculated on the base of learning sample *L*. To calculate the aggregated survival function for a given example $\mathbf{x}$ from the learning set *L*, only such DP_i ($i = 1,2,\ldots,200$) were taken into consideration, for which $\mathbf{x}$ was not belonged to the learning set L_i (i.e. $\mathbf{x}$ did not participate in the learning process of the DP_i).

The analysis was conducted on the base on two datasets. The first one is from the Mayo Clinic trial in primary biliary cirrhosis (*PBC*) of the liver conducted between 1974 and 1984 [3]. 312 patients participated in the randomized trial. Survival time was taken as a number of days between registration and death, transplantation or study analysis time in July 1986. Patients are described by the following features: age(*AGE*), sex, presence of edema, logarithm of serum bilirubin [mg/dl] (*LOGBILL*), albumin [gm/dl] (*ALBUMIN*), logarithm of prothrombin time [seconds], histologic stage of disease. Dataset contains 60 per cent of censored observations.

In table 1, the results for *PBC* dataset are presented. Three different measures of predictive accuracy were calculated for three methods: Kaplan-Meier estimator, ensemble of DNN (dipolar neural network) and ensemble of DRT (dipolar regression tree). As we can see the absolute predictive error for K-M estimator (which is equivalent to the model without covariates) is equal to 0.37 and is higher than for other two methods (0.29 - indirect approach (0.26 - direct approach) for EDNN and 0.23(0.22) for EDRT). Comparing the results received for EDNN and EDRT we can noticed that for the model with all covariates as well as for model with only one feature the predictive measures are better for EDRT. Brier score for EDNN is equal to 0.17 and is bigger by 0.1 than Brier score for EDRT. In case of indirect and direct APE - explained variation for EDNN is smaller (0.22 (0.26)) than for EDRT -

0.39(0.41). Taking into account individual factors, the order of them is the same for both methods. The most important prognostic factor is logarithm of serum bilirubin for which the explained variation is equal to 0.25 (0.25) and 0.34 (0.35) for EDNN and EDRT respectively. The influence of age and albumin for prediction of survival probability is less important.

Table 1. Measures of predictive accuracy for *PBC* dataset

Model	*BS* (12years)	Indirect *APE*/ Explained variation	Direct *APE*/ Explained variation
K-M Estimator	0.23	0.37	0.37
Ensemble of DNN			
all covariates	0.17	0.29/0.22	0.27/0.26
AGE	0.22	0.36/0.036	0.36/0.038
LOGBILL	0.17	0.28/0.25	0.28/0.25
ALBUMIN	0.22	0.33/0.11	0.33/0.12
Ensemble of DRT			
all covariates	0.16	0.23/0.39	0.22/0.41
AGE	0.18	0.33/0.11	0.33/0.12
LOGBILL	0.16	0.24/0.34	0.24/0.35
ALBUMIN	0.18	0.28/0.24	0.28/0.24

The other analyzed data set contains the information from the Veteran's Administration (*VA*) lung cancer study [7]. In this trial, male patients with advanced inoperable tumors were randomized to either standard (69 subjects) or test chemotherapy (68 subjects). Only 9 subjects from 137 were censored. Information on cell type (0 - squamous, 1 - small, 2 - adeno, 3 - large) - CELL TYPE, prior therapy, performance status at baseline (Karnofsky rating - KPS), disease duration in months (TIME) and age in years at randomization (AGE), was available.

The measures of predictive accuracy for *VA lung cancer* data was shown in table 2. The unconditional absolute predictive error is 0.335. The ensemble of DNN, built on the base of all the covariates, reduces the error by 0.035 or 0.045 for indirect and direct approach respectively. The ensemble of DRT reduces the error by 0.145 and 0.185. As for *PBC* dataset case the results are better for EDRT, but the order of prognostic factors is the same. The most important prognostic factor is KPS (error equal to 0.3 (029) and 0.25(0.25), for EDNN and EDRT respectively). Explained variation is 11 (13) and 25 (25) per cent. Taking into account the EDNN, other variables have the marginal influence on the prediction of survival probability, but in case of EDRT also the cell type is quite important (explained variation equal to 13 (12) per cent).

Table 2. Measures of predictive accuracy for *VA lung cancer* data

Model	*BS* (100 days)	Indirect *APE*/ Explained variation	Direct *APE*/ Explained variation
K-M Estimator	0.24	0.335	0.335
Ensemble of DNN			
all covariates	0.18	0.3/0.11	0.29/0.14
AGE	0.24	0.32/0.034	0.33/0.013
CELL TYPE	0.24	0.33/0.002	0.33/0.006
KPS	0.19	0.3/0.11	0.29/0.13
TIME	0.24	0.33/0.003	0.33/0.003
Ensemble of DRT			
all covariates	0.09	0.22/0.35	0.18/0.46
AGE	0.2	0.3/0.09	0.3/0.1
CELL TYPE	0.19	0.29/0.13	0.3/0.12
KPS	0.18	0.25/0.25	0.25/0.25
TIME	0.22	0.31/0.07	0.31/0.07

8. Conclusions

In the paper, prognostic abilities of two ensemble of dipolar predictors (neural networks and regression trees) were compared. The prognostic ability of the models was verified by several measures, such as the Brier score and direct and indirect estimators of absolute predictive errors: $\tilde{D}_{S,x}$, $\hat{D}_x$. In all cases the measures were better for ensemble of dipolar regression trees then for ensemble of neural networks. For *VA lung cancer* data the explained variation was equal to 0.35 (0.46) for EDRT and 0.11 (0.14) (indirect (direct approach)) for EDNN. Similarly for *PBC* dataset, the explained variation received for EDRT - 0.39 (0.41) was greater than for EDNN - 0.22 (0.26). It worth noticed than two method distinguished the same risk factors. The feature that influence the survival the most is Karnofsky rating in case of *VA lung cancer* data and serum bilirubin for *PBC* dataset. The results suggest that the way of creating the consecutive hyper-planes in regression trees approach allows using the information from the given learning sample in the better manner.

References

[1] Bobrowski L., Krętowska M., Krętowski M.: Design of neural classifying networks by using dipolar criterions. Proc. of the Third Conference on Neural Networks and Their Applications, Kule, Poland (1997) 689–694

[2] Cox D.R.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society B **34** (1972) 187–220

[3] Fleming T.R.: Harrington D.P., Counting Processes and Survival Analysis. John Wiley & Sons, Inc. (1991)

[4] Graf E., Schmoor C., Sauerbrei W., Schumacher M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine **18** (1999) 2529–2545

[5] Hothorn T., Lausen B., Benner A., Radespiel-Troger M.: Bagging survival trees. Statistics in medicine **23** (2004) 77–91

[6] Hothorn T., Buhlmann P., Dudoit S., Molinaro A. M., van der Laan M. J.: Survival ensembles. [URL http://www.bepress.com/ucbbiostat/paper174] U.C. Berkeley Division of Biostatistics Working Paper Series **174** (2005)

[7] Kalbfleisch J.D., Prentice R.L.: The statistical analysis of failure time data. John Wiley & Sons, New York (1980)

[8] Kaplan E.L., Meier P.: Nonparametric estimation from incomplete observations, Journal of the American Statistical Association **5** (1958) 457–481

[9] Krętowska M.: Dipolar regression trees in survival analysis. Biocybernetics and biomedical engineering **24** (3) (2004) 25–33

[10] Krętowska M., Bobrowski L.: Artificial neural networks in identifying areas with homogeneous survival time, L.Rutkowski et al. (Eds.): ICAISC 2004, LNAI 3070, (2004) 1008–1013

[11] Krętowska M.: Random forest of dipolar trees for survival prediction, L.Rutkowski et al. (Eds.): ICAISC 2006, LNAI 4029, (2006) 909–918

[12] Krętowska M.: Ensemble of Dipolar Neural Networks in Application to Survival Data, L.Rutkowski et al. (Eds.): ICAISC 2008, LNAI 5097, (2008) 78–88

[13] Schemper M.: Predictive accuracy and explained variation. Statistics in Medicine **22** (2003) 2299–2308

[14] Schemper M., Henderson R.: Predictive accuracy and explained variation in Cox regression. Biometrics **56** (2000) 249–255

[15] Ridgeway G.: The state of boosting. Computing Science and Statistics **31** (1999) 1722–1731

[16] Blake, C., Merz, C.: UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[17] Duda O.R., Heart P.E., Stork D.G.: Pattern Classification, Second edition, John Wiley & Sons, 2001.

[18] Quinlan J.: Induction of decision trees, Machine Learning 1(1), 1986, pp. 81-106.

ZDOLNOŚCI PROGNOSTYCZNE KOMITETÓW BAZUJĄCYCH NA DIPPOLACH - ANALIZA PORÓWNAWCZA

Streszczenie W pracy przedstawiona została analiza porównawcza własności prognostycznych komitetów bazujących na sieciach neuronowych oraz drzewach regresyjnych. Tworzenie kolejnych sięć przestrzeni cech w obu metodach polega na minimalizacji odpowiednio skonstruowanego kryterium dipolowego. Do porównania metod wykorzystano indeks Brier'a oraz pośrednią i bezpośrednią miarę jakości predykcji. Eksperymenty wykonane zostały w oparciu o dwa rzeczywiste zbiory danych: pacjentów z pierwotną marskością żółciową wątroby oraz z rakiem płuc. W obu przypadkach wyniki otrzymane dla komitetu drzew regresyjnych były lepsze niż dla komitetu sieci neuronowych. Dotyczyło to zarówno badania jakości całego modelu, do którego wzięte zostały wszystkie dostępne w zbiorze cechy, jak też jakości prognostycznej pojedynczych cech. Natomiast uszeregowanie poszczególnych cech jako czynników ryzyka było podobne w obu metodach. Podsumowując można powiedzieć, że sposób podziału przestrzeni cech zaproponowany w drzewach regresyjnych w lepszy sposób wykorzystuje informacje zawarte w zbiorze uczącym.

Słowa kluczowe: analiza przeżyć, kryterium dipolowe, komitety drzew decyzyjnych, komitety sieci neuronowych

Artykuł zrealizowano w ramach pracy badawczej W/WI/4/08.