Random Forest of Dipolar Trees for Survival Prediction

Małgorzata Krętowska

Faculty of Computer Science
Białystok Technical University
Wiejska 45a, 15-351 Białystok, Poland
mmac@ii.pb.bialystok.pl

Abstract. In the paper the method of using the ensemble of dipolar trees for survival prediction is presented. In the approach the random forest is applied to calculate the aggregated Kaplan-Meier survival function for a new patient. The induction of individual dipolar regression tree is based on minimization of a piece-wise linear criterion function. The algorithm allows using the information from censored observations for which the exact survival time is unknown. The Brier score is used to evaluate the prediction ability of the received model.

1 Introduction

The development of prognostic tools is one of the major tasks in survival analysis. The physicians are concerned not only with the prediction of the exact survival time for a given patient but also with discovering the factors that influence the survival. The most common statistical method used in analysis of survival data is Cox's proportional hazard model [4]. Its application is limited by additional assumptions required for the analyzed phenomenon. This limitation concerns also other statistical methods. If the requirements are difficult to obey some other techniques are adopted. Among them artificial neural networks and regression trees are considered as ones of the most promising tools.

The analysis of survival data may be treated either as the regression or classification task. In both cases the problem how to treat censored data arises. Censored observations include incomplete knowledge about the exact time of event occurrence. We only know that the true survival time is not less than their follow-up time.

The proposed algorithms of regression trees induction include modifications which allow coping with censored data. Its application in survival analysis aimed at identifying subgroups that are homogeneous in their survival experience. Marubini *et al.* [13] proposed an approach based on Cox's proportional hazards model and partial likelihood approach. Similar method was developed by Ciampi *et al.* [3]. Davis and Anderson [5] assumed an exponential model for the survival distribution and as a goodness-of-split criterion exploited exponential log-likelihood. Krętowska [12] proposed induction of the multivariate tree based on the minimization of a dipolar criterion function.

L. Rutkowski et al. (Eds.): ICAISC 2006, LNAI 4029, pp. 909–918, 2006.

The problem that arises while analyzing the results calculated on the base of the single tree is instability, especially in discovering the risk factors. To stabilize the predictions, the ensembles of several models are used. Hothorn *et al.* [8] proposed boosting survival trees to create aggregated survival function. The approach proposed by Ridgeway [14] allows minimizing the partial likelihood function (boosting Cox's proportional hazard model). The Hothorn *et al.* [9] developed two approaches for censored data: random forest and gradient boosting. Breiman [2] provided the software that allows induction of the random forest for censored data.

In the paper the random forest consisting of dipolar survival trees is analyzed. The method of building the ensemble of trees is based on the approach proposed by Hothorn *at al.* [8]. The method enables calculation the aggregated Kaplan-Meier survival function for a new patient. The Brier score [7] is used to evaluate the prediction ability of the received model.

The paper is organized as follows. Section 2 describes the survival data and introduces the idea of Kaplan-Meier survival function. In Section 3 induction of dipolar survival tree is presented. Section 4 contains the algorithm how to build the aggregated survival function based on random forest. Experimental results are presented in Section 5. The experiments 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 [6], the other includes the information from the Veteran's Administration lung cancer study [10]. Section 6 summarizes the results.

2 Survival Data

Let T^0 denotes the true survival time and C denotes the true censoring time with distribution functions F and G respectively. We observe random variable $O = (T, \Delta, \mathbf{X})$, where $T = \min(T^0, C)$ is the time to event, $\Delta = I(T \leq C)$ is a censoring indicator and $\mathbf{X} = (X_1, ..., X_N)$ denotes the set of N covariates from a sample space χ. We observe the 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.

The distribution of random variable T may be described by the marginal probability of being event free up to a time $t > 0$ ($S(t) = P(T > t)$). The estimation of the survival function $S(t)$ may be done by using the Kaplan-Meier product limit estimator [11]. The Kaplan-Meier function is calculated on the base of learning sample 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 $\mathbf{x}_n$ is denoted by $\hat{S}(t|\mathbf{x}_n)$.

3 Dipolar Survival Tree Induction

Hierarchical and sequential structure of a tree recursively partitions the feature space. The tree consists of terminal nodes (leaves) and internal (non-terminal) nodes. An internal node contains a split, which tests the value of an expression of the covariates. In the proposed approach the split is equivalent to the hyper-plane $H(\mathbf{w}, \theta) = \{(\mathbf{w}, \mathbf{x}) :< \mathbf{w}, \mathbf{x} >= \theta\}$. For given covariate vector $\mathbf{x}$, the result of the test is equal to 0, if the inner product $< \mathbf{w}, \mathbf{x} >$ is less then θ and 1, otherwise. Each distinct outcome 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 tree induction aims at establishing the structure of the tree (the number of internal nodes) and the values of hyper-planes parameters. The proposed algorithm [12] is based on the concept of dipoles [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:

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

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

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.

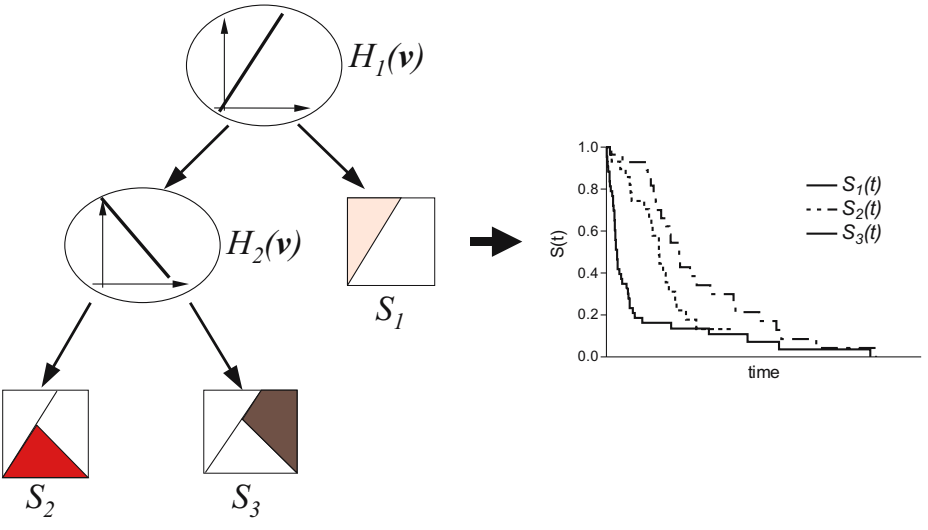

Fig. 1. An example of survival tree

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 (see Fig. 1), 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. The hyper-planes in the internal nodes of the tree are computed by minimization of dipolar criterion function, starting from the root. 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.

4 Random Forest Algorithm

The random forest method [8] allows estimation the conditional survival function $\hat{S}(t|\mathbf{x}_n)$ on the base of k learning samples $(L_1, L_2, \ldots, L_k)$ drawn with replacement from the given sample L. For each learning sample L_i $(i = 1, 2, \ldots, k)$ the set of observations $L_i(\mathbf{x}_n)$ which are close to covariates vector $\mathbf{x}_n$ is distinguished.

The dipolar survival tree is calculated for each learning set L_i, $i = 1, 2, \ldots, k$. The covariates vector $\mathbf{x}_i$ is included to the set $L_i(\mathbf{x}_n)$ when it belongs to the same leaf of the survival tree as $\mathbf{x}_n$ itself. Having k sets $L_i(\mathbf{x}_n)$, aggregated sample $L_A(\mathbf{x}_n)$ is built:

$$L_A(\mathbf{x}_n) = [L_1(\mathbf{x}_n); L_2(\mathbf{x}_n); \ldots; L_k(\mathbf{x}_n)]$$

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

To summarize the above considerations, the random forest algorithm leading to receive 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 survival tree $T(L_i)$ based on each bootstrap sample L_i
3. Build aggregated sample $L_A(\mathbf{x}_n) = [L_1(\mathbf{x}_n);\ L_2(\mathbf{x}_n), \ldots, L_k(\mathbf{x}_n)]$
4. Compute the Kaplan-Meier aggregated survival function for a new observation $\mathbf{x}_n$: $\hat{S}_A(t|\mathbf{x}_n)$.

For the evaluation of prediction ability of the method the Brier score introduced by Graf *at al.* [7] was used. The Brier score as a function of time is defined by

$$BS(t) = \frac{1}{n}\sum_{i=1}^{N}(\hat{S}(t|\mathbf{x}_i)^2 I(t_i \le t \wedge \sigma_i = 1)\hat{G}(t_i)^{-1} + (1 - \hat{S}(t|\mathbf{x}_i))^2 I(t_i > t)\hat{G}(t)^{-1}) \quad (5)$$

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.

5 Experimental Results

The analysis was conducted on the base on two datasets. The first data is from the Mayo Clinic trial in primary biliary cirrhosis (*PBC*) of the liver conducted between 1974 and 1984 [6]. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 106 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. The analysis

was done on the base of 418 patients described by the following features: age, sex, presence of edema, serum bilirubin in mg/dl, albumin in gm/dl, platelets per cubic ml/1000, prothrombin time in seconds, histologic stage of disease. The number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986 was available.

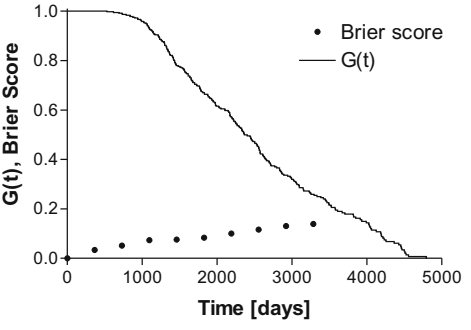

Fig. 2. Kaplan-Meier estimator $G(t)$ of the censoring distribution $G(t) = P(C > t)$; Brier score for selected values of t

In Fig. 2 Kaplan-Meier estimator of the censoring distribution together with the Brier score received for selected values of time (1, 2, 3..., 10 years) are presented. We can see that the Brier score values are quite small for the lower values of time and increase over time. The maximum value of Brier score for t=10 years is equal to 0.14.

In Fig. 3 we can observe the differences between Kaplan-Meier survival functions for 50 years old women and men with two different levels of serum bilirubin and four histologic stages of disease. Other features, which were not considered in the analysis, were fixed to their median values: absence of edema; albumin = 3; platelets = 251; prothrombin time = 11.

The impact of the level of serum bilirubin for survival can be observed in the figures. The survival for patients with lower value of serum bilirubin is much better than for patients with the value equal to 5. Taking into account two upper figures (see Fig. 3) one can see that the differences between survival functions for the first three histologic stages are not significant. The worse survival prediction is for men with histologic stage 4 (median survival time equal to 4079 days). The survival prediction for serum bilirubin equal to 5 is worse in all analyzed cases. For men we can see significant differences between the function for the first histologic stage and other stages. Median survival time is equal to 3244, 1657, 1478 and 1462 [days] for the consecutive histologic stages. The survival functions for women with different stages of disease are more diverse. The best prediction is for the first stage and is getting worse as the stage number increases. Median survival time is equal to 3358 (1st stage), 2081, 1297 and 1152 for the 4th histologic stage. We can say that women response better for the given treatment

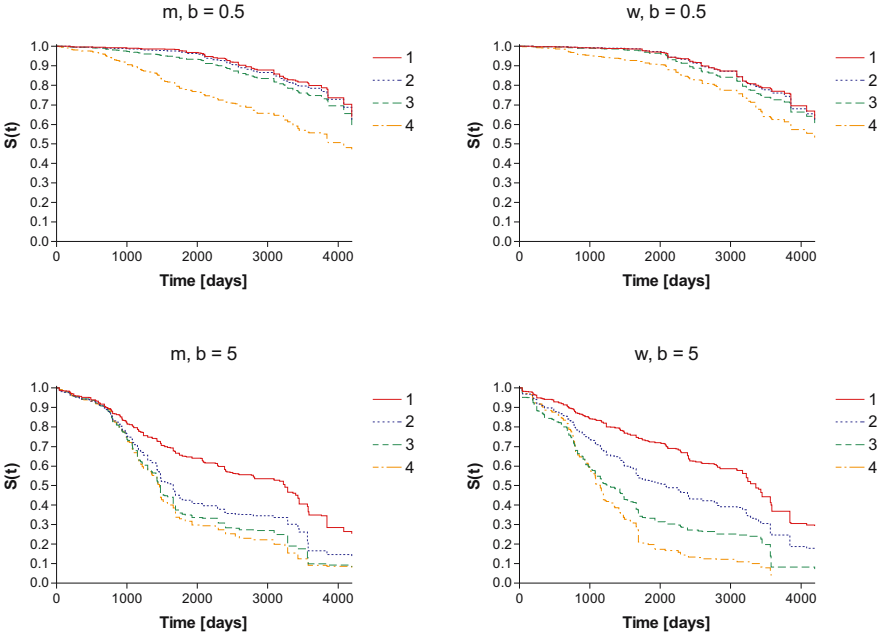

Fig. 3. Kaplan-Meier survival functions for women (w) and men (m) with two different values of serum bilirubin (b=0.5 and b=5) for each histologic stage of disease

for the first and second stage of disease. For the third and fourth stage of disease better survival prediction is for men.

The other analyzed dataset contains the information from the Veteran's Administration (*VA*) lung cancer study [10]. 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), prior therapy, performance status at baseline (Karnofsky rating - KPS), disease duration in months and age in years at randomization, was available.

In Fig. 4 the Brier score for selected times (0, 50, ..., 950 days) and estimated curve of censoring distribution for *VA* lung cancer study are presented. In contrast to the *PBC* dataset, the Brier score values decrease over time. It is due to the small number of censoring cases in the dataset. The shape of function $G(t)$ suggests that there are no censored cases with survival times greater than 250 days.

The analysis aims at discovering the factors that influence the survival. Therapy, KPS and cell type were taken into account. The estimated survival functions for 50 years old patients without prior therapy and five months of disease duration are shown in Fig. 5.

The results suggest that the cell type does not influence the survival, especially when the standard therapy is applied. Median survival times (Table 1) obtained

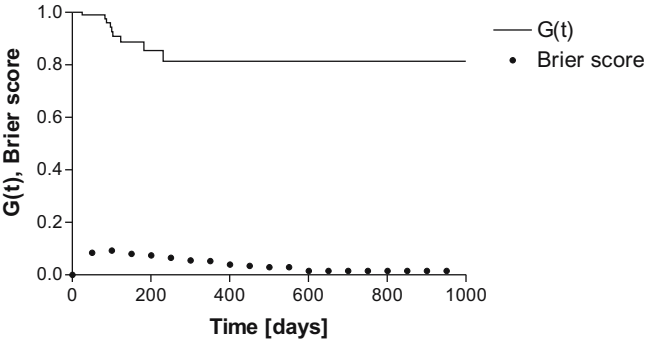

Fig. 4. Kaplan-Meier estimator $G(t)$ of the censoring distribution $G(t) = P(C > t)$ for *VA* lung cancer study; Brier score for selected values of t

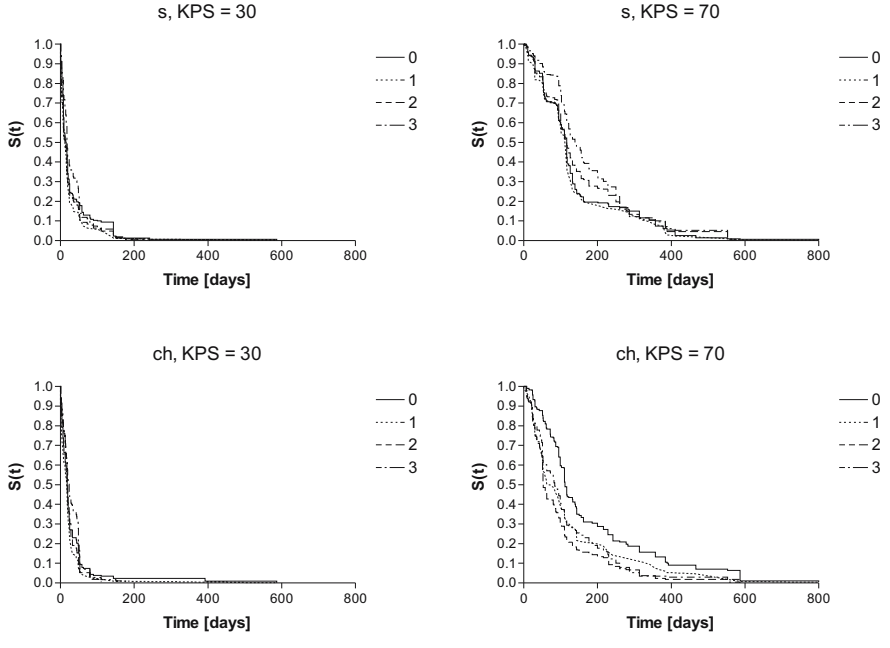

Fig. 5. Kaplan-Meier survival functions for standard (s) and chemotherapy (ch) with two different values of KPS (KPS=30 and KPS=70) for each cell type

for different cell types (with the same values of the remaining attributes) do not differ significantly. The only difference may be observed for patients with chemotherapy with KPS equal to 70. The survival function received for patients with squamous cell type indicates better prognosis (median survival time equal to 112 days) than for other patients (median survival times equal to 72, 53 and 83, respectively).

Table 1. Median survival times for *VA* lung cancer study

	Cell type			
	0	1	2	3
Standard th.				
KPS=30	16	12	13	19
KPS=70	117	112	117	139
Chemotherapy				
KPS=30	19	18	19	24
KPS=70	112	72	53	84

Significant differences between survival curves are visible for different values of KPS. As we can see the survival functions obtained for either standard and test chemotherapy for KPS equal to 30 are similar. The median survival time varies from 12 to 24 days for patients with large cell type and chemotherapy. Much better prognosis is for patients with performance status at baseline equal to 70. The smallest median survival time is for patients with adeno cell type and chemotherapy (53 days) and the best one for individuals with large cell type and standard therapy (139 days).

6 Conclusions

In the paper the random forest consisting of dipolar survival trees was analyzed. The applied method enables calculation of the aggregated Kaplan-Meier survival function for a new patient. The survival tree induction as well as the estimation of aggregated survival function enables using the information from censored cases.

Experiments 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, the other includes the information from the Veteran's Administration lung cancer study. The method was used to analyze the influence of histologic stage and serum bilirubin for the survival of patients with primary biliary cirrhosis of the liver and the influence of therapy, KPS and cell type for survival of patients form *VA* lung cancer study. The goodness of prediction was calculated using Brier score. As it was noticed, the values of the Brier score for the first dataset are increasing over time. It may mean that the prediction is better for short time prognosis, but on the other hand the percentage of censored cases increases over time, what also affects the value of Brier score. The *VA* lung cancer dataset contains only 6% of censored cases with respectively short follow-up time. The Brier score values are increasing over the first, short period of time and then starting to decrease. In this case, the method performs well even for long time prognosis.

Acknowledgements. This work was supported by the grant W/WI/4/05 from Białystok Technical University.

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. Breiman L., How to use survival forest. [URL http://stat-www.berkeley.edu/users/breiman]
3. Ciampi A., Negassa A., Lou, Z., Tree-structured prediction for censored survival data and the Cox model. Journal of Clinical Epidemiology **48**(5) (1995) 675-689
4. Cox D.R., Regression models and life tables (with discussion). Journal of the Royal Statistical Society B **34** (1972) 187–220
5. Davis R. B., Anderson J. R., Exponential survival trees. Statistics in Medicine **8**(1989) 947–961
6. Fleming T.R., Harrington D.P., Counting Processes and Survival Analysis. John Wiley & Sons, Inc. (1991)
7. 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
8. Hothorn T., Lausen B., Benner A., Radespiel-Troger M., Bagging survival trees. Statistics in medicine **23** (2004) 77–91
9. 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)
10. Kalbfleisch J.D., Prentice R.L., The statistical analysis of failure time data. John Wiley & Sons, New York (1980)
11. Kaplan E.L., Meier P., Nonparametric estimation from incomplete observations, Journal of the American Statistical Association **5** (1958), 457–481
12. Krętowska M., Dipolar regression trees in survival analysis. Biocybernetics and biomedical engineering **24** (3) (2004) 25–33
13. Marubini E., Morabito A., Valsecchi G., Prognostic factors and risk groups: Some results given by using an algorithm suitable for censored survival data. Statistics in Medicine **2** (1983) 295–303
14. Ridgeway G., The state of boosting. Computing Science and Statistics **31** (1999) 1722–1731