Biocybernetics and Biomedical Engineering
2007, Volume 27, Number 3, pp. 67–75

Ensembles of Dipolar Trees for Prediction of Survival Time

MAŁGORZATA KRĘTOWSKA*

Białystok Technical University, Faculty of Computer Science, Białystok, Poland

In the paper, the application of random forest for prediction of survival time is presented. The observed data loss function is based on inverse probability of censoring weights. The random forest consists of the sequence of multivariate regression trees created on the base of the learning sets, randomly generated from the given dataset. The applied regression trees use minimization of dipolar criterion function for finding the splits in the internal nodes.

K e y w o r d s : survival analysis, random forest, dipolar criterion function

1. Introduction

The ensembles of several models are a very promising and effective tool for classification and regression. The motivation was a procedure that combines the outputs of many "weak" models to produce a powerful "committee" [1]. The algorithm AdaBoost [2] is widely known in classification problems. The method sequentially fits simple classifiers to different weightings of the observation in a dataset. The observations that were poorly predicted by the previous classifiers receive greater weights in the next iteration. The final result is the weighted average of the outputs received from all the simple classifiers. From the statistical point of view the algorithm is an optimization method for finding the classifier that minimizes a particular exponential loss function.

The method has a straightforward application in regression problems. In such problems the loss function may be defined as the squared-error loss $L(y,f(x)) = (y-f(x))^2$ or the absolute loss $L(y,f(x))=|y-f(x)|$. The population solutions are $f(x)=E(Y/x)$ for the first function and $f(x) = \text{median}(Y/x)$ for the other one [1].

* Correspondence to: Małgorzata Krętowska, Białystok Technical University, Faculty of Computer Science, Wiejska 45a, Białystok, Poland, e-mail: mmac@ii.pb.bialystok.pl

Ensemble methods like bagging [3], random forest [4] and boosting are also used in survival analysis. Their application in prediction problems stabilizes the results, especially in discovering the risk factors. The approach proposed by Ridgeway [5] allows minimizing the partial likelihood function (boosting Cox's proportional hazard model). A new method to calculate the aggregated Kaplan-Meier survival functions is proposed by Hothorn et al. [6]. The Hothorn et al. [7] proposed two approaches for censored data: random forest and gradient boosting. Breiman [8] introduced the software for induction a random forest for censored data.

In the paper the random forest approach proposed by Hothorn et al. [7] is applied to the analysis of survival data. As a single tree, the multivariate dipolar tree proposed by Krętowska [9] is used. The induction of regression tree is based on minimization of a dipolar criterion function, which allows using censored data.

The paper is subdivided as follows: Section 2 describes the survival data, in Section 3 induction of dipolar survival tree is presented, Section 4 contains the random forest algorithm and presents how to calculate the predicted log-survival times for patients. In Section 5 the experimental results, on the base of the Veteran's Administration (VA) lung cancer study and primary biliary cirrhosis trial, are presented. Section 6 summarises 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 $O = (Y = \log(T), \Delta, \mathbf{X})$, where $T = \min(T^0, C)$ is the time to event, $\Delta = I(T \leq C)$ is the censoring indicator and $X = (X_1, \ldots, X_N)$ denotes the set of N covariates from sample space χ. We observe the learning sample $(\boldsymbol{x}_i, y_i = log(t_i), \delta_i)$ $i = 1, 2, \ldots, n$, where $\boldsymbol{x}_i$ is N-dimensional covariate vector, t_i — survival time, and δ_i — failure indicator, which is equal to 0 for censored cases and 1 for uncensored ones.

3. Induction of Dipolar Tree

The proposed dipolar tree [9] is based on the concept of dipoles [10]. The dipole is a pair of different covariate vectors $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the learning set. Mixed and pure dipoles are distinguished. Mixed dipoles are formed between objects that should be separated, while pure dipoles between objects that are similar from the point of view of the analyzed criterion.

The aim is to find such a hyper-plane $H(\boldsymbol{v})$ that would divide possibly a high number of mixed dipoles and possibly a low number of pure ones. It is done by minimization of the dipolar criterion function. Two types of piece-wise linear and convex (CPL) penalty functions $\varphi_i^+(\boldsymbol{v})$ and $\varphi_j^-(\boldsymbol{v})$ are considered:

$$\varphi_j^+(\mathbf{v}) = \begin{cases} \delta^j - < \mathbf{v}, \mathbf{y_j} > & if < \mathbf{v}, \mathbf{y_j} >< \delta^j \\ 0 & if < \mathbf{v}, \mathbf{y_j} > \geq \delta^j \end{cases}$$

$$\varphi_j^-(\mathbf{v}) = \begin{cases} \delta^j + < \mathbf{v}, \mathbf{y_j} > & if < \mathbf{v}, \mathbf{y_j} >> -\delta^j \\ 0 & if < \mathbf{v}, \mathbf{y_j} > \leq -\delta^j \end{cases}$$

where $\boldsymbol{y} = [1, \boldsymbol{x}^T]^T$ is an augmented covariate vector, $\boldsymbol{v} = [-\theta, w_1, w_2, \ldots, w_N]^T$ is an augmented weight vector and δ^j is the margin. Each mixed dipole $(\boldsymbol{y}_i, \boldsymbol{y}_j)$, which should be divided, is associated with a function being a sum of two functions with opposite signs

$$\varphi_{ij}^m(\boldsymbol{v}) = \varphi_i^-(\boldsymbol{v}) + \varphi_j^+(\boldsymbol{v}) \text{ or } \varphi_{ij}^m(\boldsymbol{v}) = \varphi_i^+(\boldsymbol{v}) + \varphi_j^-(\boldsymbol{v}).$$

With pure dipoles, which should remain undivided, we associate a function:

$$\varphi_{ij}^p(\boldsymbol{v}) = \varphi_i^-(\boldsymbol{v}) + \varphi_j^-(\boldsymbol{v}) \text{ or } \varphi_{ij}^p(\boldsymbol{v}) = \varphi_i^+(\boldsymbol{v}) + \varphi_j^+(\boldsymbol{v}).$$

The dipolar criterion function is a sum of penalty functions associated with each dipole:

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

where α_{ij} determines relative importance (price) of the dipole $(\boldsymbol{y}_i, \boldsymbol{y}_j)$, I_p and I_m are the sets of pure and mixed dipoles, respectively. Because $\Psi(\boldsymbol{v})$ is the convex, piece-wise linear function, basis exchange algorithms [11], similar to linear programming, are used as a minimization method.

Hierarchical and sequential structure of a tree recursively partition 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. Each distinct outcome of the test generates one child node, which means that all non-terminal nodes have two or more child nodes. A terminal node generates no descendant.

The proposed method of regression tree induction aims at dividing the feature space into such areas, which would include patients with similar survival times. It may be done by appropriate rules of dipoles formation. 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:

a) a pair of feature vectors $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ forms the pure dipole, if

- $\delta_i = \delta_j = 1$ and $| t_i - t_j | < \eta$;

b) a pair of feature vectors (x_i, x_j) forms the mixed dipole, if

- $\delta_i = \delta_j = 1$ and $| t_i - t_j | > \zeta$;
- $(\delta_i = 0, \delta_j = 1$ and $t_i - t_j > \zeta)$ or $(\delta_i = 1, \delta_j = 0 \wedge t_j - t_i > \zeta)$.

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 the dipolar criterion function, starting from the root. The function in a given node is designed on the base of those feature vectors that 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 reaches the node consists of less than 5 uncensored cases. The improvement of generalizing ability may be received by using the pruning technique, which applies modified rank correlation coefficient D [12] to assess the quality of the tree.

4. Random Forest Algorithm

The method proposed by Hothorn et al. [7] defines the observed data loss function as the application of inverse probability of censoring weights:

$$L(Y,\varphi(X) \mid G) = L(Y,\varphi(X))\frac{\Delta}{G(T|X)}$$

where $G(T|X) = \mathrm{P}(C > T|X)$ is the conditional censoring survival function and φ: $X \rightarrow R$ is the candidate estimator of Y.

The corresponding empirical risk is the weighted average:

$$\hat{E}_{Y,\Delta,X} L(Y,\varphi(X) \mid G) = n^{-1} \sum_{i=1}^{M} L(y_i,\varphi(x_i))\frac{\delta_i}{H(t_i|x_i)} \tag{1}$$

Conditional censoring survival function $G(T|X)$ is unknown and needs to be replaced by an estimate H. In the approach applied in the paper, the function is estimated by using the Kaplan-Meier survival function.

The random forest algorithm minimizes the empirical risk (1) by induction of many single trees based on the bootstrap samples from the learning set. For each observation from the learning set, the weight $w_i = \delta_i H^{-1}(t_i|x_i)$ is calculated. Weight w_i is interpreted as the resampling probability of observation i.

The random forest algorithm:

1. Set $p = 1$ and fix $P > 1$.

2. Draw a random vector of case counts $v_p = [v_{p1},..., v_{pM}]^T$ from the multinomial distribution with parameters n and $\mathbf{w}(\mathrm{sum}_i\, \mathrm{w}_i)^{-1}$.

3. Construct single survival tree with $K(p)$ terminal nodes $R_{p1},..., R_{pK(p)}$ using bootstrap sample 4. p: = p + 1 and repeat step 2 and 3 until $p = P$.

For prediction the log-survival time for a patient with covariate x, the prediction weight $a_i(x)$ is assigned to each observation i from the learning set. The prediction weight measures the similarity of x to x_i:

$$a_i(x) = \sum_{p=1}^{P} v_{pi} \sum_{k=1}^{K(p)} I(x_i \in R_{pk} \text{ and } x \in R_{pk}); i = 1,...,M.$$

Taking into account quadratic loss the prediction is the weighted average of the observed log-survival times

$$\hat{y} = \hat{E}(Y / X = x) = \left(\sum_{i=1}^{M} a_i(x) \right)^{-1} \sum_{i=1}^{M} a_i(x) y_i.$$

It is worth noting that w_i and $a_i(x)$ are zero for censored observations.

5. Experimental Results

The first analyzed dataset contains information from the Veteran's Administration (VA) lung cancer study [13]. 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 the cell type, prior therapy, performance status at baseline (Karnofsky rating — KPS), disease duration in months and age in years at randomization was available.

Figure 1 shows mean-difference plots and scatterplots for two random forests with $P = 50$ and $P = 200$ single dipolar trees. The weighed mean for the first model ($P = 50$) equals 4.199 and for the other one – 4.24. The average sum of squares of differences between predicted and observed values is 0.55 and 0.51, respectively.

As it can be seen in the scatterplots, the method performs quite well for the cases with the observed survival time around its median. For the observations with relatively small or large survival times, the difference between the predicted and observed survival times is greater. It may be also noticed that the diversity of predicted values is greater for the observed log-survival times below 3.

In Figures 2, 3 and 4 the log-survival time for different values of age and KPS is presented. The patients with squamous cell type, without prior therapy and with 5 months of disease duration were taken into account. In Figure 2 and 3 it can be seen that smaller values of KPS decrease the survival time, either for standard and chemotherapy. Figure 4 shows that the survival time for patients with chemotherapy is better for greater values of KPS. In all other cases the use of the standard therapy leads to better prediction.

The other dataset is the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 [14]. A total of 424 PBC patients, referred to the Mayo Clinic during that ten-year interval, met the eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases

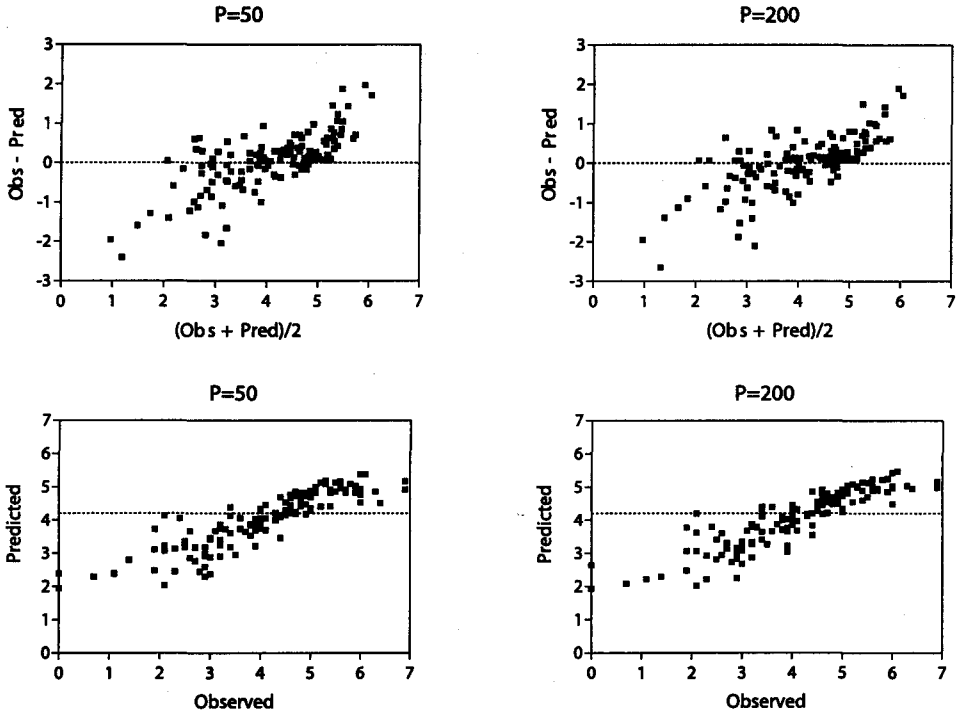

Fig. 1. Mean-difference plots and scatterplots of observed and predicted log-survival time of random forest for P = 50 and P = 200 for VA lung cancer data. The dashed horizontal line (bottom) is the weighted mean of the log-survival time

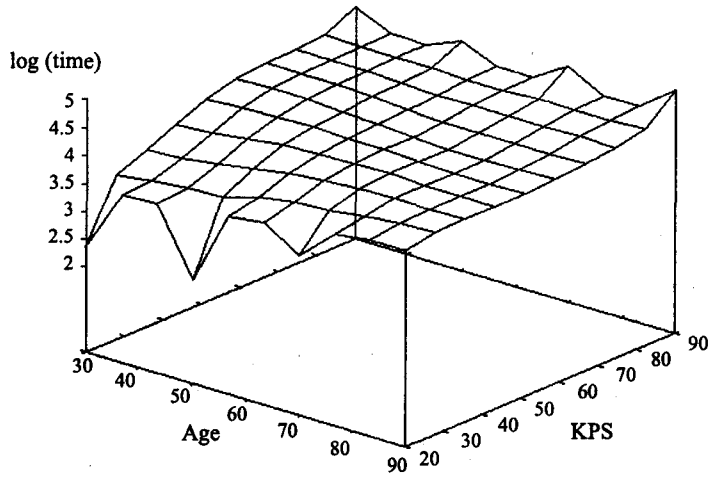

Fig. 2. Log-survival time for patients with standard therapy

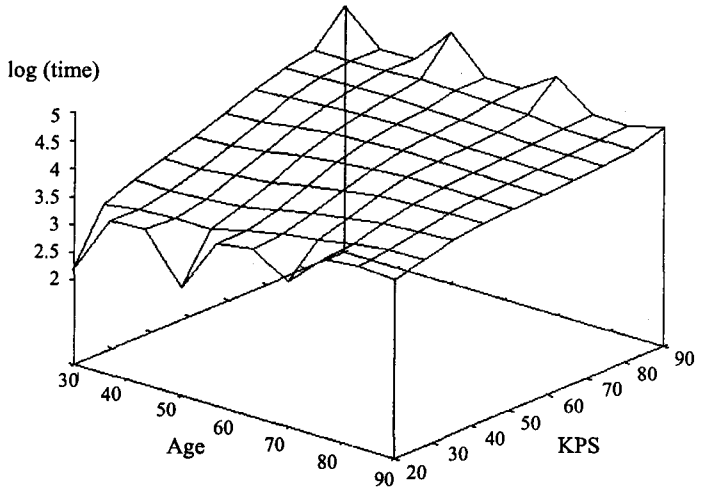

Fig. 3. Log-survival time for patients with chemotherapy

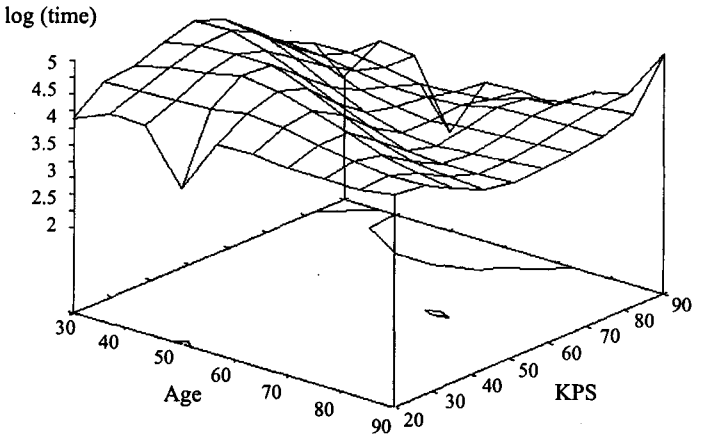

Fig. 4. Difference between log-survival times for patients with standard and chemotherapy

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 up 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. The data contain 257 censored observations.

Figure 5 shows mean-difference plots and scatterplots for the random forests with P = 100 single trees. Only uncensored observations are visible in the figure. As it can be seen, in the presence of many censored cases the prediction is comparable to the previous example. The weighed mean for the model equals 7.03 and the average sum of squares of differences between the predicted and observed values is –0.2. The differences are particularly significant for the observations with smaller survival times.

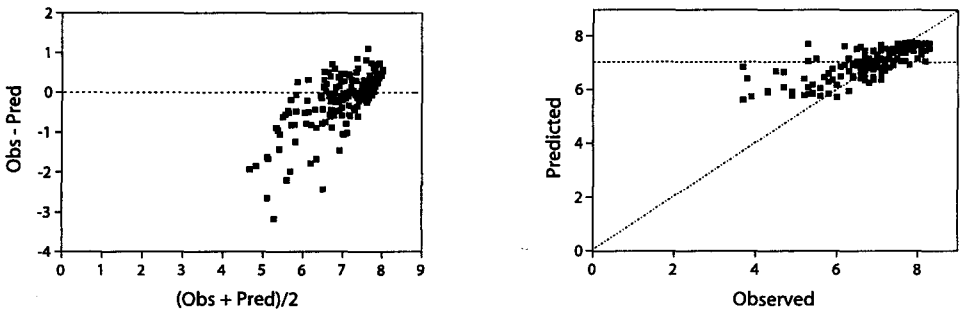

Fig. 5. Mean-difference plots and scatterplots of observed and predicted log-survival time for PBC data. The dashed horizontal line (on the left) is the weighted mean of log-survival time

6. Conclusions

In the paper, the application of the random forest for log-survival time prediction is presented. The approach creates a sequence of simple dipolar trees and the final results are calculated as the weighted average of the results received from all the trees. The method performed better than single regression tree, and it was shown that the results of the random forest with a greater number of single trees ($P = 200$) were similar to the results of the model with $P = 50$. The preliminary analysis of two datasets, with 6.6 and 61.5 percent of censored cases, respectively shows that the number of censored cases does not significantly affect the prediction.

Acknowledgment

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

References

1. Hastie T., Tibshirani R., Friedman J.: The elements of statistical learning. Data mining, inference and prediction, Springer-Verlag, 2001.

2. Freund Y., Schapire R.: A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 1997, 55, 1, 119–139.
3. Breiman L.: Bagging predictors, Machine Learning, 1996, 24, 123–140.
4. Breiman L.: Random forests, Machine Learning, 2001, 45, 5–32.
5. Ridgeway G.: The state of boosting, Computing Science and Statistics, 1999, 31, 1722–181.
6. Hothorn T., Lausen B., Benner A., Radespiel-Troger M.: Bagging survival trees, Statistics in medicine, 2004, 23, 77–91.
7. Hothorn T., Buhlmann P., Dudoit S., Molinaro A. M., van der Laan M. J.: Survival ensembles, U.C. Berkeley Division of Biostatistics Working Paper Series, 174, 2005 URL http://www.bepress.com/ucbbiostat/paper174.
8. Breiman L.: How to use survival forest. URL http://stat-www.berkeley.edu/ users/breiman.
9. Krętowska M.: Dipolar regression trees in survival analysis, Biocybernetics and Biomedical Engineering, 2004.
10. Bobrowski L., Krętowska M., Krętowski M.: Design of neural classifying networks by using dipolar criterions, Proceedings of the Third Conference "Neural Networks and their Applications", Częstochowa, Poland, 1997, 689–694.
11. Bobrowski L.: Design of piecewise linear classifiers from formal neurons by some basis exchange technique, Pattern Recognition, 1991, 24, 9, 863–870.
12. Korn E. L., Simon R.: Measures of explained variation for survival data, Statistics in medicine, 1990, 9, 487–503.
13. Kalbfleisch J.D., Prentice R.L.: The statistical analysis of failure time data. John Wiley & Sons, New York 1980.
14. Fleming T.R., Harrington D.P.: Counting Processes and Survival Analysis. John Wiley & Sons, Inc., 1991.