The Influence of Censoring for the Performance of Survival Tree Ensemble

Małgorzata Krętowska

Faculty of Computer Science
Bialystok University of Technology
Wiejska 45a, 15-351 Białystok, Poland
m.kretowska@pb.edu.pl

Abstract. One of the main objectives in survival analysis is prediction the time of failure occurrence. It is done on a base of learning sets, which contain incomplete (censored) information on patients failure times. Proposed predictors should allow to cope with censored data. In the paper the influence of censoring for the performance of dipolar tree ensemble was investigated. The prediction ability of the model was verified by several measures, such as direct and indirect estimators of absolute predictive errors: $\tilde{D}_{S,x}$, $\hat{D}_x$ and explained variation. The analysis is conducted on the base of artificial data, generated with different values of censoring rate.

1 Introduction

Censoring is a term used in survival analysis. It describes the data with incomplete information about failure time. In survival data, each patient is characterized by the feature vector $\mathbf{x}$ and corresponding survival time t, which is counted from the beginning event (e.g. surgery). If we define the event of interest (e.g death, disease relapse) and the cut-off date (how long the patients will be under investigation), time t has two different meanings:

- *failure time* - if the observation is finished with the event of interest,
- *follow-up time* - if the patient is observed to the cut-off date or the observation is finished by another event not connected directly with the aim of medical experiment

Observations, for which only the follow-up time is given are called *censored*. They do not contain the exact knowledge of the failure time. We only know, how long the patient was observed. To distinguish between the failure and the follow-up time, a new binary variable is introduced - *failure indicator* δ, which is equal to 0 for censored cases, and 1 otherwise.

The presence of censored cases causes some problems in analysis of survival data. Because the percentage of censored observations may be high (e.g. Melanoma Malignum dataset [1] - 72%, Primary Biliary Cirrhosis dataset [5] - 60% of censored observations), they should not be ignored. There exist a number

L. Rutkowski et al. (Eds.): ICAISC 2010, Part II, LNAI 6114, pp. 524–531, 2010.

of statistical methods which were developed to analyze censored data (e.g. Cox's proportional hazards model [3]), but they require some additional assumptions to fulfill. If the requirements are difficult to obey, other non-statistical techniques are proposed. Here, neural networks and regression trees are the most common ones. Such methods do not always are adapted directly to censoring data. In many cases the authors use standard tools with omitting the majority of censored cases or with earlier estimation of the failure time for censored cases, so called *imputation*. Lapuerta *et al.* [11] proposed an additional neural network structure for failure time estimation, the use of Cox's regression model is introduced by De Laurentiis and Ravdin [4] and Ripley [12]. Ignoring the censored cases as well as the imputation process allow receiving the biased results. In the first case, only the part of information is used, in the other, new information is added to the data.

In the paper the influence of censoring on the performance of dipolar based ensemble is investigated. The dipolar tree ensemble [8] as well as the neural network ensemble [9], use censored cases during the learning process, while creating the dipolar criterion function. Because better prediction ability was received for dipolar tree ensemble [10], this model is taken into account. The analysis is conducted on the base of artificial datasets, generated with different values of censoring rate ($0 - 70\%$). Predictive ability of the models is evaluated using the indirect an direct estimators of absolute predictive errors and explained variation ([14,13]).

The paper is organized as follows. Section 2 describes the distribution functions of failure time and introduces the idea of Kaplan-Meier survival function. In Section 3 the idea of dipoles and dipolar tree ensemble is presented. Measures of predictive ability for censored data are described in Section 4. Experimental results are presented in Section 5. Section 6 summarizes the results.

2 Distribution Functions of Survival Time

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 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 may be described by several functions:

- survival function
$$S(t) = P(T > t) \tag{1}$$
where $P(\bullet)$ means probability, $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$
- density function
$$f(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t)}{\triangle t} \tag{2}$$

where $f(t)dt$ is the unconditional probability of failure in the infinitesimal interval $(t, t+dt)$.

– hazard function

$$\lambda(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \tag{3}$$

where $\lambda(t)dt$ is the probability of failure in the in infinitesimal interval $(t, t+dt)$, given survival at time t.

The estimation of survival function $S(t)$ may be done by using the Kaplan-Meier product limit estimator [6], which is calculated on the base of learning sample L and is denoted by $\hat{S}(t)$:

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(\frac{m_j - d_j}{m_j} \right) \tag{4}$$

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 Survival Tree Ensemble

Individual survival tree being a part of the complex predictor [8] is a kind of binary regression tree. Each 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 $H(\mathbf{w}, \theta) = \{(\mathbf{w}, \mathbf{x}) :< \mathbf{w}, \mathbf{x} >= \theta\}$.

Establishing the structure of the tree (the number of internal nodes) and the values of hyper-planes parameters $(\mathbf{w}, \theta)$ are based on the concept of dipoles [2]. 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. 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
   - $\delta_i = \delta_j = 1$ and $|t_i - t_j| < \eta$
2. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   - $\delta_i = \delta_j = 1$ and $|t_i - t_j| > \zeta$
   - $(\delta_i = 0, \delta_j = 1$ and $t_i - t_j > \zeta)$ or $(\delta_i = 1, \delta_j = 0$ and $t_j - t_i > \zeta)$

Parameters η and ζ are equal to quartiles of absolute values of differences between uncensored survival times. Basing on the earlier experiments, the parameter η is fixed as 0.2 quartile and ζ - 0.6. An example of dipoles construction by censored and uncensored observation is presented in Fig. 1

The increasing number of censored cases may decrease the number of pure dipoles as well as the mixed ones.

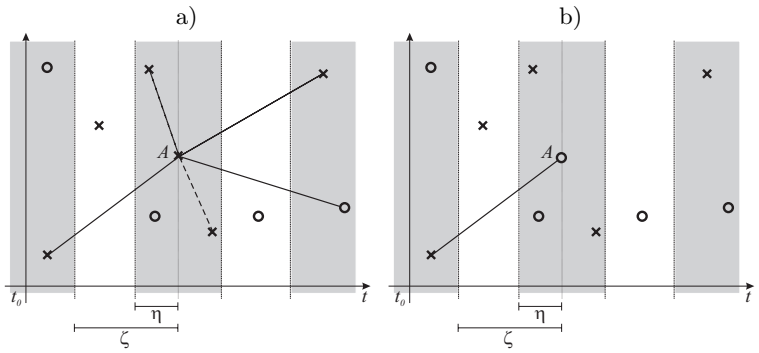

Fig. 1. Construction of pure (solid line) and mixed (dotted line) dipoles by a) uncensored observation; b) censored observation

The hyper-planes $H(\mathbf{w}, \theta)$ in the internal nodes of a tree are calculated by minimization of dipolar criterion function (detailed description may be fond in [8]). This is equivalent with division of possibly high number of mixed dipoles and possibly low number of pure ones constructed for a given dataset. The tree induction algorithm starts from the root, so in the root node, the dipolar criterion function is calculated on the base of dipoles created for the whole learning set. The dipolar criterion function for consecutive nodes of a tree are designed on the base on 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 reach the node consists less than 5 uncensored cases.

The survival tree ensemble algorithm leading to receive the aggregated survival function $\hat{S}(t|\mathbf{x}_n)$ 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. For each tree $T(L_i)$, distinguish the set of observations $L_i(\mathbf{x}_n)$ which belongs to the same terminal node as $\mathbf{x}_n$
4. Build aggregated sample $L_A(\mathbf{x}_n) = [L_1(\mathbf{x}_n), L_2(\mathbf{x}_n), \ldots, L_k(\mathbf{x}_n)]$
5. Compute the Kaplan-Meier aggregated survival function for a new observation $\mathbf{x}_n$ as $\hat{S}_A(t|\mathbf{x}_n)$.

The predicted value of exact failure time for observation $\mathbf{x}_n$ may be calculated as the median value of $\hat{S}_A(t|\mathbf{x}_n)$.

4 Evaluation of Predictive Ability

Predictive ability of the model is calculated using the measures adapting for censoring. One of them is direct estimator of absolute predictive error (APE) [14], calculated for each distinct failure time $t_{(j)}$:

$$\hat{M}(t_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}\Big[I(t_i > t_{(j)})(1-\hat{S}(t_{(j)})) + \delta_i I(t_i \le t_{(j)})\hat{S}(t_{(j)}) + (1-\delta_i)I(t_i \le t_{(j)})\left\{(1-\hat{S}(t_{(j)}))\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)} + \hat{S}(t_{(j)})(1-\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)})\right\}\Big] \quad (5)$$

where $I(condition)$ is equal to 1 if the condition is fulfilled and 0 otherwise. 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:

$$\hat{D} = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \hat{M}(t_{(j)}) \quad (6)$$

$$\hat{D}_{\mathbf{x}} = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \hat{M}(t_{(j)}|\mathbf{x}) \quad (7)$$

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. It is calculated on the base of observations $(t_i, 1-\delta_i)$.

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

$$\tilde{M}(t_{(j)}) = 2\hat{S}(t_{(j)})(1-\hat{S}(t_{(j)})) \quad (8)$$

$$\tilde{M}(t_{(j)}|\mathbf{x}) = 2n^{-1}\sum_i \hat{S}(t_{(j)}|\mathbf{x}_i)(1-\hat{S}(t_{(j)})|\mathbf{x}_i) \quad (9)$$

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 (EV) can be defined as $\tilde{V}_S = \frac{\tilde{D}_S - \tilde{D}_{S,\mathbf{x}}}{\tilde{D}_S}$ and $\hat{V} = \frac{\hat{D}-\hat{D}_{\mathbf{x}}}{\hat{D}}$.

5 Experimental Results

The influence of censoring for the performance of survival tree ensemble was evaluated on the base of several simulated datasets. An exponential survival distribution was assumed for the proportional hazards models. Let

$$\lambda(t,\mathbf{x}) = \exp\left\{\sum_{i=1}^{N}\beta_i(t)x_i + \sum\sum_{i\neq j}\gamma_{ij}(t)x_i x_j + \sum_{i\neq j\neq k}\gamma_{ijk}(t)x_i x_j x_k\right\} \quad (10)$$

be the hazard at any time t given N covariates $\mathbf{x}$, the survival times were then generated using inverse probability transformations [15]. The following datasets were considered:

- (a) $N = 2$, with $\beta_1 = 0.25$, $\beta_2 = 0.5$; x_1 and x_2 have normal distribution $N(0, 1)$
- (b) the same as in (a), except that an interaction $\gamma_{12} = 0.2$ was assumed
- (c) $N = 4$, with $\beta_1 = 1$, $\beta_2 = 0.25$, $\beta_3 = 1$, $\beta_4 = 0.5$; x_1, x_2 and x_3 have normal distribution $N(0, 1)$ and x_4 have Bernoulli distribution with $n = 1$ and $p = 0.5$.
- (d) the same as in (c), except that interactions $\gamma_{12} = \gamma_{23} = 0.2$ and $\gamma_{123} = \gamma_{234} = 0.5$ were assumed

Each dataset was generated with 0, 10, 20, 30, 40, 50, 60 and 70 per cent of censoring, number of cases $n = 400$. Censoring time was exponentially distributed and independent of the survival time.

All the experiments were performed using the ensemble of 100 survival trees ST. 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 learning set L, only such ST_i $(i = 1, 2, \ldots, 100)$ were taken into consideration, for which $\mathbf{x}$ was not belonged to learning set L_i (i.e. $\mathbf{x}$ did not participate in the learning process of the ST_i). On average, each learning set L_i do not contain $1/3$ elements from L [7].

In figure 2 the values of described earlier predictive measures for the generated datasets are presented. In each case values of direct and indirect APE without covariates were the same ($dAPE = iAPE$), so in the figure only the indirect measure is shown. The influence of censoring rate for the predictive accuracy is similar for each data. As we can see, relatively small values of censoring rate do not influence much the predictive ability of the model. For datasets without interactions ((a), (b)) we can observe similar values of explained variation for data with censoring rate less then 30%. For dataset (a) the values of indirect (direct) explained variance is equal to 0.95(0.97), 0.9(0.91), 0.93(0.94), respectively for 0, 10 and 20 censoring rate, for dataset (c) we receive 0.93(0.96), 0.89(0.9), 0.93(0.96), respectively. For higher values of censoring rate the predictive ability of the model is poor. The value of absolute predictive error increases and hence explained variation is getting close to 0.

Similar behavior is observed for the datasets with interactions. The predictive ability of received models for the censoring rates between 0 and 30% is also high and equal to 0.85(0.89), 0.82(0.85), 0.8(0.83) and 0,84(0.87) for dataset (b) and 0.73(0.78), 0.74(0.78), 0.75(0.78), 0.76(0.8) for dataset (d). The increasing value of censoring rate decreases the predictive ability of the model.

Analyzing the results received for the models build for data with high percentage of censored cases, it was noticed that for many observations the predicted failure times were undefined. The failure time for a given patient $\mathbf{x}_{new}$ is calculated as the median value of the received aggregated Kaplan-Meier survival function. If the number of censored cases is high, the values of $\hat{S}(t|\mathbf{x}_{new})$ are above 0.5, so the median value can not be calculated.

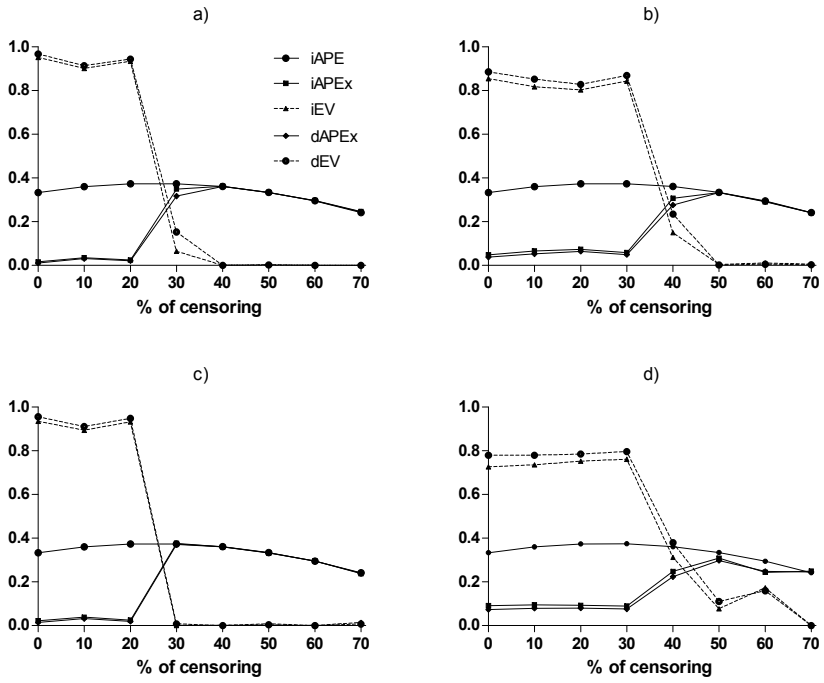

Fig. 2. Predictive measures for generated datasets with different value of censoring rate

6 Conclusions

In the paper the influence of censoring for the performance of dipolar tree ensemble was investigated. The prediction ability of the models was verified by several measures, such as direct and indirect estimators of absolute predictive errors: $\tilde{D}_{S,x}$, $\hat{D}_x$ and explained variation. The results received for artificial datasets, generated with different number of censoring rates ($0\% - 70\%$), suggest that the number of inputs or the presence of interaction do not influence substantially the performance of the predictor. Dipolar tree ensembles generated for the datasets with relatively small number of censored cases (less then 30% or 40% per cent, for datasets without and with interactions respectively) have very good predictive ability. For higher values of censoring rate the explained variation was decreasing. Because the high number of censored cases makes the exact prediction of the failure time impossible, the patient survival should be described by the aggregated Kaplan-Meier survival function.

Acknowledgements. This work was supported by the grant W/WI/4/08 from Bialystok Technical University.

References

1. Andersen, P.K., Borgan, O., Gill, R.D.: Statistical Models based on Counting Processes. Springer, New York (1993)
2. Bobrowski, L., Krętowska, M., Krętowski, M.: Design of neural classifying networks by using dipolar criterions. In: Proc. of the Third Conference on Neural Networks and Their Applications, Kule, Poland, pp. 689–694 (1997)
3. Cox, D.R.: Regression models and life tables (with discussion). Journal of the Royal Statistical Society B 34, 187–220 (1972)
4. De Laurentiis, M., Ravdin, P.M.: Survival analysis of censored data: neural network analysis detection of complex interactions between variables. Breast Cancer Research and Treatment 32(1), 113–118 (1994)
5. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. John Wiley & Sons, Inc, Chichester (1991)
6. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 5, 457–481 (1958)
7. Koronacki, J., Cwik, J.: Statistical learning systems. Wydawnictwa Naukowo-Techniczne, Warsaw (2005) (in Polish)
8. Krętowska, M.: Random forest of dipolar trees for survival prediction. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 909–918. Springer, Heidelberg (2006)
9. Krętowska, M.: Ensemble of dipolar neural networks in application to survival data. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2008. LNCS (LNAI), vol. 5097, pp. 78–88. Springer, Heidelberg (2008)
10. Krętowska, M.: Prognostic abilities of dipoles based ensembles comparative analysis. Zeszyty Naukowe Politechniki Bialostockiej. Informatyka 4, 73–83 (2009)
11. Lapuerta, P., Azen, S.P., LaBree, L.: Use of neural networks in predicting the risk of coronary artery disease. Computers and Biomedical Research 28, 38–52 (1995)
12. Ripley, R.M.: Neural networks for breast cancer prognosis. PhD thesis, Department of Engineering Science, University of Oxford (1998)
13. Schemper, M.: Predictive accuracy and explained variation. Statistics in Medicine 22, 2299–2308 (2003)
14. Schemper, M., Henderson, R.: Predictive accuracy and explained variation in Cox regression. Biometrics 56, 249–255 (2000)
15. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J., Azen, S.: Comparison of the performance of neural network methods and Cox regression for censored survival data. Computational Statistics & Data Analysis 34, 243–257 (2000)