Competing Risks and 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. In the paper the ensemble of dipolar trees for analysis of competing risks is proposed. The tool is build on the base of the learning sets, which contain the data from clinical studies following patients response for a given treatment. In case of competing risks many types of response are investigated. The proposed method is able to cope with incomplete (censored) observations and as a result, for a given set of covariates and a type of event, returns the aggregated cumulative incidence function.

1 Introduction

Survival analysis, in its basic form, aims at prediction the time of failure occurrence. The failure, in medical domain, usually means death or disease relapse. Analyzing the survival data we are often interested not only in estimation of the exact time point when the failure would occur for a given patient, but we also want to observe the failure probability during a given period of time. It may be done by estimation of a distribution function (e.g. survival function, hazard function or cumulative incidence function).

In the presence of competing risks we are not only focused on prediction of one type of event. In such kind of data there are many types of event, which may be investigated. For one patient we register only the time of the first failure occurred. If the patient has not experienced any type of event we register only the follow-up time. This kind of observation is called censored one. Censored data contain incomplete information about the time of failure occurrence of any type of event. For such kind data we only know that the failure time is greater or equal to given follow-up time.

Since the survival data is to a large extent censored, the crucial point of methods for failure time prognosis, is using the information from censored cases. The use of ensemble of simple tree structure predictors is very common, recently. Hothorn *et al.* [4] proposes boosting survival trees to create aggregated survival function. Krętowska [7] developed the approach by using the dipolar regression trees instead of the structure proposed in [4]. The technique proposed by Ridgeway [11] allows minimizing the partial likelihood function (boosting Cox's proportional hazard model). The Hothorn *et al.* [5] developed two approaches for censored data: random forest and gradient boosting. Breiman [2] provided the software that allows induction of random forest for censored data.

L. Rutkowski et al. (Eds.): ICAISC 2012, Part I, LNCS 7267, pp. 387–393, 2012.

In case of competing risk the authors develop the tools on the base of single survival tree. Similar approaches are described in [3] and [6], where induction of the proposed between-node tree is based on the difference between cumulative incidence function. Presented in [3] a within-node tree uses event-specific martingale residuals.

The results received from the single tree are instable. It means that each tree inducted for the same data produces different outcomes. The application of ensemble based methods stabilizes the results. In the paper, the survival tree ensemble proposed in [8] is modified for competing risks data. The modifications are mainly connected with the way the dipoles are formed during a single survival tree induction and the way the results are presented. The idea of aggregated cumulative incidence function is introduced as a result of survival tree ensemble. The results are presented on the base of follicular type lymphoma data, which contain 541 observations.

The paper is organized as follows. Section 2 describes the survival data with competing risks and introduces the idea of cumulative incidence function as well as the Kaplan-Meier survival function. In Section 4 induction of single dipolar survival tree and ensemble of survival tree are presented. The algorithm of calculation of aggregated cumulative incidence function is described. Experimental results are presented in Section 4. They were carried out on the base of real dataset describing the patients with follicular type lymphoma [9]. Section 5 summarizes the results.

2 Survival Data with Competing Risks

In case of survival data with competing risks, the patient is at risk of p ($p > 1$) different types of failure. Assuming that the time of occurrence of the ith failure is T_i, we are interested only in the failure with the shortest time $T = \min(T_1, T_2, \ldots, T_n)$. The learning sample L for competing risk data is defined as $L = (\mathbf{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, where $\mathbf{x}_i$ is N-dimensional covariates vector, t_i is time to the first event observed and $\delta_i = \{0, 1, \ldots, p\}$ indicates the case of failure. δ_i equals to 0 represents censored observation, which means that for a given patient has not occurred any failure. Variable t_i represents the follow-up time.

The distribution of the random variable T (time), for an event of type i ($i = 1, 2, \ldots, p$) may be represented by several functions. One of the most popular is cumulative incidence function (CIF) defined as the probability that an event of type i occurs at or before time t [10]:

$$F_i(t) = P(T \leq t, \delta = i) \tag{1}$$

or survival function:

$$S_i(t) = P(T > t, \delta = i) \tag{2}$$

The estimator of the CIF function is calculated as

$$\hat{F}_i(t) = \sum_{j|t_j \leq t} \frac{d_{ij}}{n_j} \hat{S}(t_{j-1}) \tag{3}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered uncensored time points from the learning sample L, d_{ij} is the number of events of type i at time $t_{(j)}$, n_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 $t_{(j)}$) and $\hat{S}(t)$ is the Kaplan-Meier estimator of the probability of being free of any event by time t. It is calculated as:

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

where d_j is the number of events at time $t_{(j)}$.

The "patients specific" cumulative incidence function for the event of type i is given by $F_i(t|\mathbf{x}) = P(T \leq t, \delta = i|\mathbf{X} = (x))$. The conditional CIF for the new patient with covariate vector $\mathbf{x}_{new}$ is denoted by $\hat{F}_i(t|\mathbf{x}_{new})$.

3 Survival Tree Ensemble

Individual survival tree being a part of the complex predictor [7] 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 [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. Assuming that the analysis aims at dividing the feature space into such areas, which would include the patients with the same case of failure and similar survival times, pure dipoles are created between pairs of feature vectors with the same failure type, 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
   - $\delta_i \neq 0$ and $\delta_i = \delta_j = z$ and $|t_i - t_j| < \eta_z$, $z = 1, 2, \ldots, p$.
2. a pair of feature vectors $(\mathbf{x}_i, \mathbf{x}_j)$ forms the mixed dipole, if
   - $\delta_i \neq 0$ and $\delta_i = \delta_j = z$ and $|t_i - t_j| > \zeta_z$, $z = 1, 2, \ldots, p$
   - $(\delta_i = 0, \delta_j = z$ and $t_i - t_j > \zeta_z)$ or $(\delta_i = z, \delta_j = 0$ and $t_j - t_i > \zeta_z)$, $z = 1, 2, \ldots, p$

Parameters η_z and ζ_z are equal to quartiles of absolute values of differences between uncensored survival times for zth type of failure, $z = 1, 2, \ldots, p$. Basing on the earlier experiments, the parameter η_z is fixed as 0.3 quantile and ζ_z - 0.6.

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

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 [7]). 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 reaches the node consists of less than 5 uncensored cases.

The survival tree ensemble algorithm leading to receive the aggregated cumulative incidence function $\hat{F}_i(t|\mathbf{x}_n)$ 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 $\hat{S}^A(t|\mathbf{x}_n)$.
6. Compute the aggregated CIF functions for the ith type of failure for a new observation $\mathbf{x}_n$ as $\hat{F}_i^A(t|\mathbf{x}_n)$.

The predicted value of exact time of ith type of failure for observation $\mathbf{x}_n$ may be calculated as the median value of $\hat{F}_i^A(t|\mathbf{x}_n)$.

4 Experimental Results

The experiments were done on the base of lymphoma patient data, which was created at Princess Margaret Hospital, Toronto [9]. In the experiments we use the subset of 541 patients having follicular type lymphoma, registered for treatment at the hospital between 1967 and 1996, with early stage disease (I or II) and treated with radiation alone or with radiation and chemotherapy. Each patient is described by four covariates, having the following characteristics:

age [years]: $Q_1 = 47$, $Me = 58$, $Q_3 = 67$
haemoglobin [g/l]: $Q_1 = 130$, $Me = 140$, $Q_3 = 150$
clinical stage : equal to 1 (66.9% observations) or 2
chemotherapy : 0-no (78.2% observations), 1-yes

where Q_1 is the lower quartile, Me - median, Q_3 - upper quartile.

The goal of this study was to report long-term outcome in this group of patient. The event of interest is failure from the disease: no response to treatment or relapse. Competing risk type of event is death without failure. There are 272 event of interest and 76 observations with death without relapse.

All the experiments were performed using the ensemble of 100 survival trees.

In figure 1 we can observe the cumulative incidence functions which were calculated for the patient described by the covariates equal to their median values:

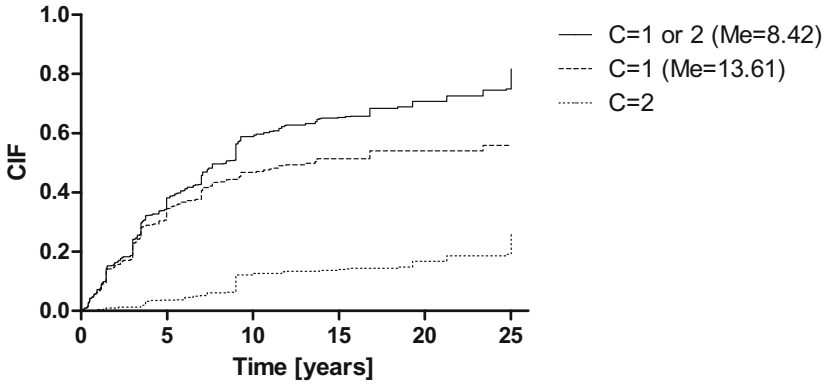

Fig. 1. CIF functions for disease failure (C=1), for competing risk (C=2) and without distinguishing two types of failure (C=1 or 2) (age=58, hgb=140, clinstg=1 and ch=0)

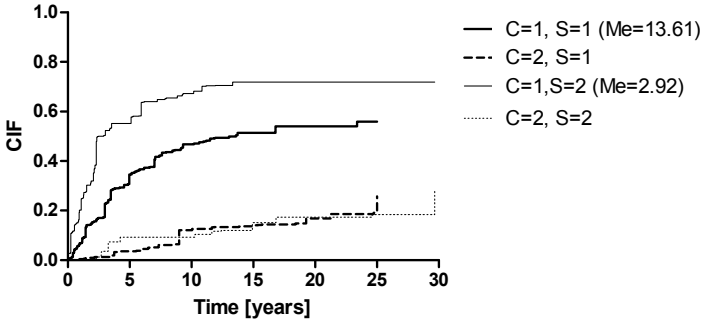

Fig. 2. CIF functions for disease failure (C=1) and for competing risk (C=2) for two values of clinical stage (S=1 and S=2) (age=58, hgb=140, ch=0)

age=58, hgb=140, clinstg=1 and ch=0. Three CIF functions are presented in figure 1: CIF function for disease failure (C=1), for competing risk (C=2) and the third one calculated for two types of failure (C=1 or 2). We can observe here that for each time point the probability of competing risk (death without failure) is lower than the probability of failure. The median value calculated for the event of interest is equal to 13.61, the median value for competing risk do not exist ($\hat{F}_2(t)$ is always less than 0.5).

Figure 2 presents the CIF functions for two values of clinical stage, other covariates were fixed to theirs median values. The functions calculated for competing risk do not differ significantly for two types of clinical stage. The difference is visible while comparing the functions calculated for disease failure. The prediction is worse for patient with clinical stage II (median value equal to 2.92 [years]), for stage I median value is 13.61 [years].

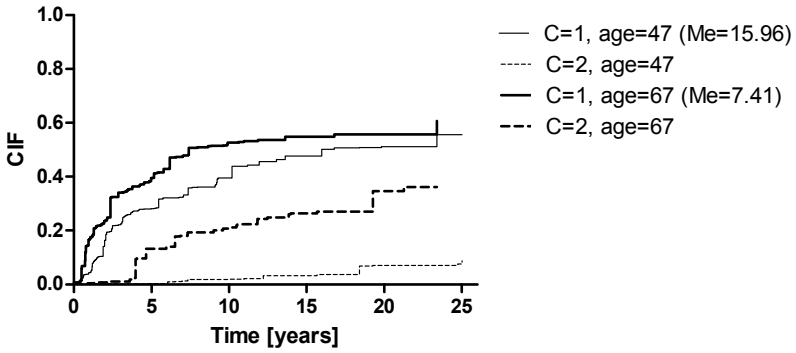

Fig. 3. CIF functions for disease failure (C=1) and for competing risk (C=2) for two values of age (hgb=140, clinstg=1, ch=0)

In figure 3 we can observe the CIF functions for two patients with two values of age: 47 and 68, other covariates were fixed to their median values. Analyzing the shape of CIF functions obtained for disease failure as well as for competing risk we can see the influence of age. The prediction for older people is worse for both types of event. The median value for disease failure is equal to 15.96 for age equals to 47 and 7.41 for older people ($age = 68$). Similar results are obtained by Pintilie [9], where the differences between CIF functions received for two groups of people ($age <= 65$ and $age > 65$) were statistically significant.

5 Conclusions

In the paper the ensemble of dipolar trees for analysis of competing risks is proposed. The method is an extension of the approach proposed in [8], where only one type of event was analyzed. In the described algorithm the information about the type of event is used during dipoles formation. Dipoles may be created only between these covariate vectors which represent the same type of event. The method produces aggregated cumulative incidence function (CIF) for a new patient described by $\mathbf{x}$. The unknown time of occurrence of a given event type may be estimated by median value of the received function.

The results are conducted on the base of the set of patients with follicular type lymphoma, where two types of event were investigated: failure from the disease and death without failure. The influence of clinical stage and age were analyzed. The graphical representation of received CIF functions shows worse prediction of the first type of event for patients with clinical stage II. Age influences the CIF function for both types of event.

Acknowledgements. This work was supported by the grant S/WI/2/08 from Bialystok University of Technology.

References

1. 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)
2. Breiman, L.: How to use survival forest, `http://stat-www.berkeley.edu/users/breiman`
3. Callaghan, F.M.: Classification trees for survival data with competing risks, Univeristy of Pittsburgh, PhD. thesis (2008)
4. Hothorn, T., Lausen, B., Benner, A., Radespiel-Troger, M.: Bagging survival trees. Statistics in Medicine 23, 77–91 (2004)
5. 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), `http://www.bepress.com/ucbbiostat/paper174`
6. Ibrahim, N.A., Kudus, A.: Decision tree for prognostic classification of multivariate survival data and competing risks. In: Strangio, M.A. (ed.) Recent Advances in Technologies (2009)
7. 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)
8. Krętowska, M.: Ensembles of dipolar trees for prediction of survival time. Biocybernetics and Biomedical Engineering 27(3), 67–75 (2007)
9. Pintilie, M.: Competing Risks: A Practical Perspective. John Willey & Sons (2006)
10. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: Competing risks and multi-stage models. Statistics in Medicine 26, 2389–2430 (2007)
11. Ridgeway, G.: The state of boosting. Computing Science and Statistics 31, 1722–1731 (1999)