Multitask Boosting for Survival Analysis with Competing Risks

Alexis Bellot and Mihaela van der Schaar

The Alan Turing Institute

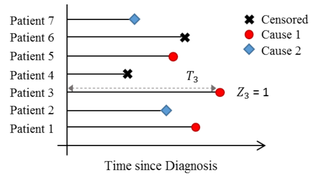
University of Oxford

Motivation

• **Biology is inter-connected:** The progressions of many diseases are related - the prediction of events of interest will be influenced by their *simultaneous* risks of developing related diseases.

▷ Leads to misdiagnoses in patients with atypical disease presentation or risk factors = **major source of patient harm**.

• **Competing Risks:** hinder occurrence of rare events and even among those observed, the large variability among patients difficults inference.



Learning from Competing Risks data

The aim is a flexible simultaneous description of the likelihood of different events over time

• **Formally:** Predict event probabilities Z over time T for patients with features X at risk of mutually exclusive events - *competing risks*. Of interest is the cumulative incidence function (CIF):

$$F_k(t|X) = p(T \leq t, Z = k|X) \tag{1}$$

▷ **Imbalance** among the multiple possible outcomes.
▷ **Complex and Heterogeneous** patient behaviour.

Impact

• **Technical Significance** First nonparametric extension of boosting architectures to survival analysis with competing risks.

▷ New notion of prediction "correctness" which successfully improves predictions of mis-estimated patients.
▷ Multivariate weak learners that encode the shared relationship in related diseases.

• **Medical Relevance** Contribution towards **"precision medicine"**.

▷ Individualized predictions from our model can better guide treatment policies and assess risk, even for atypical patients.

Try our App: mlhcprojects.shinyapps.io/survival_boosting_app

Our Approach: Multitask Boosting

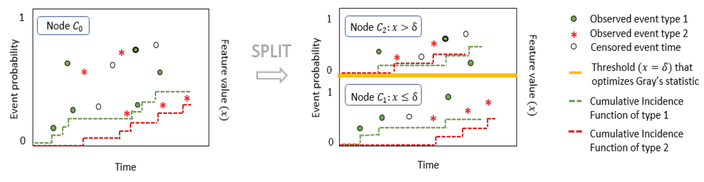
Imbalance - Find a shared representation for the subject's survival with respect to multiple correlated co-morbidities.
Heterogeneity - Develop an ensemble scheme with a special focus on atypical patients.

Multivariate weak learners

Recursively partition the data using similarity measures that involve *all* related tasks. Trees, composed of leaves and nodes.

▷ **Nodes** partition the sample space into more homogeneous subgroups using Gray's test for equality of CIFs in two groups, $H_0 : F_k^1(t) = F_k^2(t)$. Choose split that maximizes,

$$\sum_j G_j, \quad \text{where} \quad G_j := \int_0^\tau K(t)(\hat{\lambda}_j^1(t) - \hat{\lambda}_j^2(t))dt \tag{2}$$



▷ **Leaves** estimate the CIFs in (1) with the Aalen-Johansen estimator:

$$\hat{F}_k(\tau) = \int_0^\tau \hat{S}(t)d\hat{\Lambda}_k(t)$$

where $\hat{S}$ and $\hat{\Lambda}_k$ are estimates of event free survival and of the cumulative hazard respectively. *Note*: $\hat{S}$ includes inference based on all patients in that leave.

Boosting

Iteratively modify the patient distribution such as to focus and improve upon the most heterogeneous patients.

How to find heterogeneous patients? Survival analysis requires a new notion of prediction correctness to detect heterogeneous patients. Propose the following,

$$e_i = \frac{1}{K\tau} \sum_k \int_0^\tau \hat{W}_i(t) \left(I(T_i \leq t, Z_i = k) - \hat{F}(t; \boldsymbol{x}_i) \right)^2 dt \tag{3}$$

where $\hat{W}_i(t)$ are estimated inverse probability of censoring weights.

Algorithm 1 Multitask Boosting

Input: time-to-event data with multiple tasks $\mathcal{D} = \{(X_i, T_i, Z_i)\}_i$ of size n, number of iterations M, initial weights $w_i^{(1)} \propto 1$, sampling fraction s.
for $m = 1$ **to** M **do**
1. Let $\mathcal{D}^*$ be a randomly sampled fraction s of training data $\mathcal{D}$ with distribution $w^{(m)}$.
2. Learn weak model $\boldsymbol{F}^{(m)} : \mathcal{X} \times T \to [0, 1]^K$ on $\mathcal{D}^*$.
3. Prediction error $e_i^{(m)}$ for each instance i with equation (3).
4. Adjusted error of $\boldsymbol{F}^{(m)}$, $\epsilon^{(m)} = \sum_i e_i^{(m)} w_i^{(m)}$.
5. $\beta^{(m)} = \frac{\epsilon^{(m)}}{2/3 - \epsilon^{(m)}}$:
6. Update data distribution $w_i^{(m+1)} \propto w_i^{(m)} (\beta^{(m)})^{1-e_i^{(m)}}$.

end for
Output: Final predictions $\boldsymbol{F}_f$, the weighted average of $\boldsymbol{F}^{(m)}$ for $1 \leq m \leq M$ using $\log(1/\beta^{(m)})$ as the weight of model $\boldsymbol{F}^{(m)}$.

Experiments

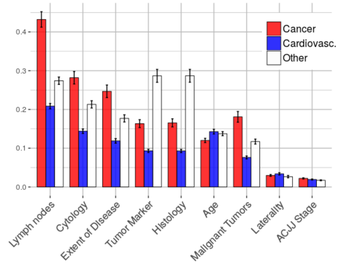
• **SEER dataset:** 72,000 patients diagnosed with breast cancer. 13.64% deaths due to breast cancer, 4.62% due to Cardiovascular diseases (CVD), and 7.3% due to other causes.

Models	Breast Cancer	CVD	Other
Cox	0.773 ± 0.02	0.639 ± 0.03	$\mathbf{0.688 \pm 0.02}$
Fine-Gray	0.777 ± 0.02	0.636 ± 0.03	0.682 ± 0.02
RSF	0.789 ± 0.03	0.722 ± 0.03	0.643 ± 0.02
DeepHit	0.800 ± 0.01	0.662 ± 0.01	0.684 ± 0.01
DMGP	0.801 ± 0.02	0.732 ± 0.03	0.646 ± 0.02
SMTBoost (sep.)	0.795 ± 0.02	0.721 ± 0.04	0.660 ± 0.03
SMTBoost	$\mathbf{0.819 \pm 0.02}$	$\mathbf{0.766 \pm 0.03}$	$\mathbf{0.688 \pm 0.02}$

Table 1: Performance with C-index - close to AUROC - (higher better).

Individual and disease-specific inference

Average predictive impact of observed covariates on survival prediction.



Cause-specific event probabilities over time for a selected patient.