A weighted random survival forest for constructing controllable models

Lev V. Utkin, Anna A. Meldo
Peter the Great St.Petersburg Polytechnic University
Saint-Petersburg clinical scientific and practical center for special types of medical care (oncology-oriented)

12th International Conference on Intelligent Data Processing
2018

Authors are from ...

1. Saint-Petersburg clinical scientific and practical center for special types of medical care (oncology-oriented)
2. Peter the Great St.Petersburg Polytechnic University

ПОЛИТЕХ

NeuroTech.lab

ОНКОЦЕНТР

Authors ...

Polytech Research Laboratory of the Neural Network Technologies and Artificial Intelligence

Lev V. Utkin

Anna A. Meldo

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

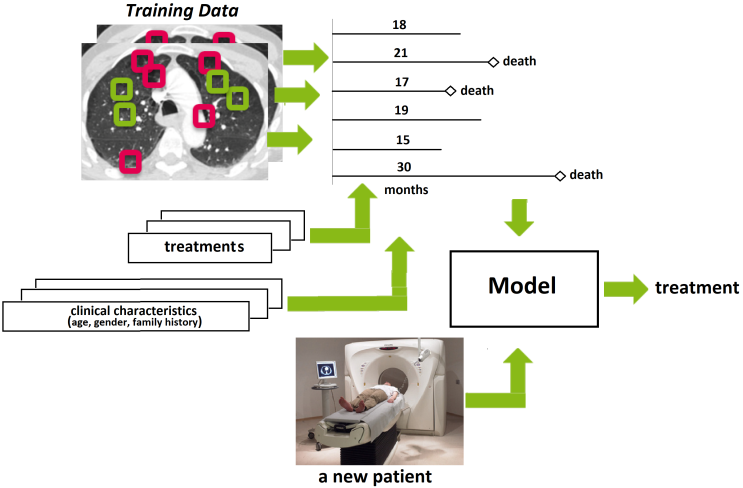
NeuroTech.lab
Лаборатория нейросетевых технологий и искусственного интеллекта СПбПУ

ОНКОЦЕНТР

Two main stages of the patient diagnostics and treatment

1. Cancer detection (computer-aided diagnostic system)
2. **Survival analysis and competing risk analysis (medical treatment recommendation system)**

Survival analysis and competing risk analysis



Formal problem statement of survival analysis

- A patient i is represented by a triplet $(\mathbf{x}_i, \delta_i, T_i)$, $\mathbf{x}_i = (x_{i1}, ..., x_{im})$ are patient characteristics (features); T_i is time to death
- $\delta_i = 1$, if death is observed (uncensored observation)
- $\delta_i = 0$, if death is not observed (censored observation)
- Training set D consists of n triplets $(\mathbf{x}_i, \delta_i, T_i)$, $i = 1, ..., n$.
- **The goal** is to estimate the time to the death T for a new patient with $\mathbf{x}$ by using D

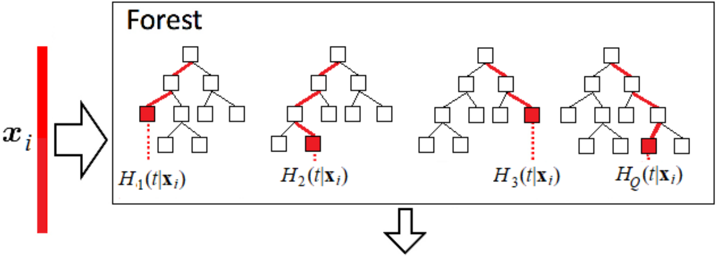
Difficulties of solving the survival analysis problem

1. there are a few training data
2. data may be cencored
3. data may be heterogeneous
4. every patient in the training set is under a single treatment (this is a fundamental problem)

Available survival models (pros and cons)

- The Kaplan-Meier model (requires a homogeneous dataset)
- The Cox proportional hazards model (covariates and time to death are liearly dependent)
- Modifications of the Cox model (Lasso, ridge, elastic net)
- A simple neural network as a basis for a non-linear proportional hazards model
- The SVM approach to survival analysis
- **Survival trees and the survival random forests**
- Deep neural networks (large amount of data)

Random survival forests (RSF)



$$H_f(t|\mathbf{x}_i) = \frac{1}{Q}\sum_{q=1}^{Q} H_q(t|\mathbf{x}_i)$$

$$S_q(t|\mathbf{x}_i) = \exp(-H_q(t|\mathbf{x}_i))$$

ПОЛИТЕХ Санкт-Петербургский политехнический университет Петра Великого
NeuroTech.lab
ОНКОЦЕНТР

Cumulative hazard function (CHF)

- Let $\{t_{j,k}\}$ be the $N(k)$ distinct death times in terminal node k of the q-th tree such that $t_{1,k} < t_{2,k} < ... < t_{N(k),k}$
- Let $Z_{j,k}$ and $Y_{j,k}$ equal the number of deaths and patients at risk at time $t_{j,k}$.
- The CHF estimate for node k is defined as (the Nelson–Aalen estimator):
$$H_k(t) = \sum_{t_{j,k} \leq t} Z_{j,k} / Y_{j,k}$$

Measure of the model quality

- Harrell's C-index or the concordance measure: agreement between the predicted and the observed survival.
- Two subjects chosen at random, the one that fails first has a worst predicted outcome.
- Estimates how good the model is at ranking survival times
- C-index is calculated as

$$C = \frac{1}{M} \sum_{i:\delta_i=1} \sum_{j:t_i<t_j} \mathbf{1}\left[S(t_i^*|\mathbf{x}_i) > S(t_j^*|\mathbf{x}_j)\right].$$

- M is the number of all admissible pairs

Pros and cons of random survival forests

1. **Pros**: They
   - belong to ensemble models with all their advantages
   - have a small number parameters
   - outperform other models by a small amount of training data
   - simple from the training and testing implementations
   - allow solving the feature selection problem
2. **Cons**: They
   - cannot compete with the deep neural networks when a dataset is large
   - some complex non-linear dependencies of features cannot be modelled

Weighted RSF

$$H_f(t, \mathbf{w}|\mathbf{x}_i) = \sum_{q=1}^{Q} H_q(t|\mathbf{x}_i)$$

$$\Downarrow$$

$$H_f(t, \mathbf{w}|\mathbf{x}_i) = \sum_{q=1}^{Q} w_q H_q(t|\mathbf{x}_i), \quad \mathbf{w} \in \Delta_Q$$

- **How to find optimal weights w**
- **What is the optimality of weights?**

Maximization of the C-index

- Optimization problem:

$$\max_{\mathbf{w}\in\Delta_Q} C(\mathbf{w}) = \max_{\mathbf{w}\in\Delta_Q} \sum_{(i,j)\in J} \mathbf{1}\left[\sum_{q=1}^{Q} w_q \left(H_q(t_j^*|\mathbf{x}_j) - H_q(t_i^*|\mathbf{x}_i)\right) > 0\right]$$

- The indicator functions $\mathbf{1}[\cdot]$ are replaced with the hinge loss function $l(x) = \max(0, x)$:

$$\max\left(0, \sum_{q=1}^{Q} w_q \left(H_q(t_i^*|\mathbf{x}_i) - H_q(t_j^*|\mathbf{x}_j)\right)\right)$$

- and the regularization term is added $R(\mathbf{w}) = \|\mathbf{w}\|^2$

Maximization of the C-index (finally)

- The quadratic optimization problem:

$$\min_{\mathbf{w},\xi_{ij}} \left\{ \sum_{(i,j)\in J} \xi_{ij} + \lambda \left\| \mathbf{w} \right\|^2 \right\}$$

subject to $\mathbf{w} \in \Delta_Q$ and

$$\xi_{ij} \geq \sum_{q=1}^{Q} w_q \left(H_q(t_i^* | \mathbf{x}_i) - H_q(t_j^* | \mathbf{x}_j) \right), \quad \xi_{ij} \geq 0, \quad \{i,j\} \in J$$

- ξ_{ij} are the slack variables

Numerical experiments

R package "randomForestSRC"

1. The Primary Biliary Cirrhosis Dataset (418 patients, 17 features): **RSF** 83.61, **WRSF** 83.72
2. Veteran's Administration Lung Cancer Trial Dataset (137 patients, 7 features): **RSF** 70.05, **WRSF** 70.25
3. The Wisconsin Prognostic Breast Cancer Dataset (198 patients, 30 features): **RSF** 76.46, **WRSF** 76.89

Conclusion

1. The proposed WRSF is a way to enhance the survival analysis accuracy as well as to make a more flexible objectives
2. The next improvement of the WRSF is to develop a controllable Deep Survival Forest (Zhou and Feng 2017, a multi-level cascade of random forests, ensemble of ensembles)
3. It can be carried out by introducing training weights of survival decision trees or by combining every random forest with a neural network of a special type.

Questions

?