

Applications of random survival forest based on pseudo-values

Ulla B. Mogensen, Thomas A. Gerds

Department of Biostatistics, University of Copenhagen

Outline

Copenhagen Stroke Study

Random forests

Pseudo-observations

New method

Benchmarking

The Copenhagen stroke study (COST)

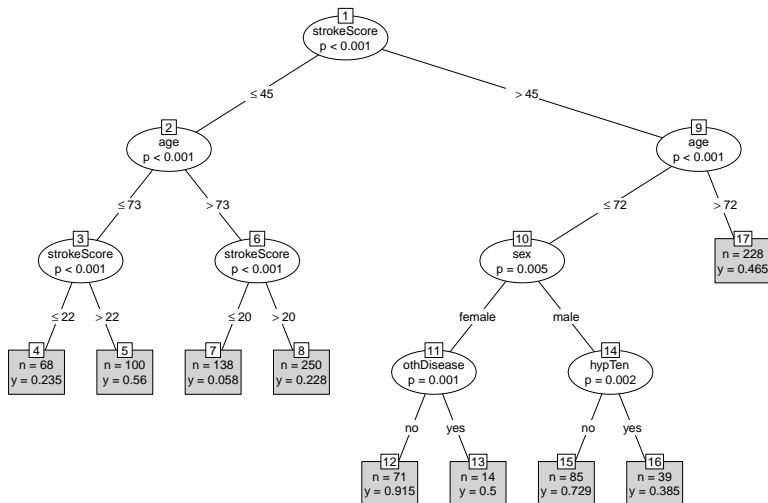
A cohort study of 993 stroke patients being admitted to a hospital in Copenhagen during March 1992 to November 1993 with a 10-year follow-up time.

response variable: survival time after stroke

potential predictors: 13 variables measured at admission

- Scandinavian stroke score (scale from 0-58)
- age, sex
- hyper tension, cholesterol, diabetes, previous stroke, intermitted claudication, myocardial infarction, hemorrhage, other disambing disease, smoke, and alcohol

Decision tree for the survival status after 5 years in the COST data



Leo Breiman's Random Forests Method¹

A random forest is an ensemble of decision trees

- A nonparametric method for prediction
- Predictions for a new patient are obtained by a majority vote of the single trees
- Works for small n large p problems
- Good possibilities to handle missing values
- Shown to perform well in many applications in machine learning

¹Breiman (2001). Random Forest. Machine Learning 45, 5-32

Growing a random forest

- Draw B bootstrap samples with or without replacement
- In each bootstrap sample grow one special decision tree:
 - use only a random subset of the available predictors to find the next split
 - no stopping criterion: grow until all final nodes are pure

- Obtain the prediction for a new patient by voting/averaging
- Assess the prediction error using the patients not used in the current bootstrap sample

Censored data

Observe $(\min(T_i, C_i), \Delta_i)$, where

T_i is the event time for patient i ,

C_i the censoring time

Δ_i indicates if T_i is observed before C_i .

Without censoring, the event time variable holds the same information as the stochastic process $Y_i(t) = I(T_i > t)$.

Pseudo-observations (Andersen & Klein et al.²)

Let $\hat{S}(t)$ be the Kaplan-Meier estimator for the survival function $S(t)$ based on the data of n patients and $\hat{S}^{(-i)}$ the Kaplan-Meier without the i 'th patient. Then define:

$$J_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{(-i)}(t)$$

Without censoring:

$$J_i(t) = Y_i(t)$$

and with censoring

$$J_i(t) \approx Y_i(t)$$

²Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 2003 90(1):15-27

Idea

Grow many random forests based on pseudo-observations as follows

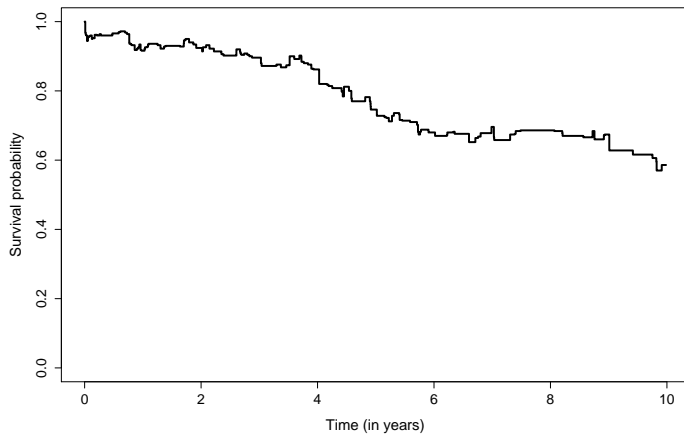
- At each time point t_1, \dots, t_K calculate the pseudo-observation and run the usual (non-survival) randomForest method

For a new patient this yields survival predictions

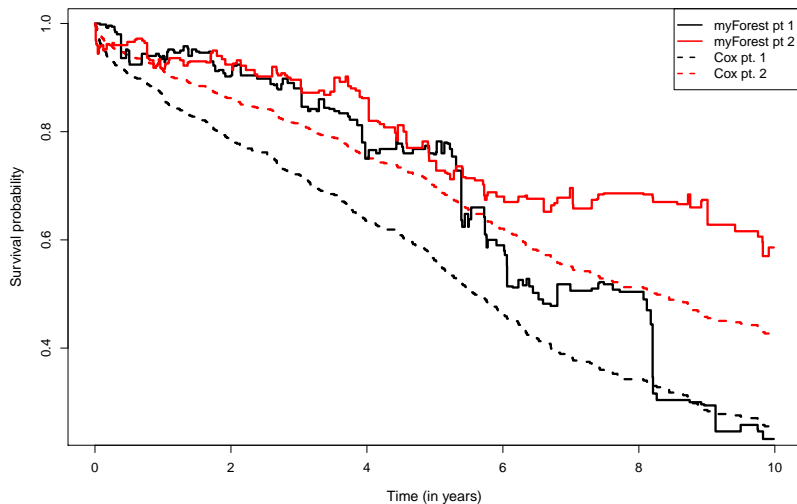
$$P(Y_i(t_k) = 1) = \begin{cases} \text{relative number of trees that predict} \\ \text{status} = 1 \text{ for patient } i \end{cases}$$

for each of the time points t_1, \dots, t_K

A first try: predicted probability for one patient in COST



Comparison of two patients



Rival models

Using the COST data (518 patients) with all available predictors build

1. a Cox model
2. RandomSurvivalForests³ with 500 trees
3. the new method with 500 trees

Compare the models with the time-dependent prediction error (R-package `pec`)

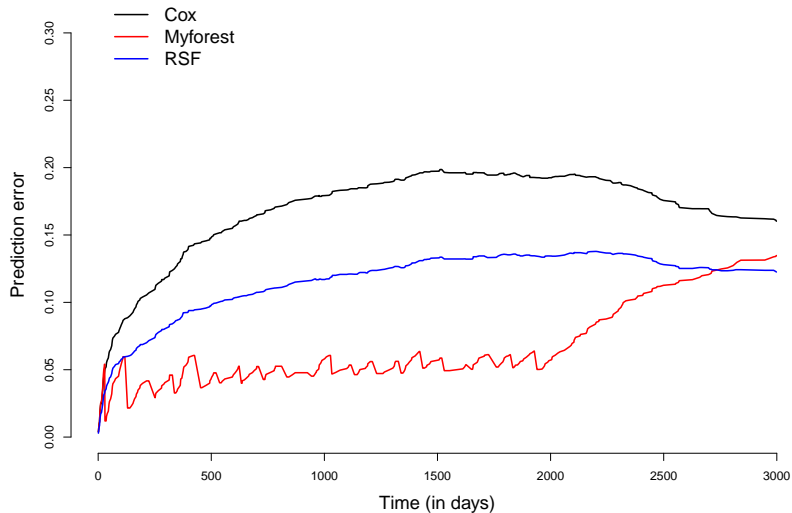
$$t \mapsto \sum_i W_i(t) \{ Y_i(t) - \widehat{pred}(t, X_i) \}^2$$

where $W_i(t)$ is a IPCW weight⁴ $B=100$

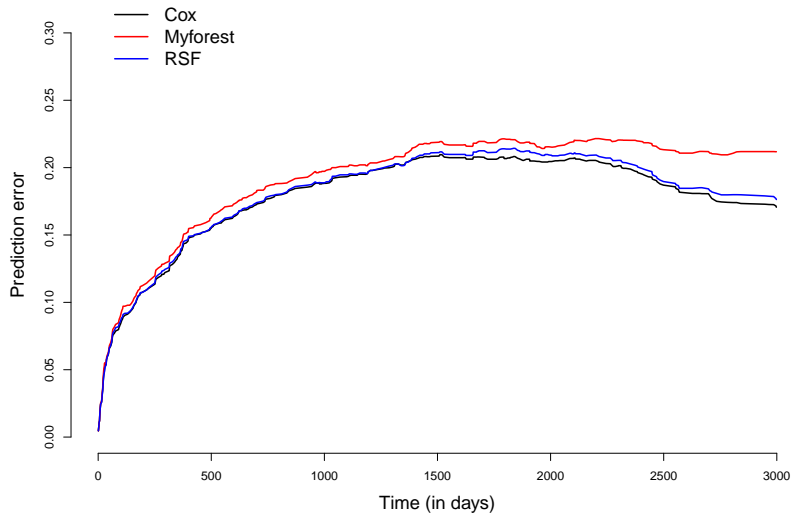
³Ishwaran et al. (2008). Random survival forests. *Ann. Appl. Statist.*, 2, 841-860.

⁴Van der Laan & Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*

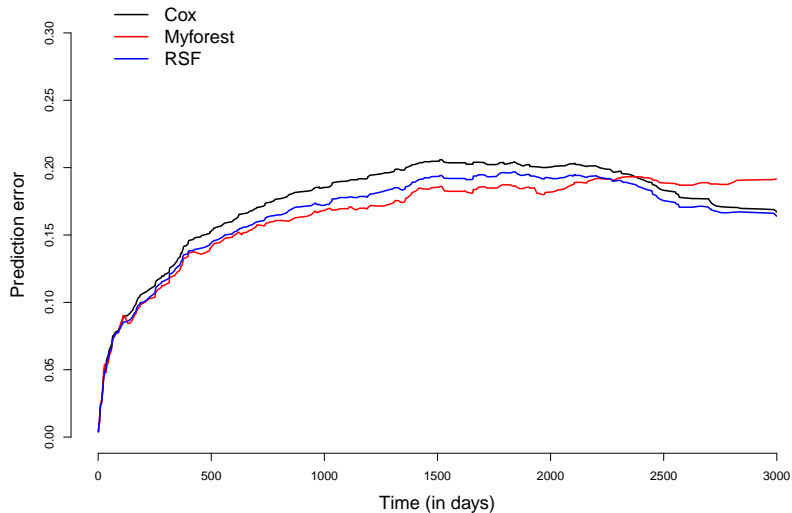
The Apparent prediction error curve



The Out-Of-Bag prediction error curve



The 0.632+ prediction error curve



Summary and Outlook

- Pseudo-observations are a powerful tool to handle censoring in survival analysis
- The new random forest method allows for time-dependent effects
- The new method needs fine tuning:
 - Connect the trees
 - How many trees are needed
 - With or without replacement?
 - Handle missing data
- The .632+ prediction error estimate needs theory. Does it work here?
- Extensions to competing risks and other multistate models