Pseudo random forests
Machine learning methods predict event risk based on high-dimensional data

Thomas A. Gerds & Ulla B. Mogensen

24 November 2012 (a Saturday!)
If you urgently need information . . .

For example, to answer the following multiple choice question:
Q: What is bagging?

1. A machine learning ensemble meta-algorithm
2. Searching in a bag
3. A special case of model averaging
4. The last name of Leo Breiman’s first Ph.D student
5. A short name for bootstrap aggregating

then . . .
...there are several strategies
Bagging

The results of the *k-nearest neighbor* method can be improved by combining the results of many neighbors, think of *asking the audience* from the well-known tv-show.

More generally, a *weak learner* can be improved by *bagging*\(^1\)

Random forests combine many decision trees (based on bootstrap) and thereby improves the predictions of a single tree.\(^2\)

Outline

- A prediction problem in survival analysis with competing risks
- Leukemia study with high-dimensional predictor space
- Jackknife pseudo-values for right censored data
- Applying random forests to pseudo-values
A prediction problem

Outcome:

\[ Y = \begin{cases} 
1 & \text{positive / disease} \\
0 & \text{negative / non-disease} 
\end{cases} \]

Predictors:

\[ X = (X^1, X^2, \ldots, X^p) \]

Parameter:

\[ P(Y = 1|X = x) \]

Data set:

\[ D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)) \quad \text{iid} \]
A survival prediction problem

Time-to-event outcome:

\[ N(t) = I\{T > t\} = \begin{cases} 1 & \text{event} \\ 0 & \text{no event} \end{cases} \]

Predictors:

\[ X = (X^1, X^2, \ldots, X^p) \]

Parameter:

\[ F(t|x) = P(N(t) = 1|X = x) \]

Data set:

\[ D_n = ((X_1, T_1 \wedge C_1, \Delta_1), \ldots, (X_n, T_n \wedge C_n, \Delta_n)) \quad \text{iid} \]
Competing risks

There may be mutually exclusive events, so that only the first is observed.

- Cancer therapy
- Relapse-related death
- Death in remission
Cumulative incidence

The aim is to predict the absolute risk of an event of type 1, say, t-years after the time origin (or landmark):

\[ F_1(t|X) = \text{cumulative incidence function} \]
Cumulative incidence

The aim is to predict the absolute risk of an event of type 1, say, \( t \)-years after the time origin (or landmark):

\[
F_1(t|X) = \text{cumulative incidence function}
\]

Common regression models for \( F_1 \) are based on:

\[
F_1(t|X) = \int_0^t \exp \left\{ - \Lambda_1(s|X) - \Lambda_2(s|X) \right\} \Lambda_1(ds|X)
\]

and alternatively on

\[
F_1(t|X) = \int_0^t P(T \land C > s|X) \frac{\Lambda_1(ds|X)}{G(s - |X)}
\]
Low dimensional predictor space

Formula 1: Combination of Cox regression models, one for each cause:

\[ \Lambda_1(t|X) = \Lambda_{01}(t) \exp(\beta_1^T X) \]
\[ \Lambda_2(t|X) = \Lambda_{02}(t) \exp(\beta_2^T X) \]
Low dimensional predictor space

Formula 1: Combination of Cox regression models, one for each cause:

\[ \Lambda_1(t|X) = \Lambda_{01}(t) \exp(\beta_1^T X) \]
\[ \Lambda_2(t|X) = \Lambda_{02}(t) \exp(\beta_2^T X) \]

Formula 2: Direct transformation models:

\[ h\{F_1(t|X)\} = F_{01}(t) + \eta_1^T X \]

- \( h(x) = \log(-\log(x)) \) (Fine-Gray model)
- \( h(x) = \log(x/(1-x)) \) (Logistic model)
- \( h(x) = \log(x) \) (Log-binomial model)

Requires

\[ G(t|X) = \text{regression model for the censoring times} \]
Bone marrow transplant data

We have 93 leukemia patients that were treated by bone marrow transplantation.

The donors are either related or unrelated, but always matched, meaning that HLA types are of the patient and the donor are the same.

For each patient-donor pair biopsies were taken and common SNPs genotyped by genom-wide microarray analysis (p≈ 1,000,000)
Motivation of the study

Due to the bone marrow transplantation the patient gets part of the immune system from the donor:

- graft vs tumor effect (reduced risk of relapse)
- graft vs host effect (acute and chronic disease)

Theory: Graft vs host disease occurs if the immune system has to deal with genetic information, which it has not seen before (dimension reduction!).

Aim: To identify nsSNP’s and epitopes (minors) associated with favourable and unwanted reactions. Findings could motivate new genetic treatment.
Information from the transplant date

```r
1 t1 <- univariateTable(~sex.patient+isRelated+sex.donor+
   ageTrans+factor(Risk.score), data=gvhd, summary.format="median(x) [range(x)]")
2 publish(summary(t1), org=TRUE)
```

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>Count/Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex.patient</td>
<td>female</td>
<td>36 (38.7)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>57 (61.3)</td>
</tr>
<tr>
<td>isRelated</td>
<td>0</td>
<td>53 (57.0)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>40 (43.0)</td>
</tr>
<tr>
<td>sex.donor</td>
<td>female</td>
<td>36 (38.7)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>57 (61.3)</td>
</tr>
<tr>
<td>ageTrans</td>
<td>median/range</td>
<td>56.00 [31 , 70]</td>
</tr>
<tr>
<td>Risk.score</td>
<td>1</td>
<td>27 (29.0)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>56 (60.2)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10 (10.8)</td>
</tr>
</tbody>
</table>

The risk score combines diagnosis and disease stage and is based on 834 previous patients (Kahl et al. (2007) Blood, Vol: 110, 7).
Overall survival

![Overall survival graph](image)

<table>
<thead>
<tr>
<th>Risk.score=1</th>
<th>27</th>
<th>25</th>
<th>23</th>
<th>22</th>
<th>18</th>
<th>16</th>
<th>12</th>
<th>9</th>
<th>7</th>
<th>5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk.score=2</td>
<td>56</td>
<td>52</td>
<td>46</td>
<td>36</td>
<td>32</td>
<td>28</td>
<td>23</td>
<td>16</td>
<td>12</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Risk.score=3</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Predicting relapse

Years since transplant

Cumulative incidence

Risk.score

0 % 25 % 50 % 75 % 100 %

27 25 23 22 18 16 12 9 7 5 3

[Risk.score=1]

56 48 43 33 28 22 18 15 12 9 6

[Risk.score=2]

10 8 6 6 6 5 5 3 3 3 3

[Risk.score=3]
Predicting treatment related morality

![Graph showing cumulative incidence over years since transplant for different risk scores.]

- Risk score 1: Cumulative incidence for years 0 to 5.
- Risk score 2: Cumulative incidence for years 0 to 5.
- Risk score 3: Cumulative incidence for years 0 to 5.

Tabulated data:

<table>
<thead>
<tr>
<th>Years since transplant</th>
<th>Risk.score=1</th>
<th>Risk.score=2</th>
<th>Risk.score=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
<td>56</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>
Genetic information

The SNP-Chip covers 1,000,000 SNPs. The per patient-donor-pair difference is approximately 300,000 SNPs.

We are only interested in non-synonymous coding SNPs (nsSNP):

▶ mean nsSNP differences in related patient-donor pairs: 3833
▶ mean nsSNP differences in unrelated patient-donor pairs: 6682
▶ mean nsSNP differences in related patient-donor pairs; only in GvH direction: 2060
▶ mean nsSNP differences in unrelated patient-donor pairs; only in GvH direction: 3801
Translation to gene level (further dimension reduction)

A minor (or epitope) is a short peptide sequence that is visible on the surface of the cell. We do have tools to predict which nsSNPs will result in possible Minors.

The predictions of epitopes were based on validated neural networks algorithms.

The number of epitopes as well as the number of nsSNPs per gene were used to predict events. There were 8662 genes for which at least one patient had at least one epitope and 9177 genes in which at least one patient had at least one nsSNP.
Crude summary

A crude summary of the genetic information (where the patient and the donor are different) is obtained by computing for each patient the average number of epitopes (epirate), respectively snprate, across all genes.
Cox regression

- Effects on relapse

<table>
<thead>
<tr>
<th></th>
<th>Hazard.ratio</th>
<th>Cl.95</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>riskscore</td>
<td>0.86</td>
<td>[0.59;1.24]</td>
<td>0.4084</td>
</tr>
<tr>
<td>scaled.epirate</td>
<td>1.58</td>
<td>[1.23;2.03]</td>
<td>0.0004</td>
</tr>
<tr>
<td>ageTrans</td>
<td>1.02</td>
<td>[0.99;1.05]</td>
<td>0.3039</td>
</tr>
</tbody>
</table>

- Effects on treatment related mortality

<table>
<thead>
<tr>
<th></th>
<th>Hazard.ratio</th>
<th>Cl.95</th>
<th>P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>riskscore</td>
<td>1.02</td>
<td>[0.74;1.43]</td>
<td>0.8898</td>
</tr>
<tr>
<td>scaled.epirate</td>
<td>1.33</td>
<td>[1.07;1.66]</td>
<td>0.0111</td>
</tr>
<tr>
<td>ageTrans</td>
<td>1.02</td>
<td>[1.00;1.05]</td>
<td>0.0966</td>
</tr>
</tbody>
</table>
Dilemma

The crude summary of nsSNP/Epitope count across genes cannot rank genes according to their importance for the prediction, and cannot recognize gene-gene interactions.

In high dimensional settings where \((p\gg n)\) multivariate regression based on maximum likelihood does not work (not without penalty).

But there are other tools . . .
If the number of predictors $p$ exceeds the sample size $n$ then, there are these black boxes

$X_i \rightarrow \text{Neural Nets}$
$\text{Support Vector Machines}$
$\text{Bump hunting}$
$Lars and his three cousins$
$\text{Cart and RandomForests}$
$\text{Bayesian networks}$

$\rightarrow \hat{F}_1(t|X_i)$
If the number of predictors $p$ exceeds the sample size $n$

then, there are these almost black boxes

- Neural Nets
- Support Vector Machines
- Bump hunting
- Lars and his three cousins
- Cart and Random Forests
- Bayesian networks

$X_i \rightarrow \hat{F}_1(t \mid X_i)$
Random forest

A random forest is a non-parametric prediction model that combines many decision trees:

- Each tree is grown on a bootstrap sample
- A random subset of predictors are candidates to find the best split in each step of the tree-growing process.
- Modest stopping criteria and no pruning: Trees are (almost) fully grown.
- Predictions for a new subject are majority votes (averages) of the single trees.

Tuning random forests:

- ntree: The number of trees
- mtry: The number of variables tried at each split
Random forest prediction

For a new subject with predictor value $x$ denote $T_b(t; x)$ for the terminal node of the $b$th regression tree in which $x$ ends up. The predicted risk from the $b$th tree at $x$ is given by the average of the pseudo-values in the terminal node:

$$\pi_{b,1}(t, x) = \frac{1}{n} \sum_{i=1}^{n} c_{ib} J_{i1}(t) I(X_i \in T_b(t; x)).$$

The risk predicted by the forest is then given by the average of the tree predictions:

$$\bar{\pi}_1(t, x) = \min(1, \max(0, \frac{1}{B} \sum_{b=1}^{B} \pi_{b,1}(t, x))).$$

The cutting at 0 and 1 is necessary since jackknife pseudo-values can take on values below zero and above one.
Right censored data

In practice the event status can be unknown due to end of follow-up (right censored).
Pseudo values

To deal with right censored data jackknife-pseudo values\(^3\) defined for the jth patient as

\[
J_{i_1}(t) = n \hat{F}_1(t) - (n - 1) \hat{F}_1^{(i)}(t)
\]

For a pre-selected prediction horizon, construct a pseudo-value for the possibly censored event status of subject i:

\[
n\hat{F}_k(t) - (n - 1)\hat{F}_k^{(i)}(t)
\]

where \(\hat{F}_k\) is the Aalen-Johansen estimator in the training set and \(\hat{F}_k^{(i)}(t)\) the same but the data of patient i removed.

\(^3\)See PK Andersen & J Klein (2003-2012)
Example: Predicting the relapse status after 3 years

```r
library(prodlim)
f <- prodlim(Hist(timeRelapse,eventRelapse,cens.code="censored")~1,data=gvhd[order(gvhd$ID),])
pv.3 <- jackknife(f,time=365.25*3)
round(cbind(head(pv.3),tail(pv.3)),4)
```

```
t.1095.75 t.1095.75
[88,]   -0.0212   0.0753
[89,]    0.0000   0.0753
[90,]   -0.0212   0.0729
[91,]    1.0000   0.0858
[92,]   -0.0212   1.0000
[93,]   -0.0212   0.1073
```
Properties of pseudo values

If the censoring distribution is marginally independent of the event time outcome and the predictors, then (with some efforts) one shows

$$E(J_{i1}(t)|X_i) = E(N_{i1}(t)|X_i) = F_1(t|X_i).$$

In particular the pseudo values have the correct mean. For example, to estimate the cumulative incidence of relapse after 3 years:

- Crude incidence: 18.28
- Naive Kaplan-Meier: 21.465
- Aalen-Johansen: 19.404
- Average pseudo-value: 19.404
Dependent censoring

If the censoring distribution is only conditionally independent of the event time outcome given the predictors, then the pseudovalues can be based on a regression model $\hat{G}$ for the conditional censoring distribution:

$$\tilde{F}_1(t) = \int_0^t \sum_i \frac{N_{1i}(ds)}{\hat{G}(s - |X_i|)}$$

and

$$\tilde{J}_{i1}(t) = n\tilde{F}_1(t) - (n - 1)\tilde{F}_1^{(i)}(t)$$
Regression tree I: relapse risk after 3 years
Regression tree II: relapse risk after 3 years
Random forest (ntree=1000): relapse risk after 3 years
Regression tree 1: treatment related mortality after 3 years
Regression tree II: treatment related mortality after 3 years
Random forest (ntree=1000): treatment related mortality after 3 years
Conclusions

▶ The jackknife pseudo-value approach allows to apply machine learning methods to complex event history data
▶ The risks of all events need be considered
▶ Random forests allows for interactions and provides accurate predictions (not shown) and variables importance rankings (not shown)
▶ Random forests are robust against high dimensional predictors spaces
▶ The propose pseudo random forest can be combined with any software implementation of random forests. This is of particular interest when computation time matters