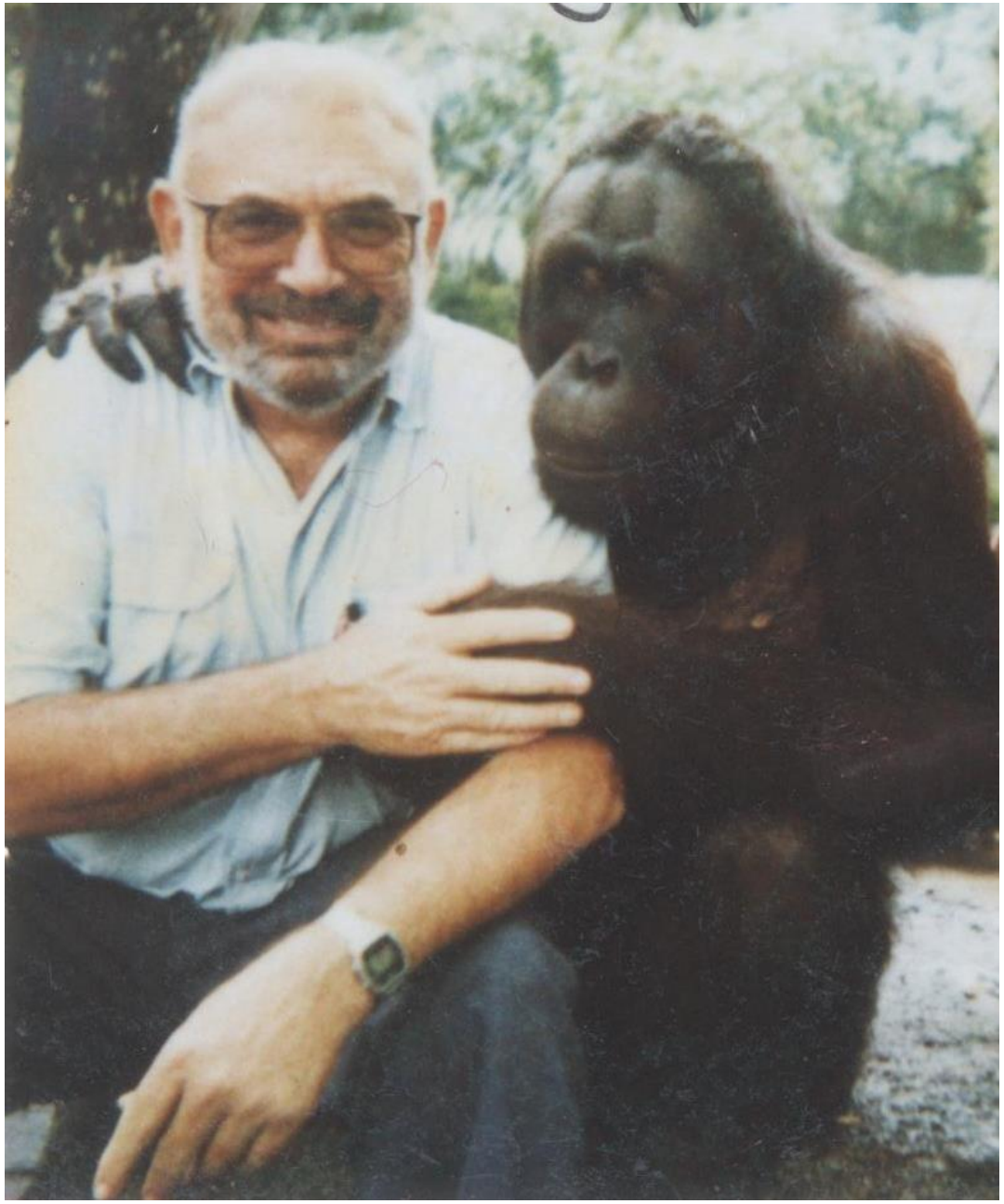
Random Forests

Adele Cutler
Utah State University



Advances in Random Forests

"A Random Forest Guided Tour" (2015)

Gérard Biau, Erwan Scornet

Lots of practical evidence that random forests perform well…

Comparisons

"Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" Jpurnal of *Machine Learning Research, 2014*

Fernandez-Delgado, Cernadas, Barro, Amorim

Compared 179 classifiers on 121 datasets,

"The classifiers most likely to be the bests are the random forest (RF) versions" (SVM with Gaussian kernel second best)

Hal Varian (Google)

"Big Data: New Tricks for Econometrics" (2014)

Journal of Economic Perspectives

Cites a conference presentation by Jeremy Howard and Mike Bowles (2012), who claim *"ensembles of decision trees (often known as 'Random Forests') have been the most successful general-purpose algorithm in modern times."*

Advances in Random Forests

"A Random Forest Guided Tour" (2015)

Gérard Biau, Erwan Scornet

Lots of practical evidence that random forests perform well… much less theory

Theory

Random Forests
Machine Learning (2001)
Leo Breiman

Upper bound on generalization error in terms of strength and correlation of individual trees.

Earyly results (Breiman 2001)

Idea: most of the trees are good for most of the data and make mistakes in different places

More formally (Breiman, 2001) the trees have

- high strength
- low correlation

Theory

Analyzing Bagging
Annals of Statistics (2002)
Peter Bühlmann and Bin Yu

Theoretical results on variance reduction, also subsampling instead of bootstrap samling

Theory

Random Forests and Adaptive Nearest Neighbors
JASA (2006)
Lin and Jeon

Show that random forests are like nearest neighbor classifiers with clever metric

Theory

Asymptotic distribution of random forests:

Biau and Devroye (2010)
Denil et al. (2003)
Meinshausen (2006)
Biau et al. (2008)
Biau (2012)
Denil (2013)
Mentch and Hooker (2014)
Wager (2014)

Theoretical Issues

Hard to prove things without simplifying the forest. E.g. Make the splits independent of the data.

Biased Variable Importance

Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a solution

BMC Bioinformatics 2007Carolin Strobl et al.

Bias in variable importance if categorical predictors have different numbers of levels and/or predictors are mixed categorical and continuous (Strobl et al. 2007, Boulesteix 2012)

Joint work with:

Rong Xia
Ph.D., University of Michigan

Biased Variable Importance

Strobl et al. (2007) report that when predictors have unequal scales it

> *"severely affects the reliability and interpretability of the variable importance measure"*

Simulations

- 1000 trees in each forest
- 100 observations in the training set
- 1000 observations in an independent test set
- 100 repetitions
- replace = FALSE for cforest (subsampling)
- default parameters unless otherwise noted

Examples 1 and 2

x1 ~ M(2)

x2 ~ M(2)

x3 ~ M(4)

x4 ~ M(10)

x5 ~ U(0, 1)

x6 ~ N(0, 1)

M is multinomial

Example 1: (main effect)

y = Bernoulli(p)

p = .3 if x1 = 0

p = .7 if x1 = 1

Example 2: (interaction)

y = 1 if x1 = x2

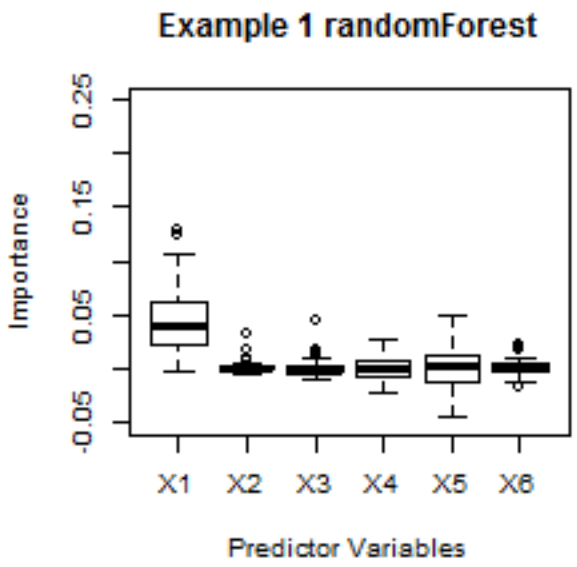
y = 0 otherwise

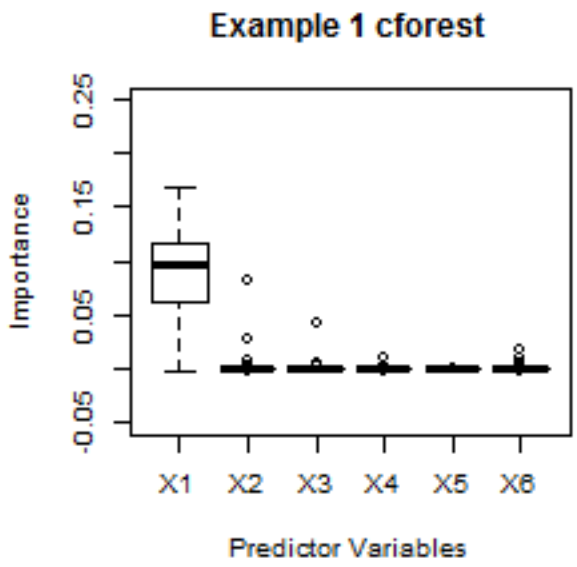
% Error rates Example 1

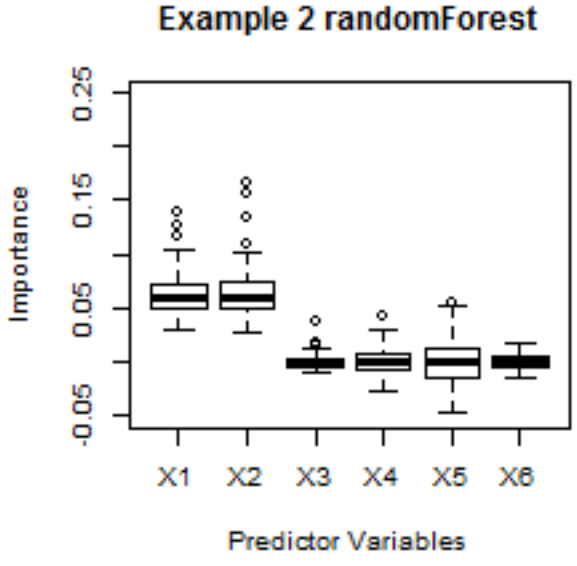
m	random forest	cforest	random forest	cforest
	mean		SE	
1	**39,1**	45,1	0,4	0,8
2	41,2	**39,3**	0,4	0,9
3	41,6	**35,4**	0,4	0,9
4	41,8	**32,9**	0,4	0,7
5	42,1	**31,7**	0,4	0,5

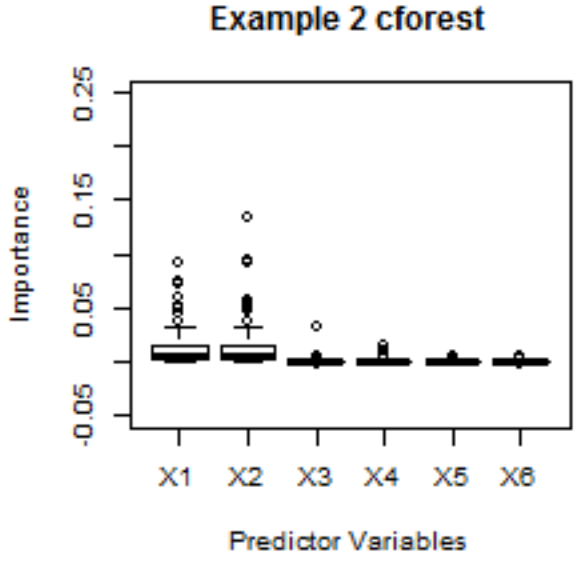
% Error rates Example 2

m	random forest	cforest	random forest	cforest
	mean		SE	
1	**17,7**	49,8	0,5	0,2
2	**24,8**	48,7	0,5	0,5
3	**36,1**	47,1	0,6	0,9
4	**40,3**	44,8	0,5	1,2
5	42,2	41,6	0,5	1,6


Example 1 randomForest
Importance
0.25
0.15
0.05
-0.05
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 1 cforest
Importance
0.25
0.15
0.05
-0.05
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 2 randomForest
Importance
0.25
0.15
0.05
-0.05
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 2 cforest
Importance
0.25
0.15
0.05
-0.05
X1
X2
X3
X4
X5
X6
Predictor Variables

Examples 3 and 4

$x1, x2, \ldots, x6 \sim N(0, 1)$

Example 3: (main effect)

$y = 0$ if $x1 > 0$

$y = 1$ otherwise

Example 4: (interaction)

$y = 0$ if $x1*x2 > 0$

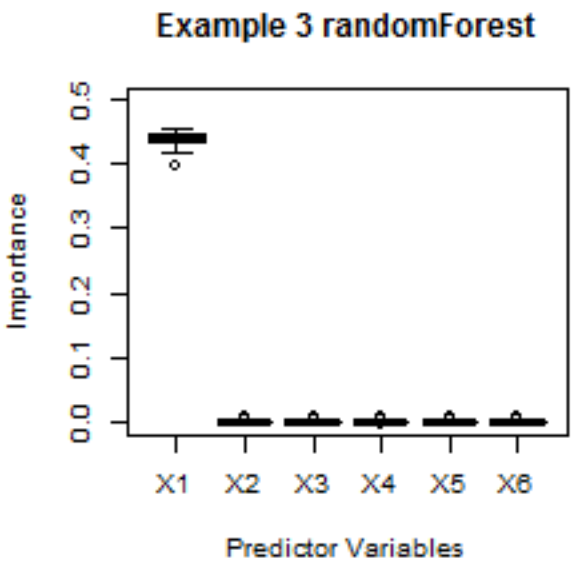
$y = 1$ otherwise

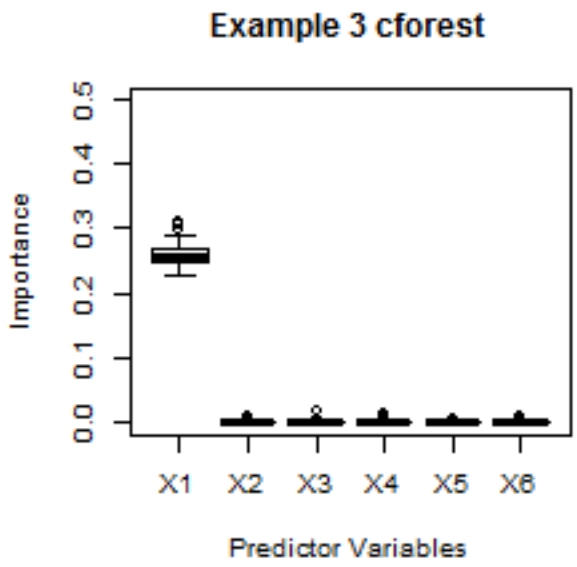
% Error rates example 3

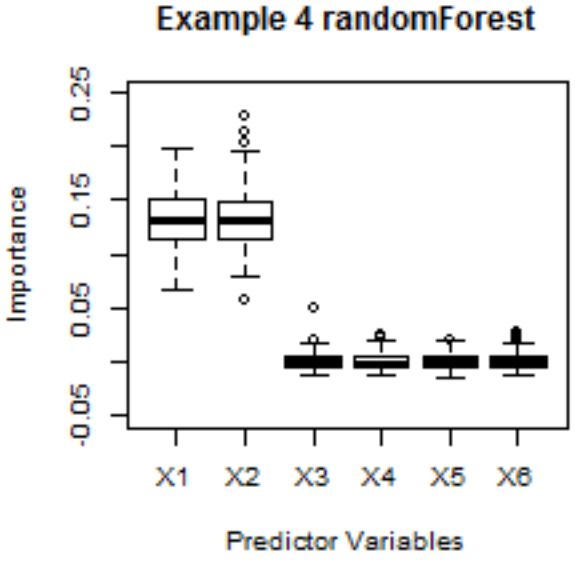
mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	**0,61**	13,86	0,05	2,10
2	**0,45**	2,42	0,04	0,72
3	**0,43**	1,21	0,04	0,13
4	**0,41**	1,10	0,04	0,12
5	**0,41**	1,06	0,04	0,12

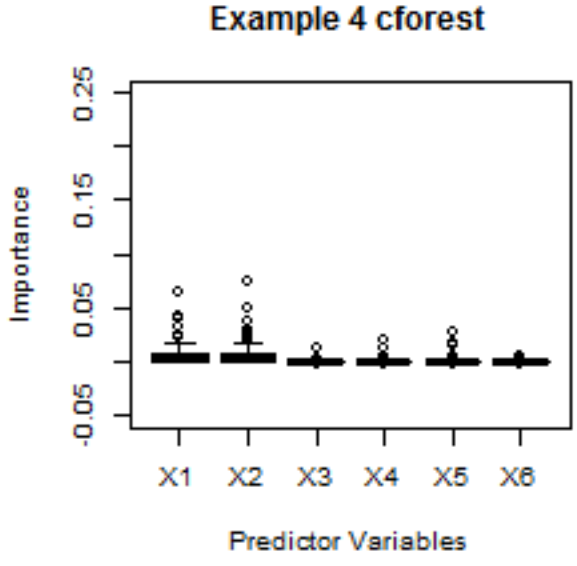
% Error rates example 4

mtry	random forest	cforest	random forest	cforest
	mean		SE	
1	**28,5**	50,0	0,4	0,1
2	**21,8**	49,8	0,5	0,2
3	**17,6**	48,8	0,6	0,4
4	**14,7**	48,0	0,7	0,5
5	**13,1**	47,0	0,7	0,7


Example 3 randomForest
Importance
0.0
0.1
0.2
0.3
0.4
0.5
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 3 cforest
Importance
0.0
0.1
0.2
0.3
0.4
0.5
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 4 randomForest
Importance
-0.05
0.05
0.15
0.25
X1
X2
X3
X4
X5
X6
Predictor Variables


Example 4 cforest
Importance
-0.05
0.05
0.15
0.25
X1
X2
X3
X4
X5
X6
Predictor Variables

Extensions

- Weighted random forests (Winham et al. 2013)
- Online forests (various authors, starting in 2009)
- Survival forests (Ishwaran, 2008)
- Quantile forests (Meinshausen, 2006)
- One class random forests (Desir, 2013)
- Bayesian forests (Chipman et al. 2008)

Extensions

- Weighted random forests (Winham et al. 2013)
- Online forests (various authors, starting in 2009)
- Survival forests (Ishwaran, 2008)
- Quantile forests (Meinshausen, 2006)
- One class random forests (Desir, 2013)
- Bayesian forests (Chipman et al. 2008)

Weighted Random Forests

Interested in genetics of complex disease and found that for very wide problems RF didn't do very well at detecting interactions (very unlikely to split on the interacting variables).

- Incorporate tree-level weights to emphasize more accurate trees.
- Can outperform RF in high-dimensional data.
- The improvements are modest

Weighted Random Forests

1. Split into 75% training set and 25% testing set.
2. Fit usual RF to the training set.
3. Use oob data to get a weight for each tree (lower weights for trees with high oob prediction error). $W_t = 1/\text{rank}(PE_t)$
4. Pass the test set down the forest and use the weights from step 3 to do a weighted aggregation.

Problem

- The OOB error rate is high if the tree is very bad, OR if there are hard points in the OOB data.
- So we are upweighting the trees which have the hard points in the bootstrap sample.
- Why should this be a good idea?

Alternative

1. Split into 75% training set and 25% testing set.
2. Do 10-fold crossvalidation on the training set:
   - Fit a random forest to the 90%
   - Get weights based on the other 10%
   - Keep the best Ntree/10 trees, discard the rest
3. Combine the Ntree trees to give a final forest
4. Predict on the test set.

Results

No better than usual random forests!

Why?

Maybe like boosting: emphasize the trees that focus on the hard parts of the data