Party on! A new, conditional variable importance measure for random forests available in `party`

Carolin Strobl (LMU München) and Achim Zeileis (WU Wien)

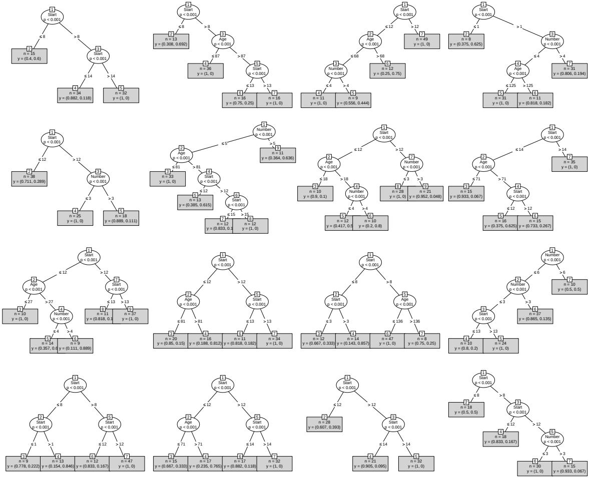
useR! 2009

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences
- can deal with "small n large p"-problems, high-order interactions, correlated predictor variables
- are used not only for prediction, but also to measure variable importance
  (advantage: RF variable importance measures capture the effect of a variable in main effects and interactions
  $\rightarrow$ smarter for screening than univariate measures)

(Small) random forest



Measuring variable importance

Measuring variable importance

- Gini importance

  mean Gini gain produced by X_j over all trees

  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)

Measuring variable importance

- Gini importance

  mean Gini gain produced by X_j over all trees

  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)

- permutation importance

  mean decrease in classification accuracy after permuting X_j over all trees

  (unbiased when subsampling is used; Strobl et al., 2007)

The permutation importance

within each tree t

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

$\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i) \;=$ predicted class before permuting

$\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j}) \;=$ predicted class after permuting X_j

$$\mathbf{x}_{i,\pi_j} = \left(x_{i,1}, \ldots, x_{i,j-1}, \;\; x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p}\right)$$

Note: $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if X_j is not in tree t

The permutation importance

over all trees:

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{x}_j)}{ntree}$$

What null hypothesis does this permutation scheme correspond to?

obs	Y	X_j	Z
1	y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	$x_{\pi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{\pi_j(n),j}$	z_n

$$H_0 : X_j \perp Y, Z \text{ or } X_j \perp Y \wedge X_j \perp Z$$

$$P(Y, X_j, Z) \stackrel{H_0}{=} P(Y, Z) \cdot P(X_j)$$

What null hypothesis does this permutation scheme correspond to?

the current null hypothesis reflects independence of X_j from both Y and the remaining predictor variables Z

$\Rightarrow$ a high variable importance can result from violation of

either one!

Suggestion: Conditional permutation scheme

obs	Y	X_j	Z
1	y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
3	y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
27	y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
6	y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
14	y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
33	y_{33}	$x_{\pi_{j\|Z=b}(33),j}$	$z_{33} = b$
⋮	⋮	⋮	⋮

$$H_0 : X_j \perp Y | Z$$

$$P(Y, X_j | Z) \overset{H_0}{=} P(Y|Z) \cdot P(X_j|Z)$$

$$\text{or} \quad P(Y | X_j, Z) \overset{H_0}{=} P(Y|Z)$$

Technically

- use any partition of the feature space for conditioning

Technically

- use any partition of the feature space for conditioning
- here: use binary partition already learned by tree

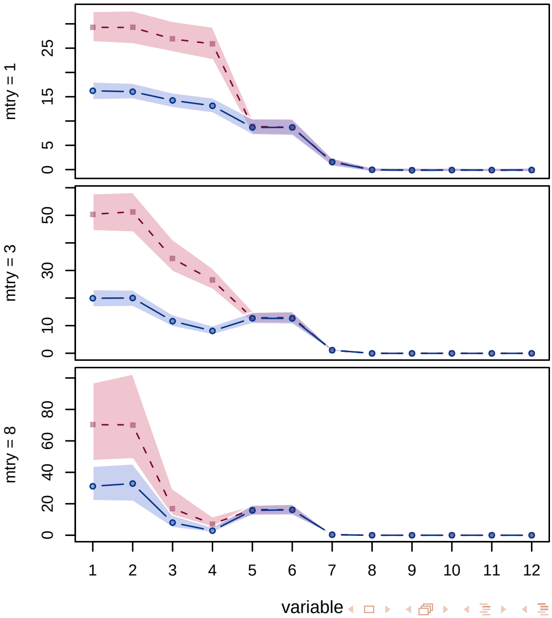
Simulation study

- dgp: $y_i = \beta_1 \cdot x_{i,1} + \cdots + \beta_{12} \cdot x_{i,12} + \varepsilon_i$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 0.5)$
- $X_1, \ldots, X_{12} \sim N(0, \mathbf{\Sigma})$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.9 & 0.9 & 0.9 & 0 & \cdots & 0 \\ 0.9 & 1 & 0.9 & 0.9 & 0 & \cdots & 0 \\ 0.9 & 0.9 & 1 & 0.9 & 0 & \cdots & 0 \\ 0.9 & 0.9 & 0.9 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

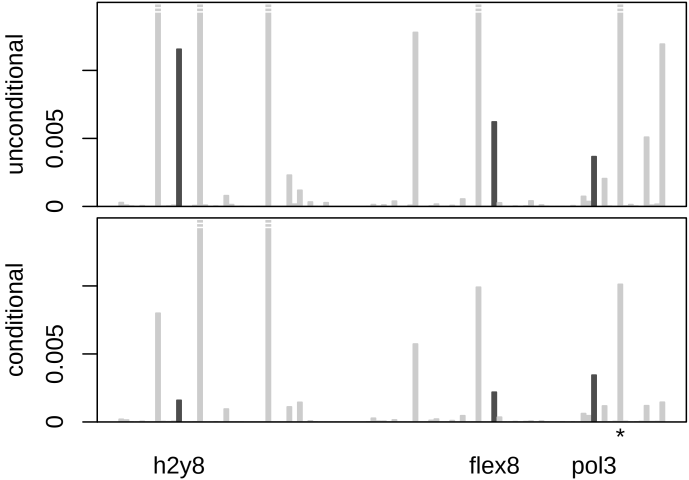
X_j	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	X_5	X_6	X_7	X_8	$\cdots$	X_{12}
β_j	**5**	**5**	**2**	**0**	-5	-5	-2	0	$\cdots$	0

Results



mtry = 1
mtry = 3
mtry = 8
0
5
15
25
0
10
30
50
0
20
40
60
80
1
2
3
4
5
6
7
8
9
10
11
12
variable

Peptide-binding data



R-Example

spurious correlation between shoe size and reading skills in school-children

```
> mycf <- cforest(score ~ ., data = readingSkills,
+                   control = cforest_unbiased(mtry = 2))
> varimp(mycf)
nativeSpeaker           age      shoeSize
     12.62926      74.89542      20.01108
> varimp(mycf, conditional = TRUE)
nativeSpeaker           age      shoeSize
    11.808192     46.995336      2.092454
```

from party 0.9-991

Conclusion

- conditional permutation is expensive
- but gets us closer to the interpretation of importance that we (statisticians) are used to
  $\longrightarrow$ beta coefficients, partial correlations
- choice of `mtry` has a high impact

General remarks

- default settings for `mtry` vary between implementations

  e.g., for classification:

  `randomForest`: `mtry`$=\sqrt{p}$

  `cforest`: `mtry`$=5$

  small values of `mtry` may often be a good choice - but not in the case of correlated predictors!

- make sure your results are stable before interpreting importance rankings

  fit another forest with a different random seed - if the ranking changes increase `ntree`

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9:307*.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25*.