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 → smarter for screening than univariate measures)
(Small) random forest

A new, conditional importance

Conclusion

References
Measuring variable importance

A new, conditional importance

Conclusion

References
Measuring variable importance

- Gini importance
  mean Gini gain produced by $X_j$ over all trees
  (can be severely biased due to estimation bias and multiple 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 multiple 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 $t$

$$VI^{(t)}(x_j) = \frac{\sum_{i \in B^{(t)}} I(y_i = \hat{y}_i^{(t)})}{|B^{(t)}|} - \frac{\sum_{i \in B^{(t)}} I(y_i = \hat{y}_{i,\pi_j}^{(t)})}{|B^{(t)}|}$$

$\hat{y}_i^{(t)} = f(t)(x_i) = \text{predicted class before permuting}$

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

$$x_{i,\pi_j} = (x_{i,1}, \ldots, x_{i,j-1}, x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p})$$

Note: $VI^{(t)}(x_j) = 0$ by definition, if $X_j$ is not in tree $t$
The permutation importance

over all trees:

\[ VI(x_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree} \]
What null hypothesis does this permutation scheme correspond to?

<table>
<thead>
<tr>
<th>obs</th>
<th>Y</th>
<th>$X_j$</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_1$</td>
<td>$x_{\pi_j(1).j}$</td>
<td>$z_1$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$i$</td>
<td>$y_i$</td>
<td>$x_{\pi_j(i).j}$</td>
<td>$z_i$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$n$</td>
<td>$y_n$</td>
<td>$x_{\pi_j(n).j}$</td>
<td>$z_n$</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>obs</th>
<th>Y</th>
<th>$X_j$</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_1$</td>
<td>$x_{\pi_j</td>
<td>Z=a(1),j}$</td>
</tr>
<tr>
<td>3</td>
<td>$y_3$</td>
<td>$x_{\pi_j</td>
<td>Z=a(3),j}$</td>
</tr>
<tr>
<td>27</td>
<td>$y_{27}$</td>
<td>$x_{\pi_j</td>
<td>Z=a(27),j}$</td>
</tr>
<tr>
<td>6</td>
<td>$y_6$</td>
<td>$x_{\pi_j</td>
<td>Z=b(6),j}$</td>
</tr>
<tr>
<td>14</td>
<td>$y_{14}$</td>
<td>$x_{\pi_j</td>
<td>Z=b(14),j}$</td>
</tr>
<tr>
<td>33</td>
<td>$y_{33}$</td>
<td>$x_{\pi_j</td>
<td>Z=b(33),j}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$H_0 : X_j \perp Y | Z$

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

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

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

\[
\Sigma = \begin{pmatrix}
1 & 0.9 & 0.9 & 0.9 & 0 & \cdots & 0 \\
0.9 & 1 & 0.9 & 0.9 & 0 & \cdots & 0 \\
0.9 & 0.9 & 1 & 0.9 & 0 & \cdots & 0 \\
0.9 & 0.9 & 0.9 & 1 & 0 & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>$X_j$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
<th>$\cdots$</th>
<th>$X_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_j$</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
<td>-2</td>
<td>0</td>
<td>$\cdots$</td>
<td>0</td>
</tr>
</tbody>
</table>
Results

![Graph showing variable importance for different mtry values.](image)
Peptide-binding data
spurious correlation between shoe size and reading skills in school-children

```r
> 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
  → beta coefficients, partial correlations
- choice of mtry has a high impact
General remarks

- default settings for \( m_{\text{try}} \) vary between implementations e.g., for classification:
  - randomForest: \( m_{\text{try}} = \sqrt{p} \)
  - cforest: \( m_{\text{try}} = 5 \)

  small values of \( m_{\text{try}} \) 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 \( n_{\text{tree}} \)