Uncovering interactions with Random Forests

Jake Michaelson
Marit Ackermann
Andreas Beyer
Random Forests

>> ensembles of decision trees

>> diverse trees trying to solve the same problem

>> used frequently for:

>> prediction (knowledge of model less important)

>> feature selection (prediction less important)
RF interactions: prior art

>> online official RF manual


>> Bureau, et al. (2005)

>> pairwise permutation importance

>> Mao and Mao (2008)

>> Jiang, et al. (2009)

>> selection with RF Gini importance, conventional (LM-based) interaction test (up to 3-way)
a typical problem
a typical problem
a typical problem
a typical problem
split symmetry

A

B

B

B

B

B

B

B
split asymmetry
testing split symmetry

>> independence of predictors A and B:

>> expect B as left daughter 50% of the time

>> expect B as right daughter 50% of the time

>> the prior (a beta density) is centered around 0.5
testing split symmetry
testing split symmetry

we update the prior density parameters with the observed left/right daughter counts:

\[
\text{a}_{\text{posterior}} = \text{a}_{\text{prior}} + AB_{\text{left}}
\]

\[
\text{b}_{\text{posterior}} = \text{b}_{\text{prior}} + AB_{\text{right}}
\]

... and take the posterior/prior density ratio at 0.5

this is the Bayes factor
testing split symmetry
building a graph

\[ P_{post} = P_h \frac{BF}{(p_h \cdot BF + 1 - p_h)} \]

>> using the Bayes factor from each pair of predictors, we calculate the posterior probability of symmetry

>> i.e. that the true proportion is 0.5

>> we use a high prior probability of the hypothesis (e.g. \( p_h = 0.999999 \))
building a graph

### Posterior Probabilities

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>I</td>
<td>0.001</td>
<td>0.001</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>0.8</td>
<td>I</td>
<td>0.99</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.99</td>
<td>0.3</td>
<td>I</td>
<td>0.003</td>
</tr>
<tr>
<td>D</td>
<td>I</td>
<td>0.89</td>
<td>0.99</td>
<td>I</td>
</tr>
</tbody>
</table>

### Adjacency Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>I</td>
<td>I</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>I</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Graph

- A → B
- B → D
- D → C
- C → B
simulations

>> 1000 binary predictor variables, 200 observations

>> 3 - 4 predictors participate in true model

>> tested ability of the method to recover the true topology of the simulated model

>> recorded TP, FP while varying mtry and ntree
test models

3 independent effects (i.e. no edges)

A
B
C

mtry

ntree

TP

FP

2500
5000
7500
10000

250
500
750
1000

250
500
750
1000

mtry

ntree
test models

3-way unordered interaction

TP

FP
test models

one main effect, one ordered 3-way interaction, one ordered 2-way interaction

TP

FP
test models

two independent, ordered two-way interactions

TP

FP

mtry

ntree

250 500 750 1000

2500 5000 7500 10000

A

C

B

D

two independent, ordered two-way interactions
real world

>> Gabrb3

>> neurotransmitter receptor subunit

>> absence (or misexpression) yields autism-like behavior

>> what mechanisms influence Gabrb3 expression?

Livet, et al. (2007)
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes \((m = 3,794)\) of the same population of mice, then extract the interaction graph
regulation of Gabrb3

grow an RF that **regresses** hippocampal Gabrb3 expression on the **genotypes** \(m=3,794\) of the same population of mice, then **extract the interaction graph**
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes \( m = 3,794 \) of the same population of mice, then extract the interaction graph
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes \((m=3,794)\) of the same population of mice, then extract the interaction graph
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes \( m = 3,794 \) of the same population of mice, then extract the interaction graph
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m=3,794) of the same population of mice, then extract the interaction graph
regulation of Gabrb3

L1 - Gabrb3 (cis effect)
L2 - Dscam (axon guidance)
L3 - Magi2 (synaptic scaffolding)
the context

Synaptic vesicle

Neurotransmitters

Neurotransmitter re-uptake pump

Voltage-gated Ca\(^{++}\) channel

Post-synaptic density

Axon terminal

Synaptic cleft

Dendritic spine
the context

- Synaptic vesicle
- Neurotransmitters
- Neurotransmitter re-uptake pump
- Neurotransmitter receptors
- Voltage-gated Ca^{++} channel
- Post-synaptic density
- Axon terminal
- Synaptic cleft
- Dendritic spine

- Dscam
- Magi2
- Gabrb3
conclusion

>> (a)symmetry of transitions between subsequently selected variables can give us clues about the degree of dependence between them

>> constructing a graph of these dependencies can illustrate the emergent dependency structure of the predictors in light of the response
forthcoming...

>> does this work for continuous and categorical predictors?

>> what about correlated predictors?

>> strategy for choosing optimal mtry and ntree?
RF is an example of a tool that is useful in doing analyses of scientific data.

But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem.

Take the output of random forests not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem.

- Breiman & Cutler
Thanks!

HELMHOLTZ
ASSOCIATION

Alliance on Systems Biology

KTF
THE KLAUS TSCHIRA FOUNDATION gGmbH

jacob.michaelson@biotec.tu-dresden.de