Decision Snippet Features

Pascal Welke, Fouad Alkhoury, Christian Bauckhage, Stefan Wrobel

ICPR 2021
A Small Dataset

<table>
<thead>
<tr>
<th>lockdown</th>
<th>rain</th>
<th>cold</th>
<th>sunny</th>
<th>drink outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>☂</td>
<td>✗</td>
<td>✗</td>
<td>🧴</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>☀</td>
<td>🧴</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>☀</td>
<td>✗</td>
<td>🧴</td>
</tr>
<tr>
<td>✗</td>
<td>☂</td>
<td>☀</td>
<td>✗</td>
<td>😊</td>
</tr>
<tr>
<td>✗</td>
<td>☄</td>
<td>☀</td>
<td>✗</td>
<td>😊</td>
</tr>
<tr>
<td>✗</td>
<td>☄</td>
<td>☂</td>
<td>✗</td>
<td>😊</td>
</tr>
</tbody>
</table>

Basically, I drink outside whenever there is no lockdown and it is not raining. We see only a random training subset, so an algorithm might come to a different conclusion.
Motivation

• Decision trees are great
  – interpretable by humans
  – fast to train and apply
  – tend to overfit

• Ensembles (i.e. Random Forests) reduce variance
  – larger model size
  – less interpretable (due to larger size)

• How can we retain the benefits of random forests and decision trees?
  – the trees in a random forest are not independent
  – arguably, common structures might result from the underlying learning problem

Let’s learn from random forests to identify a relevant smaller trained random forest
Let’s train a random forest with 20 trees on this training data.
Let’s Look at these Trees

- Three trees are found multiple times
- Substructures occur even more frequently

We will use frequent subtrees to build new (smaller) ensemble models.
Technical Issues

• Substructures may be incomplete
  – We need to add leaves

• Substructures see different data
  – We cannot use the leaf labels
Random Forests are Representation Learning + Linear Model

Decision Snippet Features

Training Process

1. Train Random Forest on Data
2. Mine Decision Snippets
3. Transform Data to Decision Snippet Feature space
4. Train a linear classifier
Conclusion

• Decision Snippet Features are based on regularities in random forests
• They work well
  – Size reductions up to orders of magnitude
  – Comparable predictive performance

Check out our paper!

☀️ 🍻