Else-Tree Classifier for Minimizing Misclassification of Biological Data

Truong X. Tran, Computer Science Dept., Marc L. Pusey, iXpressGenes Inc., Ramazan S. Aygun Computer Science Dept.

Overview
In many applications, such as biological research, inaccuracy or misclassification of machine learning algorithms can yield fatal results. A significantly high value of an evaluation measure is an indication of overfitting, and chosen classifiers are likely to have false classifications for new data. Else-Tree is a novel machine learning classifier that reduces the misclassification of data samples by labeling them as undecided rather than assigning them an incorrect class.

Impact
The Else-Tree both avoids critical mistakes and increases the trust of the user of the classifier.

Protein Crystallization Research:
- Thousands of trials may need to be set up for a successful crystalline outcome.
- The 3-D structure of a protein is initially obtained by crystallizing the protein in drug development, missing a crystalline condition may hinder its development.
- By using Else-Tree, the expert may only review undecided items rather than all samples.

Key Findings
The main feature of the else-tree is its potential to generate zero percent error without overfitting by separating hard to classify data as undecided.

Explanation
An Else-Tree classifier contains nodes, branch conditions, and leaf nodes having labels. The novelty of Else-Tree is its else branch and the final else branch leading to else-leaf. Rather than giving the wrong prediction, Else-Tree marks doubtful samples as '?' or undecided. The Else-Tree is built by analyzing pure regions of an attribute per class of the training data. The most populated contiguous regions per class are used to label leaf nodes. The rest of the data ranges are fed into the else branch to recursively build the tree.

For classifying, if the new instance falls into a leaf node, the label of the leaf node is assigned. Otherwise, the else-branch is followed for using another attribute to classify. The last else-branch takes to the else-leaf. Any data that goes into this else-leaf is classified as undecided.

Experiment results

<table>
<thead>
<tr>
<th>FP</th>
<th>C4.5</th>
<th>RF</th>
<th>ElseT</th>
<th>C4.5</th>
<th>RF</th>
<th>ElseT</th>
<th>Elset</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>0.7%</td>
<td>0%</td>
<td>0%</td>
<td>0.2%</td>
<td>0%</td>
<td>0%</td>
<td>14.55%</td>
<td></td>
</tr>
<tr>
<td>T90</td>
<td>0.6%</td>
<td>1.9%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>16.67%</td>
<td></td>
</tr>
<tr>
<td>T80</td>
<td>1.8%</td>
<td>1.5%</td>
<td>0%</td>
<td>2.2%</td>
<td>0.4%</td>
<td>0%</td>
<td>14.51%</td>
<td></td>
</tr>
<tr>
<td>T70</td>
<td>1.8%</td>
<td>0.8%</td>
<td>0%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0%</td>
<td>12.09%</td>
<td></td>
</tr>
</tbody>
</table>

First, there were two classes of Y, as 0 and 1. Then, the non-uniform range is marked as undecided.

Acknowledgements
This research was supported by National Institutes of Health (GM116283) grant and iXpressGenes, Inc.
This study was presented at MABM in 2018 IEEE International Conference on Bioinformatics and BioMedicine.