LEARNING HETEROGENEOUS ENSEMBLES

- Biomedical data are abundant.
- Systems biology and machine learning can generate predictive models from data.

IDEA: A novel ensemble selection approach based on reinforcement learning, which provides a systematic way of exhaustively exploring the many possible combinations of base predictors that can be selected into an ensemble.

REINFORCEMENT LEARNING (RL) FOR ENSEMBLE SELECTION (ES)

**Reinforcement Learning**

- Possible actions
- Explore
- Exploit
- Learn a policy

**RL Strategies for ES**

Balance between ensemble performance and diversity:

- no diversity \rightarrow no improvement
- too much diversity \rightarrow low performing ensemble

We have designed several search strategies focused on:

**performance** (Stanescu and Pandey, PSB 2017)

**diversity** (Stanescu and Pandey, arXiv 2018)

**RESULTS:**

- RL ensembles are competitively predictive with the much larger ensembles consisting of all available base predictors, while being more parsimonious.

(Stanescu and Pandey, PSB 2017)

- Ensemble diversity, measured appropriately, can be incorporated to help the RL-based framework build even more accurate and parsimonious ensembles, at nearly only 30 – 40% of the complete ensemble size.

(Stanescu and Pandey, arXiv 2018)

<table>
<thead>
<tr>
<th>RL search strategy</th>
<th>auESC</th>
<th>size_ratio @50</th>
<th>size_ratio @120</th>
<th>size_ratio @140</th>
<th>perf_ratio @50</th>
<th>perf_ratio @120</th>
<th>perf_ratio @140</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RL_greedy</strong></td>
<td>0.647</td>
<td>0.761</td>
<td>0.676</td>
<td>0.618</td>
<td>0.993</td>
<td>0.996</td>
<td>0.999</td>
</tr>
<tr>
<td><strong>RL_pessimistic</strong></td>
<td>0.647</td>
<td>0.497</td>
<td>0.292</td>
<td>0.195</td>
<td>0.999</td>
<td>0.987</td>
<td>0.998</td>
</tr>
<tr>
<td><strong>RL_backtrack</strong></td>
<td>0.545</td>
<td>0.115</td>
<td>0.069</td>
<td>0.036</td>
<td>0.818</td>
<td>0.853</td>
<td>0.803</td>
</tr>
<tr>
<td><strong>RL_diversity_cosine</strong></td>
<td>0.657</td>
<td>0.418</td>
<td>0.368</td>
<td>0.335</td>
<td>1.010</td>
<td>1.012</td>
<td>1.012</td>
</tr>
<tr>
<td><strong>RL_diversity_euclidean</strong></td>
<td>0.650</td>
<td>0.389</td>
<td>0.358</td>
<td>0.326</td>
<td>0.990</td>
<td>0.998</td>
<td>1.008</td>
</tr>
<tr>
<td><strong>RL_diversity_correlation</strong></td>
<td>0.654</td>
<td>0.456</td>
<td>0.357</td>
<td>0.316</td>
<td>0.996</td>
<td>1.011</td>
<td>1.008</td>
</tr>
<tr>
<td><strong>RL_diversity_yule</strong></td>
<td>0.648</td>
<td>0.722</td>
<td>0.642</td>
<td>0.647</td>
<td>0.994</td>
<td>0.998</td>
<td>0.995</td>
</tr>
<tr>
<td><strong>RL_diversity_kappa</strong></td>
<td>0.647</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td>0.995</td>
<td>0.996</td>
</tr>
</tbody>
</table>

**Performance of the diversity-incorporated RL-based ensemble selection algorithms [2], as well as those proposed in our previous work [1] on a variety of splice site prediction datasets.**

**REFERENCES:**


**Implementation available:**
https://github.com/GauravPandeyLabens