Task

Given a set of documents written by author A and an unknown document, find whether the latter was written by A.
- Output: probability in [0, 1]
- Evaluation: product of
  - Area under the ROC curve (AUC),
  - c@1 (accuracy with “don’t know” answer)

Approach

- Supervised classification problem
- Combining multiple learners
- Genetic algorithm:
  - Training individual learners
  - Traning meta-model

Motivations

- Experience from PAN’14
- two complementary approaches
- PAN’14 meta-classifier performance

Configurations

- Representing distinct sets of parameters in a uniform way
- Set of parameter-value pairs: \( C = \{ p_1 \mapsto v_1, \ldots, p_n \mapsto v_n \} \)
- Meta-parameters of a strategy
  - Uniquely defines how to train a model
  - Very large space of possible configs

Genetic Algorithm

- Configurations = “individuals”
- Selects optimal configs for each strategy
- Parameters (at every generation):
  - Proportion of selected breeders: 10%.
  - Elite prop.: 10%; Random: 5%.
  - Probability of mutation: 0.02.

Architecture

For every dataset, 5 strategies are trained individually with the genetic algorithm. Their output are 5xN “optimal” configurations, which are then fed to the meta-training stage. In this stage, the genetic algorithm selects an optimal combination of configurations.

Results

- Influence of the size of the sample
  - English: only one known doc by case
  - Spanish: four known docs by case
- Similar perf on training and test set
  - no overfitting (except with Spanish)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Meta test set</th>
<th>Full training set</th>
<th>Test set perf.</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>0.710</td>
<td>0.722</td>
<td>0.635</td>
<td>1st</td>
</tr>
<tr>
<td>English</td>
<td>0.405</td>
<td>0.421</td>
<td>0.453</td>
<td>6th</td>
</tr>
<tr>
<td>Greek</td>
<td>0.696</td>
<td>0.761</td>
<td>0.693</td>
<td>2nd</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.950</td>
<td>0.952</td>
<td>0.661</td>
<td>4th</td>
</tr>
</tbody>
</table>

Acknowledgments

This research is supported by Science Foundation Ireland (Grant 12/CE/2267) as part of the Centre for Next Generation Localisation (www.cnegl.ie) funding at Trinity College, University of Dublin.