Delegating Classifiers

Cèsar Ferri\textsuperscript{1}, Peter Flach\textsuperscript{2}, José Hernández-Orallo\textsuperscript{1}

\textsuperscript{1} Dpt. de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain.

\textsuperscript{2} Department of Computer Science, University of Bristol, UK.

\{ cferri@dsic.upv.es, Peter.Flach@bristol.ac.uk, jorallo@dsic.upv.es \}

\textit{ICML-2004, Banff, July 5, 2004.}
• Introduction.
• Delegation as Separate-and-Conquer
• Establishing the Thresholds
• Scenarios
• Experiments
• Discussion
• Conclusions and Future Work
Introduction

• Many collaborative views of learning:
  – combination (ensembles, stacking/cascading),
  – co-learning, ...
• Generally composed of “total”, non-specialised classifiers, usually under-utilised.
• Learner specialisation:

But who determines the areas of specialisation?
And how?

  – Pre-refereeing: meta-learning, analysis of separability, ...
  – Post-refereeing: stacking, cascading, arbitrating, grading, ...
  – Self-refereeing: ???
• Delegation:

\[ \text{Training set } Tr \text{ is split using a threshold } \tau. \]

\[ \text{The threshold } \tau \text{ defines the degree of self-confidence.} \]

\[ Tr^{>}_f = \{ e \in Tr : f_{CONF}(e) > \tau \} \]

\[ Tr^{\leq}_f = Tr - Tr^{>}_f \]

• The overall classifier is used as follows:

<table>
<thead>
<tr>
<th>Decision Rule for a Delegating Classifier with threshold ( \tau ):</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IF</strong> ( f_{CONF}^{(1)}(e) &gt; \tau ) <strong>THEN</strong> PREDICT ( f_{CLASS}^{(1)}(e) )</td>
</tr>
<tr>
<td><strong>ELSE</strong> PREDICT ( f_{CLASS}^{(2)}(e) )</td>
</tr>
</tbody>
</table>
• If the task is classification, it is a multi-classifier method.
  – **Self-refereeing**: each classifier self-assigns its area of expertise.
  – **Serial**: not parallel or hierarchical.
  – **Transferring**: each prediction is made by only one classifier (no combination).
  – **Specialised**: based on partial classifiers.
  – **Attribute-preserving**: no new attributes are generated.
Delegation as Separate-and-Conquer

- The resulting model can be comprehensible.
Establishing the Threshold

• We use the *same* threshold for prediction (and for the test set).

How do we determine this threshold?

• This threshold is very dependent on the problem.

• Instead, we define a percentage of retention.
  – Two different ways:
    • Global Absolute Percentage (GAP): retain a fraction of the $\rho$ best ranked examples.
    • Stratified Absolute Percentage (SAP): retain a fraction of the $\rho$ best ranked examples *per class*. 
We investigated three different scenarios:

- Two stages: a master classifier and a slave classifier.

- Two stages with “round rebound”.

- Iterative: several chained stages.
Experiments

- Experimental methodology:
  - 22 datasets from UCI repository
  - Trained PETs (Probability Estimation Trees):
    - Smiles and Weka J4.8 variants of C4.5.
    - Pruning disabled.
    - Probability smoothing.
  - Evaluation:
    - 20x5-fold cross-validation.
    - Accuracy and AUC used as metrics.
Experiments

• Importance of a good probability estimation.
  – Four methods of PETs:
    – With pruning and no smoothing (Pr NoSmooth)
    – No pruning and no smoothing (NoPr NoSmooth)
    – No pruning and Laplace smoothing (Pr Laplace)
    – No pruning and Mbranch smoothing (Pr Mbranch) (ECML'03)

• Two-stage scenario. GAP ($\rho=0.5$).
  – Averaged results for the 22 datasets.

<table>
<thead>
<tr>
<th></th>
<th>Pr NoSmooth</th>
<th>NoPr NoSmooth</th>
<th>NoPr Laplace</th>
<th>NoPr Mbranch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Acc</td>
<td>83.88</td>
<td>83.72</td>
<td>83.72</td>
<td>83.72</td>
</tr>
<tr>
<td>Single AUC</td>
<td>86.46</td>
<td>87.16</td>
<td>90.18</td>
<td>90.78</td>
</tr>
<tr>
<td>Del50% Acc</td>
<td>84.01</td>
<td>83.81</td>
<td>84.77</td>
<td>84.73</td>
</tr>
<tr>
<td>Del50% AUC</td>
<td>85.93</td>
<td>87.16</td>
<td>90.89</td>
<td>91.31</td>
</tr>
</tbody>
</table>

• The way in which the master classifier is able to estimate its reliability is key to the success of the delegating method.
Experiments

- The proportion $\rho$ and Global/Stratified.
  - Two-stage scenario. GAP and SAP. Varying proportions $\rho$.
  - Averaged results for the 22 datasets.

\[
\begin{array}{|c|c|c|c|c|c|c|c|c|}
\hline
& \text{None} & 20\% & 33\% & 45\% & 50\% & 55\% & 67\% & 80\% \\
\hline
\text{GAP Acc} & 83.72 & 84.23 & 84.13 & 84.67 & 84.73 & 84.72 & 84.61 & 84.60 \\
\text{GAP AUC} & 90.78 & 91.02 & 91.14 & 91.27 & 91.31 & 91.24 & 91.04 & 90.92 \\
\text{SAP Acc} & 83.73 & 84.29 & 84.42 & 84.37 & 84.34 & 84.32 & 84.48 & 84.32 \\
\text{SAP AUC} & 91.31 & 90.79 & 90.79 & 90.65 & 90.61 & 90.47 & 90.26 & 89.85 \\
\hline
\end{array}
\]

- Proportions around 0.5 are optimal.
- The improvement is obtained with just around a 50% overhead.
- Stratified thresholds do not improve the results of global thresholds in general.
Experiments

- Iterative Scenario:
  - The greater the # of iterations the better the results.

<table>
<thead>
<tr>
<th></th>
<th>50%</th>
<th>33%</th>
<th>20%</th>
<th>10%</th>
<th>5%</th>
<th>2%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLOBAL Acc</strong></td>
<td>84.73</td>
<td>85.20</td>
<td>85.33</td>
<td>85.64</td>
<td>85.82</td>
<td>85.85</td>
<td>85.93</td>
</tr>
<tr>
<td><strong>GLOBAL AUC</strong></td>
<td>91.31</td>
<td>91.40</td>
<td>91.43</td>
<td>91.61</td>
<td>91.75</td>
<td>91.82</td>
<td>91.82</td>
</tr>
<tr>
<td><strong>GLOBAL #It</strong></td>
<td>2.00</td>
<td>3.16</td>
<td>4.68</td>
<td>7.74</td>
<td>12.50</td>
<td>21.64</td>
<td>31.31</td>
</tr>
<tr>
<td><strong>STRAT. Acc</strong></td>
<td>84.34</td>
<td>84.75</td>
<td>85.09</td>
<td>85.30</td>
<td>85.42</td>
<td>85.53</td>
<td>85.58</td>
</tr>
<tr>
<td><strong>STRAT. AUC</strong></td>
<td>90.61</td>
<td>90.70</td>
<td>90.74</td>
<td>91.02</td>
<td>91.25</td>
<td>91.44</td>
<td>91.51</td>
</tr>
<tr>
<td><strong>STRAT. #It</strong></td>
<td>2.00</td>
<td>3.06</td>
<td>4.33</td>
<td>6.70</td>
<td>9.83</td>
<td>15.30</td>
<td>18.24</td>
</tr>
</tbody>
</table>

- Once again, the “Global Absolute” variant is the best one.
- Execution times for 2% and 1% are around 8 and 10 times higher, respectively, than a single classifier.
- With similar times, delegation is close to bagging and not far behind boosting.
Discussion

• Factors that affect “Delegation”:
  – Reliability estimation (confidence) crucial.
  – Cannot be justified as a reduction of variance.
  – Patterns removed iteratively, as in Sep&Conq.
  – Class distribution is modified (better specialisation?).
  – Overfitting not so crucial as expected.

• Some of these factors may also explain a better improvement for accuracy than for AUC.
Conclusions and Future Work

• Delegation is a key idea in machine learning.
  – This work has used it systematically, using learners as building blocks for different scenarios.
  – As long as classifiers perform better probability estimation, they are more reliable for self-refereeing, crucial in delegation.
  – The method is simple, general and efficient.
  – In some configurations, it can preserve the comprehensibility of the base models by pruning and grafting them.
    • Only the useful parts are maintained.
Conclusions and Future Work

• Future work:
  – Use very efficient classifiers for the first stage and then more data-intensive ones for the subsequent stages.
  – Investigate the “combination” of the predictions.
  – Apply to regression and clustering.
  – Investigate other methods to determine the threshold (e.g. AUC-based, validation dataset, …).