Feature partitioning and boosting

Miklós Kurucz, Dávid Siklósi

Data Mining and Web Search Group
Computer and Automation Research Institute
Hungarian Academy of Sciences

joint work with several colleagues from Budapest
Methodology

- Large data set only
- 15,000 features -> partitioning, selection
- Feature evaluation as a weak pre-selection only
- Expected classifier combination to perform well over *partitioned feature set*
  - Might hold with knowledge of feature meaning
  - Did help in scaling, parallelization, exploration
- 10% heldout and 10% validation data set aside
- Access to large computational power, little additional time used after fast track
- Using Weka + scripts, tested many, many classifiers - *LogitBoost w/ decision stump* wins almost everywhere
Feature Partitioning

- Most frequent value at least 49500 times
  - Y
    - Nominal
      - Binary
        - Has missing value
          - Has at least 100 negative values
            - Continuous (10000< values)
              - Most frequent value > 48500
                - Fits exponential; >100 values
                  - Fits exponential
                    - DenseExp (530)
                      - SparseExp (445)
            - NonExp (587)
        - Unbalanced (540)
          - Cont10000 (503)
            - Neg100 (85)
              - Missing (330)
                - BinNum (1190)
                  - Nomin (290)
                    - Bad (10500)
### Performance of feature subsets

<table>
<thead>
<tr>
<th></th>
<th>churn heldout</th>
<th>churn valid</th>
<th>appetency heldout</th>
<th>appetency valid</th>
<th>upselling heldout</th>
<th>upselling valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing w/ LogitBoost</td>
<td>0.7232</td>
<td>0.7318</td>
<td>0.8394</td>
<td>0.8217</td>
<td>0.8855</td>
<td>0.8931</td>
</tr>
<tr>
<td>NonExp w/ AdaBoost</td>
<td>0.7188</td>
<td>0.7359</td>
<td>0.8551</td>
<td>0.8332</td>
<td>0.8835</td>
<td>0.8815</td>
</tr>
<tr>
<td>Nominal w/ LogitBoost</td>
<td>0.6657</td>
<td>0.6696</td>
<td>0.8385</td>
<td>0.7868</td>
<td>0.7623</td>
<td>0.7649</td>
</tr>
<tr>
<td>Cont10000 w/ Logitboost</td>
<td>0.6465</td>
<td>0.6631</td>
<td>0.6564</td>
<td>0.6712</td>
<td>0.7419</td>
<td>0.7474</td>
</tr>
<tr>
<td>BinNum w/ Logitboost</td>
<td>0.6369</td>
<td>0.6187</td>
<td>0.7204</td>
<td>0.7233</td>
<td>0.8016</td>
<td>0.8126</td>
</tr>
<tr>
<td>DenseExp w/ LogitBoost</td>
<td>0.6294</td>
<td>0.6473</td>
<td>0.6398</td>
<td>0.6591</td>
<td>0.7251</td>
<td>0.7391</td>
</tr>
<tr>
<td>NonExp w/ Bayes</td>
<td>0.6230</td>
<td>0.6531</td>
<td>0.5870</td>
<td>0.6393</td>
<td>0.7330</td>
<td>0.7224</td>
</tr>
<tr>
<td>Combination w/ LogitBoost</td>
<td></td>
<td></td>
<td>0.7667</td>
<td>0.8537</td>
<td>0.9100</td>
<td></td>
</tr>
<tr>
<td>Combination of log-odds w/ LogitBoost</td>
<td></td>
<td></td>
<td>0.7583</td>
<td>0.8361</td>
<td></td>
<td>0.9026</td>
</tr>
</tbody>
</table>

Table 1: The AUC value of feature subsets and the classifier combination over our 10+10% heldout and validation sets.
Feature Partitioning

- Most frequent value at least 49500 times
  - Bad (10500)
  - Nomin (260)
  - BinNum (1190)
  - Missing (330)
  - Neg100 (85)
  - Cont10000 (503)
  - Unbalanced (540)
  - DenseExp (530)
  - SparseExp (445)
  - NonExp (587)
  - Y

- Nominal
  - Has missing value
    - Continuous (10000< values)
      - Most frequent value > 48500
        - Fits exponential; >100 values
          - Fits exponential

1. Feature evaluation: weak pre-selection
   - Many non-predictive, highly correlated features
   - Threshold hard to set
   - IG, Chi^2 overscore many unique values
   - Gain Ratio overscore few unique values

2. LogitBoost itself uses a few selected features
   - Superlinear time, even 1000 features too much
   - Used over our feature partitioning
   - Used over random partition
Partitioned vs global over our heldout ...

<table>
<thead>
<tr>
<th></th>
<th>churn</th>
<th></th>
<th>appetency</th>
<th></th>
<th>upselling</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>heldout</td>
<td>valid</td>
<td>heldout</td>
<td>valid</td>
<td>heldout</td>
<td>valid</td>
</tr>
<tr>
<td>Combination LogitBoost</td>
<td>0.7667</td>
<td></td>
<td>0.8537</td>
<td></td>
<td>0.9100</td>
<td></td>
</tr>
<tr>
<td>Logitboost by partition</td>
<td>0.7557</td>
<td>0.7649</td>
<td>0.8668</td>
<td>0.8509</td>
<td>0.9122</td>
<td>0.9099</td>
</tr>
<tr>
<td>Logitboost random</td>
<td>0.7540</td>
<td>0.7612</td>
<td></td>
<td></td>
<td>0.9064</td>
<td>0.9069</td>
</tr>
<tr>
<td>Combination log-odds LogitBoost</td>
<td>0.7583</td>
<td></td>
<td>0.8361</td>
<td></td>
<td>0.9026</td>
<td></td>
</tr>
<tr>
<td>feature evaluation LogitBoost</td>
<td>0.7335</td>
<td>0.7414</td>
<td>0.8033</td>
<td>0.7924</td>
<td>0.8935</td>
<td>0.8868</td>
</tr>
</tbody>
</table>

Table 2: AUC values over our 10+10% heldout and validation sets.
... and the Cup test set

<table>
<thead>
<tr>
<th>Method</th>
<th>churn</th>
<th>appetite</th>
<th>upselling</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner (University of Melbourne)</td>
<td>0.7570</td>
<td>0.8836</td>
<td>0.9048</td>
<td>0.8484</td>
</tr>
<tr>
<td>LogitBoost + ADTree by partition (final)</td>
<td>0.7567</td>
<td>0.8736</td>
<td>0.9065</td>
<td>0.8456</td>
</tr>
<tr>
<td>LogitBoost by partition</td>
<td>0.7496</td>
<td>0.8683</td>
<td>0.9042</td>
<td>0.8407</td>
</tr>
<tr>
<td>Combination LogitBoost</td>
<td>0.7409</td>
<td>0.8561</td>
<td>0.8894</td>
<td>0.8288</td>
</tr>
</tbody>
</table>

Table 3: The AUC value of selected final methods over the test set.
Final best solution

- LogitBoost and ADTree
- Plain average turns out better than combination by classifiers
- Final results use all training set (combination by cross-validation)
- Final results (less than 20) evaluated over the 10% feedback - no overtraining, no difference in relative order
- Understand the variance (difference between 10% and full test set +0.02% for us but lot more for other teams)?
Further directions

- Partitioning by meaning (traffic, socio-demographic etc) might work better
- Would the same methods scale for larger data (M’s of users instead of 50K)?
- Staying power (prediction for future)?
- Evaluate graph stacking? Needs call graph
Questions?

Miklós Kurucz

mkurucz@ilab.sztaki.hu
http://datamining.sztaki.hu