A Combination of Boosting and Bagging for KDD Cup 2009 – Fast Scoring on a Large Database

Jianjun Xie
Viktoria Rojkova
Siddharth Pal
Stephen Coggeshall

ID Analytics, Inc.
San Diego, California, USA

June 28, 2009
Data Analysis

• 50% of 14740 numerical variables have 1 or 2 discrete values
• 80% of 260 categorical variables have categories < 10
• 79% of 14740 numerical variables have > 98% population filled by 0
Histogram Patterns

- Check any sampling bias between training and testing
- Discover variable scramble logic
### Improve Label Balance

<table>
<thead>
<tr>
<th>Churn</th>
<th>Appetency</th>
<th>Up-selling</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>41756</td>
<td>83.51%</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>3682</td>
<td>7.36%</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>890</td>
<td>1.78%</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>3672</td>
<td>7.34%</td>
</tr>
</tbody>
</table>

Down-sampling rate for each modeling task.

<table>
<thead>
<tr>
<th>Label</th>
<th>Down-sampling on negative examples</th>
<th>Positive rate after sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Churn</td>
<td>70%</td>
<td>10.17%</td>
</tr>
<tr>
<td>Appetency</td>
<td>20%</td>
<td>8.31%</td>
</tr>
<tr>
<td>Up-selling</td>
<td>90%</td>
<td>8.12%</td>
</tr>
</tbody>
</table>
Stochastic Gradient Boosting Tree (TreeNet)

- Fits many small trees
- Uses a function of the errors as weights
- Stochastic gradient boosting
  - Randomly selects the sample for each iteration of the boosting
  - Tests the model on held out data to obtain best parameters
Combine Boosting and Bagging

• A single TreeNet model may not see complete picture of all data
  - Inside TreeNet: split training and testing, sampling at each tree split
  - Outside TreeNet: down-sampling on negative population to improve label balance

• Bagging Boosted tree models together get improved performance
  - Iterations of bootstrap sampling (typically 5)
  - Performance (AUC) improve 0.3-0.5% on most out-of-bagging validations over a single model
Modeling Workflow

1. Data preprocessing
2. Variable reduction
3. Variable selection
4. Single model building
5. Bagging
Modeling on Slow Track

• Model performance of small dataset is lower than that of large dataset
• Competition rule requires small model compete with large model
• Our strategy
  - Unscramble the small dataset
  - Map variables between small and large
  - Focus model on large dataset
  - Use small as Reference (additional 10% validation feedback)
Unscramble the Small Dataset

• Unscramble the variable mapping
  - Compare the frequency distribution of each variable (histogram)
  - Able to map 194 out of 230 variables by histogram alone.

• Unscramble the example order
  - Construct a key using mapped variables (key = Var_i Var_j ... Var_n)
  - Cut key out in original order: sequence ID, key
  - Sort key files by key
  - Paste sequence ID files together
Final Results

AUC results of our final models on test dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Churn</th>
<th>Appetency</th>
<th>Up-selling</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>100%</td>
<td>10%</td>
<td>100%</td>
</tr>
<tr>
<td>Large (fast)</td>
<td>0.7333</td>
<td>0.7565</td>
<td>0.8705</td>
<td>0.9308</td>
</tr>
<tr>
<td></td>
<td>0.7390</td>
<td>0.7614</td>
<td>0.8714</td>
<td>0.9023</td>
</tr>
<tr>
<td>Small (slow)</td>
<td>0.7612</td>
<td>0.7611</td>
<td>0.8544</td>
<td>0.9155</td>
</tr>
</tbody>
</table>
Key Factors to Achieve the Results

• Combination of boosting and bagging
• Variable preprocessing and selection
• Proper imbalanced data handling