GENERATING WELL-BEHAVED LEARNING CURVES: AN EMPIRICAL STUDY

Gary M. Weiss
Alexander Battistin
Fordham University
Motivation

- Classification performance related to amount of training data
  - Relationship visually represented by learning curve
    - Performance increases steeply at first
    - Slope begins to decrease with adequate training data
    - Slope approaches 0 as more data barely helps

- Training data often costly
  - Cost of collecting or labeling

- “Good” learning curves can help identify optimal amount of data
Exploiting Learning Curves

- In practice will only have learning curve up to current number of examples when deciding on whether to acquire more
- Need to predict performance for larger sizes
  - Can do iteratively and acquire in batches
  - Can even use curve fitting
  - Works best if learning curves are well behaved
    - Smooth and regular
Prior Work Using Learning Curves

- Provost, Jensen and Oates\(^1\) evaluated progressive sampling schemes to identify point where learning curves begin to plateau
- Weiss and Tian\(^2\) examined how learning curves can be used to optimize learning when performance, acquisition costs, and CPU time are considered
  - “Because the analyses are all driven by the learning curves, any method for improving the quality of the learning curves (i.e., smoothness, monotonicity) would improve the quality of our results, especially the effectiveness of the progressive sampling strategies.”


What We Do

- Generate learning curves for six data sets
  - Different classification algorithms
  - Random sampling and cross validation
- Evaluate curves
  - Visually for smoothness and monotonicity
  - “Variance” of the learning curve
The Data Sets

<table>
<thead>
<tr>
<th>Name</th>
<th># Examples</th>
<th>Classes</th>
<th># Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>32,561</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>Coding</td>
<td>20,000</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Blackjack</td>
<td>15,000</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Boa1</td>
<td>11,000</td>
<td>2</td>
<td>68</td>
</tr>
<tr>
<td>Kr-vs-kp</td>
<td>3,196</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>452</td>
<td>2</td>
<td>279</td>
</tr>
</tbody>
</table>
Experiment Methodology

- Sampling strategies
  - 10-fold cross validation: 90% available for training
  - Random sampling: 75% available for training
- Training set sizes sampled at regular 2% intervals of available data
- Classification algorithms (from WEKA)
  - J48 Decision Tree
  - Random Forest
  - Naïve Bayes
Accuracy is not our focus, but a well behaved learning curve for a method that produces poor results is not useful. These results are for the largest training set size (no reduction)

J48 and Random Forest are competitive so we will focus on them

<table>
<thead>
<tr>
<th>Dataset</th>
<th>J48</th>
<th>Random Forest</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>86.3</td>
<td>84.3</td>
<td>83.4</td>
</tr>
<tr>
<td>Coding</td>
<td>72.2</td>
<td>79.3</td>
<td>71.2</td>
</tr>
<tr>
<td>Blackjack</td>
<td>72.3</td>
<td>71.7</td>
<td>67.8</td>
</tr>
<tr>
<td>Boa1</td>
<td>54.7</td>
<td>56.0</td>
<td>58.0</td>
</tr>
<tr>
<td>Kr-vs-kp</td>
<td>99.4</td>
<td>98.7</td>
<td>87.8</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>65.4</td>
<td>65.2</td>
<td>62.0</td>
</tr>
<tr>
<td>Average</td>
<td>75.1</td>
<td>75.9</td>
<td>71.7</td>
</tr>
</tbody>
</table>
Results: Variances

Variance for a curve equals average variance in performance for each evaluated training set size. The results are for 10-fold cross validation. Naïve Bayes is best followed by J48. But Naïve Bayes had low accuracy (see previous slide)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>J48</th>
<th>Random Forest</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>0.51</td>
<td>0.32</td>
<td>0.01</td>
</tr>
<tr>
<td>Coding</td>
<td>9.78</td>
<td>17.08</td>
<td>0.19</td>
</tr>
<tr>
<td>Blackjack</td>
<td>0.36</td>
<td>2.81</td>
<td>0.01</td>
</tr>
<tr>
<td>Boa1</td>
<td>0.20</td>
<td>0.31</td>
<td>0.73</td>
</tr>
<tr>
<td>Kr-vs-kp</td>
<td>3.54</td>
<td>12.08</td>
<td>4.34</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>41.46</td>
<td>15.87</td>
<td>9.90</td>
</tr>
</tbody>
</table>
J48 Learning Curves (10 xval)
Random Forest Learning Curves
Naïve Bayes Learning Curves
Closer Look at J48 and RF (Adult)
A Closer Look at J48 and RF (kr-vs-kp)
Closer Look at J48 and RF (Arrhythmia)
Now let's compare cross validation to Random Sampling, which we find generates less well behaved curves.
J48 Learning Curves (Blackjack Data Set)
RF Learning Curves
(Blackjack Data Set)
Conclusions

- Introduced the notion of well-behaved learning curves and methods for evaluating this property
- Naïve Bayes seemed to produce much smoother curves, but less accurate
  - Low variance may be because they consistently reach a plateau early
- J48 and Random Forest seem reasonable
  - Need more data sets to determine which is best
- Cross validation clearly generates better curves than random sampling (less randomness?)
Future Work

- Need more comprehensive evaluation
  - Many more data sets
  - Compare more algorithms
  - Additional metrics
    - Count number of drops in performance with greater size (i.e., “blips”). Need better summary metric.
  - Vary number of runs. More runs almost certainly yields smoother learning curves.

- Evaluate in context
  - Ability to identify optimal learning point
  - Ability to identify plateau (based on some criterion)
If Interested in This Area


- Contact me if you want to work on expanding this paper (gaweiss@fordham.edu)