1 Previous Empirical Comparisons

- STATLOG (1995)
  - Did not have boosting, SVMs and other recent methods.
  - Caruana and Niculescu-Mizil (2006)
    - Included newer methods, very thorough.
    - On average, boosted trees were the best, followed by random forests.
    - Neither study considered problems of high dimensionality.
    - Are those conclusions valid in high dimensions?
    - Teaser: Yes, up to some dimensionality. But in higher dimensions things are different in a semi-obvious way.

2 Methodology

- 11 datasets, ranging from 700 to 700K dimensions, mainly from biology, text and link prediction domains.
- 10 state of the art learning algorithms
- 100's of parameter settings
- 3 metrics: accuracy, squared loss, area under the ROC
- Why not use more than these three?

3 Small difficulties

- Coping with squared loss → calibration (Platt & Isotonic).
- Coping with different baselines → standardization
- Interpretation: a standardized score of 1.02 indicates 2% improvement over typical method.

4 Implementation Tricks

- Most high dimensional data is sparse.
- Specialized implementations for handling sparse data.
- Neural Nets
  - Forward: Matrix times sparse vector multiplication
  - Backward: Sparse input implies sparse gradient

- Momentum would make the updates non-sparse
- Decision Trees: Indexing by feature
- Kernel SVMs: Specialized large scale SVM solver LaSVM
- Still, experiments took 5-6 weeks in 40 cpus.

5 Average Over All Three Metrics

<table>
<thead>
<tr>
<th>DIM</th>
<th>AVG</th>
<th>Stu</th>
<th>Cal</th>
<th>Dig</th>
<th>Is2</th>
<th>Cry</th>
<th>Kdd</th>
<th>R-S</th>
<th>Cite</th>
<th>Dse</th>
<th>Spam</th>
<th>Imdb</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>1.094</td>
<td>1.02</td>
<td>1.009</td>
<td>1.007</td>
<td>1.019</td>
<td>1.005</td>
<td>1.001</td>
<td>1.032</td>
<td>1.013</td>
<td>1.006</td>
<td>1.007</td>
<td>1.010</td>
<td></td>
</tr>
<tr>
<td>ANN</td>
<td>1.006</td>
<td>0.997</td>
<td>1.005</td>
<td>1.005</td>
<td>0.996</td>
<td>1.016</td>
<td>1.015</td>
<td>0.993</td>
<td>1.006</td>
<td>1.004</td>
<td>1.002</td>
<td>1.004</td>
<td></td>
</tr>
<tr>
<td>BST</td>
<td>0.998</td>
<td>1.040</td>
<td>0.981</td>
<td>0.981</td>
<td>1.021</td>
<td>0.987</td>
<td>0.988</td>
<td>0.985</td>
<td>1.000</td>
<td>1.001</td>
<td>1.000</td>
<td>1.003</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.992</td>
<td>0.990</td>
<td>1.003</td>
<td>1.010</td>
<td>0.997</td>
<td>0.988</td>
<td>0.988</td>
<td>0.995</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.002</td>
<td></td>
</tr>
<tr>
<td>BGT</td>
<td>1.001</td>
<td>1.043</td>
<td>0.997</td>
<td>1.003</td>
<td>1.015</td>
<td>0.992</td>
<td>0.977</td>
<td>0.989</td>
<td>0.999</td>
<td>0.989</td>
<td>0.974</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>1.002</td>
<td>0.993</td>
<td>0.886</td>
<td>1.016</td>
<td>1.003</td>
<td>1.017</td>
<td>1.017</td>
<td>1.099</td>
<td>1.013</td>
<td>1.003</td>
<td>1.002</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>1.022</td>
<td>1.000</td>
<td>1.017</td>
<td>0.946</td>
<td>0.999</td>
<td>1.006</td>
<td>0.920</td>
<td>1.052</td>
<td>1.000</td>
<td>0.962</td>
<td>0.986</td>
<td>0.992</td>
<td></td>
</tr>
<tr>
<td>BSS</td>
<td>1.012</td>
<td>1.033</td>
<td>0.890</td>
<td>0.982</td>
<td>0.988</td>
<td>1.017</td>
<td>0.993</td>
<td>0.999</td>
<td>0.994</td>
<td>0.986</td>
<td>0.999</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>PRC</td>
<td>1.096</td>
<td>0.978</td>
<td>0.883</td>
<td>0.967</td>
<td>0.993</td>
<td>0.991</td>
<td>1.016</td>
<td>0.999</td>
<td>0.993</td>
<td>1.004</td>
<td>0.983</td>
<td>0.982</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>0.961</td>
<td>0.927</td>
<td>0.799</td>
<td>0.922</td>
<td>0.958</td>
<td>0.995</td>
<td>1.000</td>
<td>0.987</td>
<td>0.943</td>
<td>0.950</td>
<td>0.949</td>
<td>0.949</td>
<td></td>
</tr>
</tbody>
</table>

6 Trends

- Random Forests on really high dimensions.
- Random Forests vs. Boosted Trees.
- Consistency of ANNs.
- Diversity of best models.
- Not apparent from this table: calibration with Isotonic Regression is almost always better than Platt's method or no calibration.

7 Conclusions

- Our results confirm the findings of previous studies in low dimensions.
- But as dimensionality increases, boosted trees fall behind random forests.
- Non-linear methods can do well in high dimensions.
  - But they need appropriate regularization. (ANNs, Kernel SVMs, Random Forests)
- Calibration never hurts and almost always helps even for methods such as logistic regression and neural nets.

8 Acknowledgments

- This work began as a group project in a graduate machine learning course at Cornell.
- We thank everyone who participated in the course and especially the following students: Sergei Fotin, Michael Friedman, Myle Ott, Raghu Ramanujan, Alec Berntson, Eric Breck, and Art Munson.
- Random forest and other tree software: http://www.cs.cornell.edu/~nk/fest