Ensemble Approaches
Ensemble Methods Comparison

Our experimental evaluation:

• Five well-known ensemble methods.
• 57 public-domain datasets.
• 5 x 2-fold cross-val and F-test.
  (5 x 2-fold rather than 10-fold because …)
• Friedman-Holm test based on ranks.

Do fancier methods improve on bagging?
Ensemble Methods Comparison

Random subspaces.

- Randomly select $N_F$ of possible features.
- Create classifier using selected features.
- Repeat $N_C$ times.

(a=1, b=3.5, c=0.5; honest)
(a=2, b=3.2, c=0.3; fraud)

(b=3.5, c=0.5; honest)
(b=3.2, c=0.3; fraud)

(a=1, b=3.5; honest)
(a=2, b=3.2; fraud)
Ensemble Methods Comparison

Bagging

• Randomly select examples with replacement to make up the training data.

• Usually, you create a bag that is the same size as the original data.

• Typically, 100 or more bagged classifiers are created.
Ensemble Methods Comparison

Random trees.

- During the tree-building process:
  - Find the best $N_S$ splits at each node.
  - Randomly select one of these $N_S$ tests.

- Repeat to create $N_C$ trees.

Issues: continuous / discrete features?
Ensemble Methods Comparison

Random forests.

• Bag to select data for creating a tree.

• In creating a tree:
  – At each node, randomly select $N_F$ features.
  – Select the best test among these features.

• Repeat to create $N_C$ trees.
Ensemble Methods Comparison

Boosting

• Focus on misclassified examples by weighting them more.

• Either use integer weights with repeated examples or incorporate weights into the learning algorithm.

• AdaBoost.M1 works as shown on the next slide.
AdaBoost.M1

• Assign equal weight to each training instance.
• For each of t iterations
  – Apply Learning algorithm to weighted dataset and store resulting model.
  – Compute error e of model on weighted dataset and store error.
  – If e equal to 0 or e >= 0.5:
    • Terminate model generation.
  – For Each instance in dataset:
    • If instance classified correctly by model:
      – Multiply weight of instance by e/(1-e).
    – Normalize weight of all instances.
AdaBoost.M1 - Classification

- Assign weight of 0 to all classes.
- For each the t (or less) models:
  - Add $-\log(e/(1-e))$ to weight of class predicted by model.
- Return class with highest weight
Experimental comparison.

• 1,000-classifier ensemble by each method.
• Boosting also evaluated at 50 classifiers.
• Accuracy of others compared to bagging.
• Compare statistically significant wins and losses in accuracy out of 57 datasets.
• Use Bonferroni correction, F-Test with 5x2 fold cross validation.
Ensemble Methods Comparison

• Rank the classifiers from 1 for the most accurate on a data set to 8 for the least accurate.

• Two tie at 3, get a rank of 3.5.

• Apply the nonparametric Freidman test to see if there are differences with many classifiers and many data sets.
Ensemble Methods Comparison

• If Freidman test indicates there is a difference, the Holm test can be used.

• The Holm test allows the comparison of one classifier (bagging) against the rest by differences in rank.

• This approach does not have a problem with overlapping training sets.
## Ensemble Methods Comparison

(statistical significance at 0.05 level)

<table>
<thead>
<tr>
<th>Method</th>
<th>Win</th>
<th>Loss</th>
<th>“Tie”</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forests-2</td>
<td>5</td>
<td>2</td>
<td>50</td>
<td>3.32</td>
</tr>
<tr>
<td>Random Forests-lg</td>
<td>6</td>
<td>0</td>
<td>51</td>
<td>3.7</td>
</tr>
<tr>
<td>Random Trees</td>
<td>2</td>
<td>4</td>
<td>51</td>
<td>4.53</td>
</tr>
<tr>
<td>Random Subspaces</td>
<td>5</td>
<td>9</td>
<td>43</td>
<td>5.39</td>
</tr>
<tr>
<td>Boosting (50)</td>
<td>6</td>
<td>0</td>
<td>51</td>
<td>5.15</td>
</tr>
<tr>
<td>Boosting (1000)</td>
<td>8</td>
<td>0</td>
<td>49</td>
<td>3.34</td>
</tr>
<tr>
<td>Bagging</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.06</td>
</tr>
</tbody>
</table>
Ensemble Methods Comparison

Conclusions.

• Boosting and RF-lg improve on accuracy of bagging in about 10% of datasets.

• Boosting appears to benefit from larger ensemble sizes than once thought.

• *Friedman-Holm tells us only boosting-50 and random subspaces fail to improve on bagging.*

• Methods to automatically choose ensemble size may be important topic to develop.
Conclusions.

- While most approaches are not much more accurate than bagging, they are consistently more accurate.