What’s better than a tree?
Random Forest

- Can often improve performance of decision tree classifiers using a set of decision trees (a forest)
- Each tree trained on a random subset of training data
- Classify a data instance using all trees
- Combine answers to make classification
  - E.g., vote for most common class
Bagging

• Idea can be used on any classifier!
• Improve classification by combining classifications of randomly selected training subsets
• Bagging = **Bootstrap aggregating**
  
  An *ensemble* meta-algorithm that can improve stability & accuracy of algorithms for statistical classification and regression

• Helps avoid overfitting
Choosing subsets of training data

- Classic bagging: select random subset of training instances with replacement
- Pasting: select random subset of training instances
- Random Subspaces: use all training instances, but with a random subset of features
- Random Patches: random subset of instances and random subset of features
- What’s best? YMMV: depends on problem, training data, algorithm
Examples

• Two examples using Weka
  – UCI Auto mpg prediction dataset
  – UCI Adult income prediction dataset

• RandomForest improves over J48 for the smaller dataset, but not for the larger

• Takeaway: more data is always best
The Auto MPG dataset is a multivariate data set with 398 instances and 8 attributes. It is categorical and real-valued. The task associated with this data set is regression. Missing values are present in the data. The dataset was donated on 1993-07-07.

The source of the data set is the StatLib library maintained at Carnegie Mellon University. The data set was used in the 1983 American Statistical Association Exposition.

Data Set Information:

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)
UCI Auto MGP Dataset (2)

• Data from 1983
• 398 instances
• Predict auto mpg from seven attributes: cylinders, displacement, horsepower, weight, acceleration, model year, and origin
Classifier:

Choose J48 -C 0.25 -M 2

Test options:

- Use training set
- Supplied test set
- Cross-validation
- Percentage split

Classifier output:

Time taken to build model: 0.01 seconds

--- Evaluation on training set ---

Time taken to test model on training data: 0 seconds

--- Summary ---

Correctly Classified Instances 230 95.8333 %
Incorrectly Classified Instances 10 4.1667 %
Kappa statistic 0.9174
Mean absolute error 0.0453
Root mean squared error 0.1505
Relative absolute error 13.4303 %
Root relative squared error 36.7193 %
Total Number of Instances 240

--- Detailed Accuracy By Class ---

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.987</td>
<td>0.025</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
<td>0.963</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>b</td>
<td>0.881</td>
<td>0.015</td>
<td>0.925</td>
<td>0.881</td>
<td>0.902</td>
<td>0.883</td>
<td>0.991</td>
<td>0.954</td>
</tr>
<tr>
<td>c</td>
<td>0.923</td>
<td>0.025</td>
<td>0.878</td>
<td>0.923</td>
<td>0.900</td>
<td>0.880</td>
<td>0.989</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Weighted Avg.:

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.958</td>
<td>0.023</td>
<td>0.959</td>
<td>0.958</td>
<td>0.958</td>
<td>0.935</td>
<td>0.995</td>
<td>0.978</td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---

```
 a  b  c <- classified as
157 1  1 | a = 1
 1 37  4 | b = 2
 1 2  36 | c = 3
```
Classifier:

RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options:

- Use training set

Classifier output:

Time taken to build model: 0.1 seconds

Evaluation on training set:

Correctly Classified Instances 240 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0.0674
Root mean squared error 0.114
Relative absolute error 19.9659 %
Root relative squared error 27.8064 %
Total Number of Instances 240

Detailed Accuracy by Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

weighted Avg.: 1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000

Confusion Matrix:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>159</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>39</td>
</tr>
</tbody>
</table>

- classified as:
  - a = 1
  - b = 2
  - c = 3

Status:

OK
Results

• Relatively small dataset allows construction of a DT model that does very well
• Using Random Forest still improves on it
• This is considered to be poor methodology since it overfits to the particular training set
Classifier:

Choose J48 -C 0.25 -M 2

Test options:

- Supplied test set

Classifier output:

Size of the tree: 49

Time taken to build model: 0.02 seconds

Evaluation on test set:

Time taken to test model: 0 seconds

Summary:

| Correctly Classified Instances | 112 | 84.8485 % |
| Incorrectly Classified Instances | 20  | 15.1515 % |
| Kappa statistic | 0.7255 |
| Mean absolute error | 0.1198 |
| Root mean squared error | 0.2915 |
| Relative absolute error | 32.9443 % |
| Root relative squared error | 66.1432 % |

Total Number of Instances: 132

Detailed Accuracy By Class:

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.907</td>
<td>0.127</td>
<td>0.916</td>
<td>0.987</td>
<td>0.950</td>
<td>0.877</td>
<td>0.967</td>
<td>0.962</td>
</tr>
<tr>
<td>2</td>
<td>0.650</td>
<td>0.063</td>
<td>0.650</td>
<td>0.650</td>
<td>0.650</td>
<td>0.588</td>
<td>0.851</td>
<td>0.660</td>
</tr>
<tr>
<td>3</td>
<td>0.657</td>
<td>0.062</td>
<td>0.793</td>
<td>0.657</td>
<td>0.719</td>
<td>0.735</td>
<td>0.887</td>
<td>0.690</td>
</tr>
</tbody>
</table>

Weighted Avg.:

| Weighted Avg. | 0.848 | 0.100 | 0.843 | 0.848 | 0.843 | 0.769 | 0.928 | 0.844 |

Confusion Matrix:

```
 a  b  c  --- classified as
76  0  1  | a = 1
 2 13  5  | b = 2
 5  7 23  | c = 3
```
Classifier

Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

- Use training set
- Supplied test set

Classifier output

Debugging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.09 seconds

Time taken to test model on test set: 0.01 seconds

Summary

Correctly Classified Instances 115 87.1212%
Incorrectly Classified Instances 17 12.8788%
Kappa statistic 0.7653
Mean absolute error 0.1642
Root mean squared error 0.2605
Relative absolute error 45.1528%
Root relative squared error 59.0951%
Total Number of Instances 132

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.974</td>
<td>0.164</td>
<td>0.893</td>
<td>0.974</td>
<td>0.932</td>
<td>0.831</td>
<td>0.988</td>
<td>0.992</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0.750</td>
<td>0.036</td>
<td>0.789</td>
<td>0.750</td>
<td>0.769</td>
<td>0.730</td>
<td>0.961</td>
<td>0.838</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.714</td>
<td>0.041</td>
<td>0.862</td>
<td>0.714</td>
<td>0.781</td>
<td>0.718</td>
<td>0.965</td>
<td>0.910</td>
<td>3</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.871 0.112 0.869 0.871 0.867 0.785 0.978 0.947

Confusion Matrix

a  b  c  --- classified as
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>25</td>
<td>3</td>
</tr>
</tbody>
</table>
AUTO MPG Results (2)

• Using an independent test set shows more realistic balanced F1 score of .843
• Using Random Forest raises this to .867
• While the increase is not large, it is probably statistically significant
• F1 scores this tish are difficult to increase dramatically
Adult Data Set

Download Data Folder Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate</th>
<th>Number of Instances: 48842</th>
<th>Area:</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Categorical, Integer</td>
<td>Number of Attributes: 14</td>
<td>Date Donated: 1996-05-01</td>
<td></td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Classification</td>
<td>Missing Values?: Yes</td>
<td>Number of Web Hits: 1470139</td>
<td></td>
</tr>
</tbody>
</table>

Source:

Donor:
Ronny Kohavi and Barry Becker
Data Mining and Visualization
Silicon Graphics,
ea-mail: ronnyk[@]live.com for questions.

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AGE>16) & (AGI>100) & (AFNLWGT>1) & (HRSWK>0))
Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:
UCI Adult Dataset (2)

- Data on adults from 1994 census data
- 48,842 instances
- Predict if person makes over $50K/year
  - Equivalent to ~$86K/year today
- 14 features including age, education, marital status, occupation, race, sex, native country, ...
  - Mixture of numeric (e.g., age) and nominal (e.g., occupation) values
Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set

Classifier output

Size of the tree: 911

Time taken to build model: 2.64 seconds

Time taken to test model on training data: 0.16 seconds

Summary

- Correctly Classified Instances: 42803 (87.6356%)
- Incorrectly Classified Instances: 6039 (12.3644%)
- Kappa statistic: 0.6325
- Mean absolute error: 0.1861
- Root mean squared error: 0.3048
- Relative absolute error: 51.1076%
- Root relative squared error: 71.4388%
- Total Number of Instances: 48842

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.631</td>
<td>0.046</td>
<td>0.810</td>
<td>0.631</td>
<td>0.710</td>
<td>0.680</td>
<td>0.907</td>
<td>0.792</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>0.954</td>
<td>0.369</td>
<td>0.891</td>
<td>0.954</td>
<td>0.921</td>
<td>0.840</td>
<td>0.907</td>
<td>0.960</td>
<td>&lt;=50K</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.876</td>
<td>0.292</td>
<td>0.872</td>
<td>0.767</td>
<td>0.871</td>
<td>0.640</td>
<td>0.907</td>
<td>0.920</td>
<td></td>
</tr>
</tbody>
</table>

Confusion Matrix

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>--- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>7375</td>
<td>4312</td>
<td>a = &gt;50K</td>
</tr>
<tr>
<td>1727</td>
<td>3542</td>
<td>b = &lt;=50K</td>
</tr>
</tbody>
</table>
RandomForest -P 100 -l 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Use training set

Correctly Classified Instances 48774  99.8608%
Incorrectly Classified Instances 68  0.1392%
Kappa statistic 0.9962
Mean absolute error 0.0737
Root mean squared error 0.1263
Relative absolute error 20.2565%
Root relative squared error 29.6022%
Total Number of Instances 48842

Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.995</td>
<td>0.000</td>
<td>1.000</td>
<td>0.995</td>
<td>0.997</td>
<td>0.996</td>
<td>1.000</td>
<td>1.000</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>1.000</td>
<td>0.005</td>
<td>0.998</td>
<td>1.000</td>
<td>0.999</td>
<td>0.996</td>
<td>1.000</td>
<td>1.000</td>
<td>&lt;=50K</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.999  0.004  0.999  0.999  0.999  0.996  1.000  1.000  <=50K

Confusion Matrix

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>11624</td>
<td>63</td>
</tr>
</tbody>
</table>
| 5  | 37150 | classified as

a = >50K
b = <=50K
Result

- Significant increase on F1 scores when both trained and evaluated on training set
- This is considered to be poor methodology since it overfits to the particular training set
Create train and test collection

• Train has \(~95\%\) of data, test 5%
• Trained models for J48 and random forest using train dataset
• Tested on test data set
• Results were that random forest was (at best) about the same as J48
• Large dataset reduced problem of overfitting, so random Forest did not help
Correctly Classified Instances   2155   86.2 %
Incorrectly Classified Instances 345   13.8 %
Kappa statistic                  0.5988
Mean absolute error              0.1191
Root mean squared error          0.3196
Relative absolute error          52.5531 %
Root relative squared error      74.1954 %
Total Number of Instances        2500

Detailed Accuracy By Class:

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;50K</td>
<td>0.611</td>
<td>0.057</td>
<td>0.780</td>
<td>0.611</td>
<td>0.686</td>
<td>0.606</td>
<td>0.895</td>
<td>0.759</td>
<td>&gt;50K</td>
</tr>
<tr>
<td>&lt;=50K</td>
<td>0.944</td>
<td>0.360</td>
<td>0.882</td>
<td>0.944</td>
<td>0.912</td>
<td>0.606</td>
<td>0.895</td>
<td>0.953</td>
<td>&lt;=50K</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.862</td>
<td>0.307</td>
<td>0.857</td>
<td>0.862</td>
<td>0.856</td>
<td>0.606</td>
<td>0.895</td>
<td>0.905</td>
<td></td>
</tr>
</tbody>
</table>

Confusion Matrix:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>376</td>
<td>239</td>
</tr>
</tbody>
</table>
| 106| 1779| classified as
a = >50K
b = <=50K
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Re-evaluation on test set ===

User supplied test set
Relation: adult
Instances: unknown (yet). Reading incrementally
Attributes: 15

=== Summary ===

Correctly Classified Instances 2146 85.84%
Incorrectly Classified Instances 354 14.16%
Kappa Statistic 0.59
Mean absolute error 0.195
Root mean squared error 0.3272
Total Number of Instances 2500

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.610</td>
<td>0.066</td>
<td>0.767</td>
<td>0.610</td>
<td>0.679</td>
<td>0.596</td>
<td>0.893</td>
<td>0.765</td>
<td>&gt;50K</td>
</tr>
<tr>
<td></td>
<td>0.940</td>
<td>0.396</td>
<td>0.881</td>
<td>0.940</td>
<td>0.909</td>
<td>0.596</td>
<td>0.893</td>
<td>0.959</td>
<td>&lt;=50K</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.858 0.309 0.881 0.940 0.909 0.596 0.893 0.959

=== Confusion Matrix ===

   | a  | b  |
---|----|----|
| a | 375| 240|
| b | 114| 1771|
Conclusions

• Bagging can help, especially of the amount of data is adequate but not as large as it should be

• While we explore it using decision trees, it can be applied to any classifier
  – Scikit-learn has a general module for bagging

• In general, using any of several ensemble approaches to classification is often very helpful