Dynamic Model Selection and Combination in Forecasting: an Empirical Evaluation of Bagging and Boosting

The 31st Annual International Symposium on Forecasting | June 26-29, 2011 | Prague | The University of Economics

Devon K. Barrow
Dr. Sven F. Crone
Combination vs. Bagging vs. Boosting?

- Motivation
  - Model Selection or Combination?
  - The Gap: forecasting vs. machine learning
  - Introduction to Bagging and Boosting
- Empirical Evaluation
  - Experimental Design: NN3 Data
  - Experimental results
- Conclusion and Future Work
Competing models and uncertainty
- Model selection or model combination?
- *Ex post* or *Dynamic* combination?

Forecasting research and industry best practice
- (Individual) Model selection (see also Fildes, 89):
  - Information criteria (e.g. AIC, BIC) using MLE (e.g. R, Hyndman, 2008, etc.)
  - 1-step-ahead in-sample errors, based on \( \min C(e) \) (e.g. SAP APO)
  - Cross validation \((k\text{-fold})\) possibly with early stopping e.g. Neural networks
  -> Contradiction of recommendations & best practices in model selection!
- Model combination:
  - M3-Competition (Makridakis et al. 2000) \(\rightarrow\) simple average (CombS-H-D) outperforms others
  - More robust and accurate than individual forecast. (Newbold and Granger, 74) (Palm and Zellner, 92) etc...
  - Generally leads to improved accuracy. (Stock and Watson, 2004) (Fildes, Nikolopoulos et al. 2008) etc...
  -> Contradiction between model combination & model selection (special case of comb.?)

Machine Learning
- **Bagging** and **Boosting** are most established, with excellent track record
- Predictive classification models are regularly combined in a *dynamic* way, not ex post
  -> Limited evaluations in time series prediction & forecasting

-> Contradiction of best practices in (automatic) model selection
-> No systematic evaluation of empirical accuracy across domains \(\rightarrow\) GAP in research
Results of literature survey (ISI Web of Knowledge Database)

- 800 papers in classification & regression → 15 apply boosting to time series forecasting.
- Limited scope of data
  - 8 of 15 focus in Finance (volatility prediction)
  - 4 forecast the sunspot or Mackey glass synthetic data
  - 7 apply AdaBoost: 1 uses multiple time series.
  → Focus primarily on (marginal) extensions of the algorithms
- Limitations of experimental design
  - None evaluate > 11 real time series and 12 evaluate < 2
  - 2 use rolling origin evaluation (Tashman, 2000)
  - MAE, Squared errors – do not compare across time series
  → Invalid and unreliable empirical evidence on method performance
- Limitations of comparisons across combination families
  - Only 3 papers compare Bagging vs. Boosting (however limited data)
  - No comparison against benchmark forecasting methods e.g. ExSmoothing
  → Limits evidence of performance gains from boosting over alternatives

→ Significant omissions in application domain
→ Substantial shortcomings in evaluative design
Bootstrap and Aggregating

Training Set $S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \}$
Bootstrap and Aggregating

1) $S_1'(x,y) \rightarrow M_1(x) \rightarrow$  

2) $S_2'(x,y) \rightarrow M_2(x) \rightarrow$  

3) $S_3'(x,y) \rightarrow M_3(x) \rightarrow$
Boosting

1) $W^1$

$S(x, y)$

Errors$^1$, $\beta^1$

2) $W^2$

$S(x, y)$

Errors$^2$, $\beta^2$

3) $W^3$

$S(x, y)$

Weights

Value

1 4 7 10 13 16 19 22

0.00 0.02 0.04 0.06 0.08

0.10

0.10
and after three iterations ...

Bagged output:

\[ M_e(x) = M_1(x) + M_2(x) + \ldots + M_3(x) \]

Boosted output:

\[ M_e(x) = \beta_1 M_1(x) + \beta_2 M_2(x) + \ldots + \beta_3 M_3(x) \]
## Agenda

<table>
<thead>
<tr>
<th>Combination vs. Bagging vs. Boosting?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motivation</strong></td>
</tr>
<tr>
<td>• Model Selection or Combination?</td>
</tr>
<tr>
<td>• The Gap: forecasting vs. machine learning</td>
</tr>
<tr>
<td>• Introduction to Bagging and Boosting</td>
</tr>
<tr>
<td><strong>Empirical Evaluation</strong></td>
</tr>
<tr>
<td>• Experimental Design: NN3 Data</td>
</tr>
<tr>
<td>• Experimental results</td>
</tr>
<tr>
<td><strong>Conclusion and Future Work</strong></td>
</tr>
</tbody>
</table>
Experimental Design

**Times series data**
- NN3 competition complete dataset (NN3, 2007).

<table>
<thead>
<tr>
<th>Length</th>
<th>Seasonal</th>
<th>Non-seasonal</th>
<th>HARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long</td>
<td>25</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>Short</td>
<td>25</td>
<td>25</td>
<td>7</td>
</tr>
</tbody>
</table>

**Experimental Setup**
- The following experimental setup is used:
  - Forecast horizon: 12 months
  - Holdout period: 18 months
  - Error Measures: SMAPE and MAE.
  - Rolling origin evaluation (Tashman, 2000).

**Established benchmark dataset** (taken from M3 data)

**CONDITIONS** where Boosting / Bagging works (long vs. short, seasonal vs. non seasonal)
## Experimental Design

### Model Design

<table>
<thead>
<tr>
<th>Dynamic Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaBoost.R2 (Boost AR2)</td>
</tr>
<tr>
<td>AdaBoost.RT (Boost ART)</td>
</tr>
<tr>
<td>Bagging (Bagg)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Static Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Averaging (ModAvg)</td>
</tr>
</tbody>
</table>

### Combination versus selection

- **Aggregate Selection**
  - Random Walk
  - Seasonal Random Walk
- **Individual Selection**
  - Exponential Smoothing (ExSm)
  - Single MLP (Model Selection)

### Network design:
- Univariate Multiplayer Perceptron (MLP) with $Y_t$ up to $Y_{t-13}$ lags.
- 30 runs using random initialisations
- $^1$Family of exponential smoothing algorithms.
### Experimental Results

**Dynamic combination versus single MLP model**

<table>
<thead>
<tr>
<th>Method</th>
<th>SMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
</tr>
<tr>
<td>Single MLP</td>
<td>13.1%</td>
</tr>
<tr>
<td>Boost AR2</td>
<td>12.3%</td>
</tr>
<tr>
<td>Boost ART</td>
<td>13.4%</td>
</tr>
<tr>
<td>Bagging</td>
<td>12.7%</td>
</tr>
</tbody>
</table>

Bagging and Boosting improve forecast accuracy. Bagging outperforms Boosting. Significance (Freidman Nemenyi) confirmed by test (not shown).

Results averaged across 111 time series, 30 initialisations.
## Experimental Results

**Model combination versus single model – Gains in forecast accuracy**

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>MAE % Reduction (increase)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Validation</td>
</tr>
<tr>
<td>Single MLP</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boost AR2</td>
<td>602.95</td>
<td>720.21</td>
</tr>
<tr>
<td>Boost ART</td>
<td>675.03</td>
<td>710.16</td>
</tr>
<tr>
<td>Bagg</td>
<td>626.55</td>
<td>668.90</td>
</tr>
</tbody>
</table>

Results averaged across 111 time series, 30 initialisations

- Quantifiable gains in accuracy from Bagging and boosting
- Significance (Freidman Nemenyi) confirmed by test (not shown)
## Experimental Results

**Model Selection** versus **Model Combination** versus **Benchmarks**

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Set</th>
<th>Rank</th>
<th>Validation Set</th>
<th>Rank</th>
<th>Test Set</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>12.16%</td>
<td>4</td>
<td>14.78%</td>
<td>5</td>
<td>16.23%</td>
<td>6</td>
</tr>
<tr>
<td>AR2</td>
<td><strong>10.32%</strong></td>
<td><strong>1</strong></td>
<td>12.63%</td>
<td>2</td>
<td>15.51%</td>
<td>3</td>
</tr>
<tr>
<td>ART</td>
<td>12.44%</td>
<td>5</td>
<td>13.42%</td>
<td>3</td>
<td>15.40%</td>
<td>2</td>
</tr>
<tr>
<td>Bag</td>
<td>11.67%</td>
<td>3</td>
<td><strong>12.62%</strong></td>
<td><strong>1</strong></td>
<td><strong>15.08%</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>EnsAvg</td>
<td>11.55%</td>
<td>2</td>
<td>13.72%</td>
<td>4</td>
<td>16.13%</td>
<td>4</td>
</tr>
<tr>
<td>NaiveL</td>
<td>23.33%</td>
<td>8</td>
<td>–</td>
<td>–</td>
<td>21.19%</td>
<td>8</td>
</tr>
<tr>
<td>NaiveS</td>
<td>19.03%</td>
<td>7</td>
<td>–</td>
<td>–</td>
<td>17.75%</td>
<td>7</td>
</tr>
<tr>
<td>ExSm</td>
<td>18.82%</td>
<td>6</td>
<td>–</td>
<td>–</td>
<td>16.19%</td>
<td>5</td>
</tr>
</tbody>
</table>

Results averaged across 111 time series

→ Bagging and boosting perform best across all time series
→ AdaBoost.R2 is best on training data – possible overfitting
### Experimental Results

#### Comparing against NN3 Competition results

<table>
<thead>
<tr>
<th>Competition Code</th>
<th>Method Name</th>
<th>SMAPE</th>
<th>Rank</th>
<th>All Candidates</th>
<th>NN Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>B09</td>
<td>Wildi</td>
<td>14.84</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B07</td>
<td>Theta</td>
<td>14.89</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C27</td>
<td>Illies</td>
<td>15.18</td>
<td>3</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B03</td>
<td>ForecastPro</td>
<td>15.44</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>** Bagging</td>
<td></td>
<td><strong>15.82</strong></td>
<td>5</td>
<td></td>
<td><strong>2</strong></td>
</tr>
<tr>
<td>B16</td>
<td>DES</td>
<td>15.90</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B17</td>
<td>Comb S-H-D</td>
<td>15.93</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B05</td>
<td>Autobox</td>
<td>15.95</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C03</td>
<td>Flores</td>
<td>16.31</td>
<td>9</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>** AdaBoost.RT</td>
<td></td>
<td><strong>16.39</strong></td>
<td>10</td>
<td></td>
<td><strong>4</strong></td>
</tr>
<tr>
<td>B14</td>
<td>SES</td>
<td>16.42</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B15</td>
<td>HES</td>
<td>16.49</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>** AdaBoost.R2</td>
<td></td>
<td><strong>16.80</strong></td>
<td>13</td>
<td></td>
<td><strong>5</strong></td>
</tr>
<tr>
<td>** MLP</td>
<td></td>
<td><strong>16.94</strong></td>
<td>15</td>
<td></td>
<td><strong>7</strong></td>
</tr>
</tbody>
</table>

Bagging improves base MLP model by 10 places, Boosting 5 places.
Conclusions and Future Work?

- **Conclusions**
  - Bagging and boosting improve forecasting accuracy of a single base model
  - Bagging is however better than boosting (both algorithms)
  - **Some conflicts with previously reported findings**
    - AdaBoost.RT outperformed AdaBoost.R2 (Shrestha and Solomantine, 06)
      - **Limitation**: few time series - 2 hydrological, laser data (Santa Fe).
    - AdaBoost.R2 outperforms bagging (Drucker, 97)
      - **Limitation**: small number of time series - laser data and Mackey-Glass
    - Bagging and boosting (dynamic) performed better than
      - Model selection (select the best MLP)
      - Neural Network Model averaging
    - Boosting is prone to overfitting - noise and outliers

- **Future work**
  - Investigate performance of boosting
  - Meta parameter analysis of boosting algorithms
    - Loss function, combination method and stopping criteria
    - Can the performance of boosting be improved? **Yes**
  - Bagging and boosting: diversity, base learner, combination method and size


These slides are from a presentation by Dr. Sven F. Crone and Devon Barrow and subject to copyright ©. You may use these slides for internal purpose, so long as they are clearly identified as being created and copyrighted © by Dr. Sven F. Crone and Devon Barrow, Lancaster Centre for Forecasting. You may not use the text and images in a paper, tutorial or external training without explicit prior permission from Dr. Crone.