An Empirical Comparison of Algorithms for Aggregating Expert Predictions

Varsha Dani, U. Chicago
Brian Galebach, ProbabilitySports
Omid Madani, Yahoo! Research
David Pennock, Yahoo! Research
Sumit Sanghai, U. Washington
Aggregating Experts

• The problem
  • Predict the outcomes of a series of uncertain events
  • Given: probability estimates from experts

• Methods
  • Opinion pools (belief aggregation) [GZ]
  • Experts algorithms (worst case) [CBFD]
  • Machine learning (avg case) [GKV,Kah]
  • Prediction markets [WZ,SSWPG,CCMP]
An Empirical Study

- We analyze the performance of
  - Average
  - Experts algorithms
  - Machine learning
  - Market simulation
  - New: Variance algorithm
- On a unique real-world dataset
  - ProbabilitySports.com:NFL Football + NCAA
  - Actual prob assessment for 1319 games
  - Contestants rewarded with quadratic score
- According to
  - 0/1 accuracy
  - Quadratic score: 100-400(pred-actual)^2
Example

- ProbabilitySports.com
- ProbabilityFootball.com

<table>
<thead>
<tr>
<th></th>
<th>Pittsburgh</th>
<th>Seattle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contestant 1</td>
<td>67</td>
<td>33</td>
</tr>
<tr>
<td>Contestant 2</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Algorithm</td>
<td>71</td>
<td>29</td>
</tr>
</tbody>
</table>

QScore
The Data

- Many experts are bad; Median expert bad
- Experts are miscalibrated (overconfident)
- “Expertness” does carry over across games
Experts algorithms

• Basic: Online algorithm
  • Predict weighted avg of experts
  • Update weights multiplicatively based on performance (|pred - actual|)

• For best known worst-case guarantees
  • Predict $F(r = \text{weighted avg of experts})$ s.t.
    \[
    1 + \frac{\ln((1 - r)\beta + r)}{2 \ln(2/(1 + \beta))} \leq F_\beta(r) \leq \frac{-\ln(1 - r + r\beta)}{2 \ln(2/(1 + \beta))}
    \]
  • Update $w \leftarrow w \times U(q = |\text{pred} - \text{actual}|)$ s.t.
    \[
    \beta^q \leq U_\beta(q) \leq 1 - (1 - \beta)q.
    \]

• Q: Is optimizing worst-case sensible?
Machine Learning

• Variety of algorithms for different settings
• Experimented with:
  • Experts algorithms
  • Decision trees
  • Perceptron and Winnow algorithms
  • Support vector machines
  • Committees of the above
• Cross-validation experiments
• 0/1 error as well as quadratic loss
Exponential Gradient

• Update expert weights on every instance (game)

\[ w_i \leftarrow w_i (0.2 x_i \delta), \text{ where } \delta = y - p \]

• Allow multiple passes
• Used 3 passes in experiments
The “Variance” Algorithm

• Assume: For each game there is some true probability of occurrence
• Model each expert as a stationary Gaussian centered around that probability
• Each expert has its own variance
• Estimate the variances from games seen so far (EM)
Each Expert is Modeled as a Gaussian

\[ P_t = \frac{\sum_i w_i p^i_t}{\sum_i w_i}, \quad w_i = \frac{1}{\sigma_i}, \quad \sigma_i = \sqrt{\frac{\sum_t (p_t - p^i_t)^2}{T}} \]
Market Simulation

- Agents correspond to experts
- They buy/sell security: “$1 if Team A”
- Equilibrium where agents posteriors equal weighted avg of prior & market
- Form of wealth-weighted average [PW]
- YAEA (Yet Another Experts Algorithm)
- Arbitrarily bad worst-case bound
- Average case?
Experiments

- Cross Validation
- Online
  - Single year periods
  - Multi year periods
- US NFL Football
- US NCAA Basketball Playoffs
## Results: 0/1 Error

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home Team</td>
<td>0.4363</td>
<td>0.4479</td>
<td>0.412</td>
<td>0.3895</td>
</tr>
<tr>
<td>Top Expert</td>
<td>0.3514</td>
<td>0.3127</td>
<td>0.3521</td>
<td>0.3221</td>
</tr>
<tr>
<td>Average</td>
<td>0.3552</td>
<td>0.3436</td>
<td>0.3708</td>
<td>0.3109</td>
</tr>
<tr>
<td>SVM</td>
<td>0.3808</td>
<td>0.3571</td>
<td>0.3826</td>
<td>0.3179</td>
</tr>
</tbody>
</table>
## Results: Quadratic Score

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of experts</td>
<td>625</td>
<td>786</td>
<td>1257</td>
<td>1969</td>
<td>2231</td>
</tr>
<tr>
<td>Top Expert</td>
<td>3185</td>
<td>3445</td>
<td>3339</td>
<td>4218</td>
<td>3747</td>
</tr>
<tr>
<td>Average</td>
<td>2561</td>
<td>2574</td>
<td>2562</td>
<td>3298</td>
<td>3371</td>
</tr>
<tr>
<td>Average (30)</td>
<td>2864</td>
<td>2589</td>
<td>2529</td>
<td>3731</td>
<td>2986</td>
</tr>
<tr>
<td>Variance</td>
<td>2979</td>
<td>2660</td>
<td>2627</td>
<td>3498</td>
<td>3456</td>
</tr>
<tr>
<td>Variance (20)</td>
<td>3187</td>
<td>2662</td>
<td>2611</td>
<td>3881</td>
<td>3344</td>
</tr>
<tr>
<td>Experts</td>
<td>2801</td>
<td>2541</td>
<td>2406</td>
<td>3343</td>
<td>3099</td>
</tr>
<tr>
<td>Expert MD</td>
<td>2875</td>
<td>2644</td>
<td>2505</td>
<td>3442</td>
<td>3346</td>
</tr>
<tr>
<td>Exp Gradient</td>
<td>2827</td>
<td>2563</td>
<td>2616</td>
<td>3371</td>
<td>3137</td>
</tr>
<tr>
<td>Market Sim</td>
<td>3090</td>
<td>2482</td>
<td>2381</td>
<td>3397</td>
<td>3203</td>
</tr>
</tbody>
</table>
## Results: Multi-year

<table>
<thead>
<tr>
<th>Period</th>
<th>2000-3</th>
<th>2001-3</th>
<th>2002-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Expert</td>
<td>9910</td>
<td>8533</td>
<td>6782</td>
</tr>
<tr>
<td>Average</td>
<td>11169</td>
<td>8867</td>
<td>6221</td>
</tr>
<tr>
<td>Variance</td>
<td>11512</td>
<td>8774</td>
<td>6120</td>
</tr>
</tbody>
</table>
Results: NCAA

<table>
<thead>
<tr>
<th></th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Expert</td>
<td>2251</td>
<td>2401</td>
<td>2487</td>
</tr>
<tr>
<td>Average</td>
<td>1804</td>
<td>1607</td>
<td>1631</td>
</tr>
<tr>
<td>Experts</td>
<td>1827</td>
<td>1595</td>
<td>1643</td>
</tr>
<tr>
<td>Variance</td>
<td>1844</td>
<td>1706</td>
<td>1719</td>
</tr>
</tbody>
</table>
Conclusions

• Unique dataset useful for putting experts algorithms to the test
• Median expert bad, average very good; miscalibration, but calibrating not helpful
• Unweighted average hard to beat (experts, ML, MarketSim, wAvg fail)
• “Variance” slightly (but stat sig) better
• “Variance” does not update on performance!
• Many (most) experts are bad, yet averaging them is good; even averaging bad experts (Qscore < 0) does fairly well!
• “Wisdom of crowds”
Future Work

• Appropriate regularization of exponential gradient/regression?
• Understand and extend variance
• Compare to market (betting)
• ProbabilitySports now requires money to enter -> low participation
• Nearest neighbor method