Crowdsourced Data Analytics: A Case Study of Predictive Modeling Competition

Yukino Baba (National Institute of Informatics/JST, ERATO)
Nozomi Nori (Kyoto University)
Shigeru Saito (Opt, Inc.)
Hisashi Kashima (Kyoto University)
BACKGROUND

Predictive modeling competitions

● Predictive modeling
  - acquires latent patterns in data and exploits them to predict future data

● Finding the best suited predictive model is quite laborious for a single data scientist

● Predictive modeling competitions (e.g., kaggle)
  - allow us to leverage crowds of data scientist to examine a large number of models
  → Crowdsourced data analytics
MOTIVATION

Question:
Is crowdsourcing a promising approach to obtain good predictive models?

- Are participants in predictive competitions comparable to expert data scientist?
- Does prediction performance improve when we aggregate the models submitted by participants?
OVERVIEW

We conduct an experiment on a real platform for predictive modeling competitions

Our results show:

- Models submitted by about a half of participants outperformed those built by experts

- Aggregated model outperformed the winning model even if aggregating early submitted models
COMPETITION OUTLINE

● Platform: CrowdSolving
● Date: Aug. 14 – Sep. 15 (33 days)
● Num. of participants: 16
● Num. of submissions: 134
● Prize:
  - $500, $200, $150, $100 and $50 for first five place winners
● Task: Link prediction for Wikipedia articles
COMPETITION PROCEDURE

- Participants are asked to build a model and submit the predictions on test samples
- **Public score:** calculated using 50% of test samples
  - used to provide feedback to participants
- **Private score:** calculated using all of test samples
  - used to determine the winner
COMPETITION TASK

Link prediction for Wikipedia articles

- Task: predicting whether a hyperlink exists from one Wikipedia article to another

- Training data:
  - 45,209 pairs of articles having a link
  - 39,541-dimensional features of all the 23,269 articles

- Test data:
  - 78,426 pairs of articles
EXPERIMENT 1: Participant vs. expert
Method used by participants

Winners used simple heuristic techniques

- 1st place winner used Random Forest with heuristic features such as sum of feature vector of each article
- 2nd place winner designed his/her own link occurrence measure

<table>
<thead>
<tr>
<th>Place</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Supervised learning</td>
<td>Designing link measure</td>
<td>Supervised learning</td>
<td>Designing link measure</td>
<td>Designing link measure</td>
</tr>
</tbody>
</table>
EXPERIMENT 1: Participant vs. expert

Method used by expert

<table>
<thead>
<tr>
<th>Expert used a <strong>state-of-the-art</strong> method and <strong>general-purpose</strong> method</th>
</tr>
</thead>
<tbody>
<tr>
<td>● <strong>state-of-the-art method</strong> [Nori et al., AAAI’12]</td>
</tr>
<tr>
<td>- learns an embedding of features into a latent space that ensures the closeness between the features of linked articles</td>
</tr>
<tr>
<td>● <strong>general-purpose method</strong></td>
</tr>
<tr>
<td>- combines four common link measures and learns the weight of each measure</td>
</tr>
</tbody>
</table>
EXPERIMENT 1: Participant vs. expert

Results

<table>
<thead>
<tr>
<th>AUC</th>
<th>P01</th>
<th>P02</th>
<th>P03</th>
<th>P04</th>
<th>P05</th>
<th>P06</th>
<th>P07</th>
<th>P08</th>
<th>P09</th>
<th>P10</th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Submission date:
- Aug 19
- Aug 26
- Sep 02
- Sep 09
- Sep 16

Comparison between:
- NLDR
- LM
EXPERIMENT 1: Participant vs. expert

Results

General purpose model scored AUC=0.743

State-of-the-art model scored AUC=0.717

AUC

Submission date

Aug 19 Aug 26 Sep 02 Sep 09 Sep 16

P01 P02 P03 P04 P05 P06 P07 P08 P09 P10 P11 P12 P13 P14 P15 P16 NLDR LM

Scored AUC=0.743

Scored AUC=0.717
A participant’s model overtook the experts’ models on **4th day**.
EXPERIMENT 1: Participant vs. expert

Results

Winning model scored
AUC=0.946
EXPERIMENT 1: Participant vs. expert

Results

44% of participants created models that outperformed ones built by experts
EXPERIMENT 2: Aggregated vs. single model

Model aggregation procedure

We created an **aggregated model** from all submissions.

<table>
<thead>
<tr>
<th></th>
<th>Submission_1</th>
<th>Submission_2</th>
<th>Submission_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_1</td>
<td>Prediction_1_1</td>
<td>Prediction_1_2</td>
<td>Prediction_1_3</td>
</tr>
<tr>
<td>Sample_2</td>
<td>Prediction_2_1</td>
<td>Prediction_2_2</td>
<td>Prediction_2_3</td>
</tr>
</tbody>
</table>

Use submissions as features and train a new model (**stacking**).

Aggregated model
EXPERIMENT 2: Aggregated vs. single model

Results

<table>
<thead>
<tr>
<th>Submission date</th>
<th>AUC</th>
<th>Aggregated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 19</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Aug 26</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Sep 02</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>Sep 09</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Sep 16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The graph shows the AUC values for different submission dates, comparing aggregated and individual models.
EXPERIMENT 2: Aggregated vs. single model

Results

Aggregated model scored AUC=0.982 and overtook the winning model
EXPERIMENT 2: Aggregated vs. single model

Results

Aggregated model constructed from submissions in first 6 days outperformed the final winning model.
SUMMARY

- Results show the power of crowds for predictive modeling
  - **Quality**
    - A half of participants submitted the models that outperformed the models built by experts
    - Aggregation further improves the performance
  - **Speed:** Aggregated model built from early submissions overtook the winning model

- Model aggregation is one of Human Computation approaches for predictive modeling

→ Future investigation of such approaches is a promising research direction