Crowdsourced Data Analytics: A Case Study of Predictive Modeling Competition

Yukino Baba (NII/ JST, ERATO), Nozomi Nori (Kyoto Univ.)
Shigeru Saito (Opt, Inc.), Hisashi Kashima (Kyoto Univ.)

Overview Is crowdsourcing a promising approach to obtain good predictive models?
- Finding the best suited predictive model is 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.
- We conduct an experiment on a real predictive modeling competition platform.
- Our results show the power of crowds for predictive modeling in quality and speed.

Task: Link prediction for Wikipedia articles

- Training data: 45,209 pairs of articles having a link and 39,541-dimensional features of all the 23,269 articles
- Test data: 78,426 pairs of articles
- Competition period: 33 days
- Num. of participants: 16, Num. of submissions: 125

Model aggregation Train a new model by using submissions for features

<table>
<thead>
<tr>
<th>Training sample</th>
<th>Test sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample_1</td>
<td>Prediction_1_1</td>
</tr>
<tr>
<td>Sample_2</td>
<td>Prediction_2_1</td>
</tr>
</tbody>
</table>

Results

1. Models made by competition winners outperformed those built by experts.
2. Aggregated model outperformed the winning model even if aggregating early submissions.

- 44% of participants created models that outperformed the ones created by experts.
- Aggregated model scored AUC=0.982 and overtook the winning model (AUC=0.946).
- Aggregated model constructed from submissions in first 6 days outperformed the final winning model.

([AUC graph])