Crowdsourced Data Analytics: A Case Study of a Predictive Modeling Competition

Yukino Baba
National Institute of Informatics
JST, ERATO, Kawarabayashi Large Graph Project

Nozomi Nori
Kyoto University

Shigeru Saito
OPT, Inc.

Hisashi Kashima
Kyoto University

Abstract

Predictive modeling competitions provide a new data mining approach that leverages crowds of data scientists to examine a wide variety of predictive models and build the best performance model. In this paper, we report the results of a study conducted on CrowdSolving, a platform for predictive modeling competitions in Japan. We hosted a competition on a link prediction task and observed that (i) the prediction performance of the winner significantly outperformed that of a state-of-the-art method, (ii) the aggregated model constructed from all submitted models further improved the final performance, and (iii) the performance of the aggregated model built only from early submissions nevertheless overtook the final performance of the winner.

Introduction

Predictive modeling aims to acquire rules and relations in data and exploit them to predict future data. Various predictive modeling methods have been continuously developed to address the growing variety of data types and prediction tasks in data analytics research. However, in practice, when focusing on a particular application domain, the use of novel and sophisticated techniques is not always the best method to achieve the best predictive performance. Performance improvement is often brought about by the combination of a careful selection of models, feature engineering, and data sampling, most of which are driven by domain knowledge, experiences, and human inspiration. This fact is supported by the "no free lunch" theorem that tells us that there is no universally good method .

Although the broad exploration of models best suited to the data at hand is quite laborious for a single data scientist, the recent expansion of platforms for *predictive modeling competitions*, such as Kaggle, allows us to leverage crowds of data scientists to examine a large number of models. Using such competitions, a host can employ the power of crowdsourcing to obtain accurate prediction models on his or her own data. The participants of competitions build models and compete for monetary rewards. They are often allowed to submit their models multiple times during the competition period, and the prediction performance of each submission is fed back to them via a live leaderboard. Based on this feedback, the participants continue to improve their models. In other words, they explore the model spaces to find better models. Additionaly, the host of the competition can aggregates the models in order to obtain accurate models as early as possible during the course of a competition.

In this paper, we report the results of our experimental study of a predictive modeling competition. We hosted a competition on CrowdSolving, a predictive modeling competition platform in Japan, and collected submissions from its participants. We first compared the prediction accuracy of the models submitted by the participants with a model we built using a state-of-the-art method, and observed that the submitted models using simple heuristic techniques outperformed the state-of-the-art method. This result indicates the effectiveness of exploiting the power of crowds for the broad exploration of model spaces. Next, we applied a supervised learning technique to aggregate all the submitted models and compared the aggregated model with the individual ones. Our results show that the aggregated model further improved the final performance of the winner. Additionally, we observed that the performance of the aggregated model constructed only from submissions in the first several days nevertheless overtook the final performance of the winning model. These findings underline the fact that aggregating models derived by the power of crowds is an expeditious way to obtain a high quality model. The novelty of this study lies in the real experiment where a host of the competition aggregates the submitted models and the investigation of the power of crowds for predictive modeling, both in the quality of obtained models and its expedition to achieve it.

Case Study of a Predictive Modeling Competition

We hosted a competition on CrowdSolving from August 14th to September 15th, 2013, and received 134 prediction models made by 16 participants. We chose a link prediction problem on Wikipedia articles for our competition. We asked the participants to predict whether a hyperlink exists from one Wikipedia article to another. Specifically, the training dataset consisted of the $39,541$-dimensional attributes of all the $23,269$ articles and $45,209$ pairs of articles having a link between them. The participants were requested

to build a model predicting whether links among the other $78,426$ pairs in the test dataset existed or not, and to submit estimated probabilities for each link. We employed the Area Under the ROC Curve (AUC) to evaluate the prediction performance of each submission.

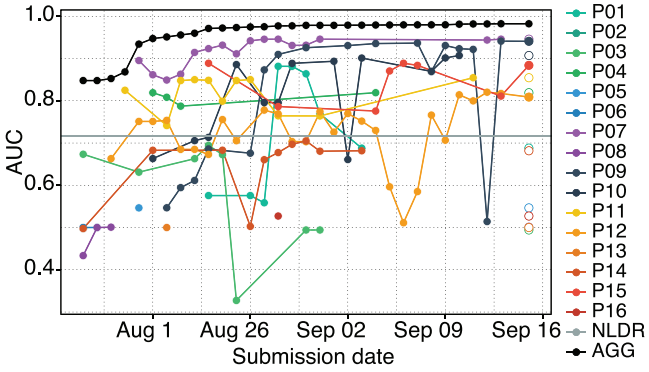

Figure 1: Time series comparing of the predictive performance of submitted models during the competition period, model built using the state-of-the-art method, and aggregated model. Data labeled P01 to P16 indicate the AUC of the model submitted by each participant on each day. White circles represent the AUC of the models used to determine the final ranking (those for the participants who submitted their model on the final day are omitted from the plot). Labels NLDR and AGG represent the state-of-the-art method and the aggregated model, respectively.

Results

Submitted Models vs. State-of-the-art Method

We employed a link prediction method using nonlinear dimension reduction (Nori, Bollegala, and Kashima 2012) as a representative of the state-of-the-art link prediction methods. A performance comparison of the submitted models and a model built using the state-of-the-art method (NLDR) is shown in Figure 1. The state-of-the-art method scored an AUC of 0.7166 and exceeded the submitted models on the first three days. However, P11 overtook them on the fourth day, and about half of the submitted models surpassed them by the end of the competition. Although the first three winners adopted common predictive models and heuristic feature engineering techniques, the models created by them achieved a score of over 0.90; such a substantial increase in the prediction performance might not be achievable by improvements to a single predictive model alone. The combination of a common model and careful feature engineering that captures characteristics of the data achieves the high prediction performances. This fact notably illustrates the benefits of predictive modeling competitions that leverages the power of crowd data scientists.

Submitted Models vs. Aggregated Model

We compared the submitted models with the aggregated model that was constructed by an ensemble learning technique called *stacking* that uses the outputs of classifiers as features for another classifier (Wolpert 1992). For the aggregation classifier, we adopted the gradient tree boosting (Friedman 2001) that uses decision trees as its weak learner. Note that the aggregated model used some of the ground truth labels of the test dataset in addition to the originally provided training data, and the predictions by the submitted models for the test samples were used as the features for model aggregation.

Figure 1 compares each submitted model and the aggregated model. On each day, the aggregated model was trained using all of the models submitted up to that day. The prediction performance of the best individual model submitted on the first day was 0.6735, while the aggregated model had already achieved an AUC of 0.8479, almost equivalent to the best individual performance on the fifth day. At the end of the competition, the final performance of the aggregated model was 0.9823, significantly outperforming the AUC of 0.9459 achieved by the winning model. It is not surprising that the performance of the aggregated model improves as the performance of the individual models improve. However, it is worth noting that the aggregated model of the first day achieved such a high accuracy, even though it was based on only five models. Moreover, the aggregated model constructed only from the submissions of the first eight days overtook the performance of the winning model that took 33 days to build. This result indicates the power of crowds for predictive modeling not only in the quality of the obtained model, but in its expedition to achieve it.

Conclusion

To examine the effectiveness of crowdsourcing competitions for predictive modeling, we hosted a competition and conducted an experiment. While most of the existing studies on competition-style crowdsourcing focus on the theory of incentive design for active participation (Abernethy and Frongillo 2011), the present study experimentally investigated the power of data scientist crowds for fast and high-quality predictive modeling as well as the possibility of combining them with machine learning for further improvements.

Acknowledgments

The authors would like to thank Kazuki Fukue (Infocom Corporation) for his support in the experiments.

References

Abernethy, J., and Frongillo, R. M. 2011. A collaborative mechanism for crowdsourcing prediction problems. In *Advances in Neural Information Processing 24*, 2600–2608.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 1189–1232.

Nori, N.; Bollegala, D.; and Kashima, H. 2012. Multinomial relation prediction in social data: A dimension reduction approach. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, 115–121.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.