Crowdsourcing: can workers compete with experts in generating predictors?

Jing Song, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara *
(Graduate School of Information Science and Technology, Hokkaido University)†

1 Introduction

Large companies in developed countries usually send their jobs to India and China since there are large and cheap labor forces there.[1] Based on the same idea, people can outsource Human Intelligence work (HITs) to the "crowd" using online platforms, such as Amazon Mechanical Turk (AMT). The task is divided into some micro-tasks and the users can earn money by finishing these micro-tasks. The process of crowdsourcing is usually fast and of low cost.

One type of human intelligence which is difficult for computers to realize is that human can predict or decide the causation of specified phenomenon. Even though computers can find out the correlation coefficients between two items, they cannot interpret how the relationship works between them. For example, a computer can find out that the average temperature of Singapore, which is 23~31 degree Centigrade, relates to its popularity as a resort, but cannot decide whether the average temperature contributes to its popularity as a resort or its popularity contributes to its weather. However, for human beings, it is apparent that the popularity as a resort cannot affect its local temperature.

In this paper, we propose to use crowdsourcing as a way to generate predicators instead of experts to make use of human intelligence in inferring causation. To control the quality of generated predicators, we propose to hire another group of micro-task workers to give a score to the generated predictors. In the two-way process, we want to find out whether crowdsourcing could be used as a common platform to generate predictors instead of experts and also whether crowdsourcing could be used as a way to give evaluation of HITs.

2 Related work

Bevelander[3] has proposed a method that use crowdsourcing to get childhood predictors of adult obesity. He made users answer questions on a website designed for the research. Besides of answering questions, the users can put forward new questions freely. The demerit of the setting of the process he argues is that users will not answer questions proposed later, which makes some features of the users cannot be obtained.

Paulheim[2] use DBpedia and two other dataset to generate interpretations for the ranking of quality of living and different countries' corruption perception. Since there is almost no human-interruption in the whole process, some generated interpretations are not acceptable for people.

Baba and Kashima[4] proposed to use two stages, the creation stage and the review stage, to control the quality of general crowdsourcing tasks. They carried on research about 3 tasks: logo designing, image description and language translation. The experiment result shows that the two-stage process works well compared with recent well-known statistical methods.

3 Crowdsourcing procedure

We want to use crowdsourcing to generate predictors about the score people will get in an English exam, e.g. TOEIC. Crowdsourcing workers should propose questions which they think are related to the final score they get in the exam. After gathering the questions, we will divide the generated questions into several batches, e.g. 10 questions in one batch, and hire other crowdsourcing workers to give an evaluation of them. We should note that a worker may give score to only one or several batches of questions, not the whole ones. After gathering the score, we want to use statistical methods to analyze the true value of every question and give them a rank. In the last step, we will hire workers to tell us the grade he or she got in the TOEIC exam and answer the questions we have gathered. We expect the workers answer all the questions which we have gathered in the step before. Finally, we want to use regression methods to analyze the correlation coefficients between the predictors and the target exam grade to see whether top ranked predictors work better than the low ranked ones. The whole procedure is illustrated in Figure 1. There are three crowdsourcing tasks in the whole process: generating predictors, evaluating predictors, answering questions.

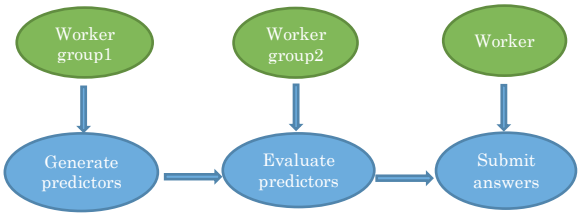

Figure 1. Flow chart of the crowdsourcing procedure

3.1 Task 1: Generating Predictors

We will use online crowdsourcing platforms to gather predictors for the score people can get in the TOEIC exam. In this process, we will not control the process users propose questions unless there are some offensive or inappropriate words.[3] We hope to get various kinds of predictors and find something interesting and new. We decide to gather predictors using the online crowdsourcing platforms such as Lancers for about two weeks.

3.2 Task 2: Evaluating Generated Predictors

We will hire different group (Group 2) of workers to evaluate those questions we have gathered, in other words, using crowdsourcing as a way to evaluate crowdsourcing results. We set the full grade as 5, and let the crowdsourcing workers to give a score to each proposed question. After gathering the score, we will use statistical methods to give the generated predictors a rank.

3.3 Task 3: Submitting answers

Considering the problem in Kirsten's experiment, we expect the workers answer as many questions as possible in order to get enough features for every worker. Besides, we need the worker to tell us how much score they got in the TOEIC exam. If they have taken the exam for many times, we expect the last score they got.

After gathering the answers of different workers, we want to use regression methods to map the features of each worker to the target value to see whether the proposed questions can predict the target value effectively. Besides, we want to compare the predicted result using different groups of predictors based on the rank we calculated in last step to see whether the crowdsourcing evaluation process works.

Through the experiment, we hope to find some effective predictors for the grade people will get in the TOEIC exam. We hope the experiment result could give students some direction on English learning.

4 Preliminary Research

To test whether it is feasible to carry on this research, we did a pre-experiment, in which we prepared 6 questions by ourselves and asked 8 people to answer them. We listed a few questions blow:

1. Do you watch TV play by native English speakers?
2. Do you want to learn another foreign language except English?
3. To learn English, have you attended cram school?

From the answers we gathered from 8 people, we found that question 2 almost have no value in predicting the score in TOEIC because only one student wants to learn another foreign language. Besides, the students who give most questions negative answers tend to have low grade in the exam. From the preliminary experiment, we see two points which we think are worthy to carry on further research. First, whether common people can give predictors like experts. Secondly, whether the evaluation given by the crowd can work.

5 Conclusion

In this pater, we propose to use crowdsourcing as a way to generate predictors and also use it as a way to evaluate crowdsourcing work. We want to see whether the generating stage and the evaluation stage could work well in low cost.

Referencing

[1] Howe J. The rise of crowdsourcing. *Wired magazine*, 2006, 14(6): 1-4.

[2] Paulheim H. Generating possible interpretations for statistics from linked open data./ *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, 2012: 560-574.

[3] Bevelander K E, Kaipainen K, Swain R, et al. Crowdsourcing Novel Childhood Predictors of Adult Obesity. *PloS one*, 2014, 9(2): e87756.

[4] Baba Y, Kashima H. Statistical quality estimation for general crowdsourcing tasks. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013: 554-562.

*{jingsong, oyama, haru, kurihara}@complex.ist.hokudai.ac.jp