Improving Reproducibility of Crowdsourcing Experiments

Kohta Katsuno
University of Tsukuba
kohta.katsuno.2019b@mlab.info

Masaki Matsubara
University of Tsukuba
masaki@slis.tsukuba.ac.jp

Chiemi Watanabe
Tsukuba University of Technology
chiemi@a.tsukuba-tech.ac.jp

Atsuyuki Morishima
University of Tsukuba
mori@slis.tsukuba.ac.jp

Abstract

This paper reports our preliminary investigation into which kinds of metadata are effective in improving the reproducibility of crowdsourcing experiments. We examined the effectiveness of using an ability distribution of workers, regarding their solving of tasks. One issue, however, is that identifying the ability required to solve a set of tasks can be difficult. We applied *item response theory* (IRT) to find said undefined ability, allowing us to calculate the ability distribution of the workers. In our preliminary experiment, the standard deviation of the results quality was reduced to 60%. This implies that we can reduce the number repetitions of the experiment to 32.9%, while still maintaining the same power level.

Introduction

The reproducibility of crowdsourcing experiments is one of the most pressing concerns in the field of the science of crowdsourcing,(Jiang and Wang 2016),(Paritosh 2012). One of the factors that has obstructed the improvement of reproducibility is that the set of people in the crowd are unknown. Two attempts of the same experiment could produce different results if the different sets of workers joined, because of the inherent differences of people in general (Daniel et al. 2018). Therefore, in order to persuade reviewers and readers that the results of crowdsourcing experiments are trustworthy, researchers in this field often repeat the same experiment many times, and therefore argue that the difference is statistically significant. Reproducing experimental results through this process is not easy, however, in terms of both the time taken and the monetary cost.

In this study, our approach is to add metadata concerning the kinds of workers that joined the experiment to the experiment description. This would then allow others to replicate the experiment with a similar set of workers. As a first step, we examine the effectiveness of using an ability distribution of workers for the solving of a given task (Figure 1).

However, it can often be difficult to identify the ability distribution required to solve a set of tasks. The workers' ability to understand written sentences of day-to-day conversations in English, for example, would be an English language skill. However, the workers' ability to solve tasks in general depends on the given task, and there is no explicit answer, unless it is a well-known problem, such as the above example.

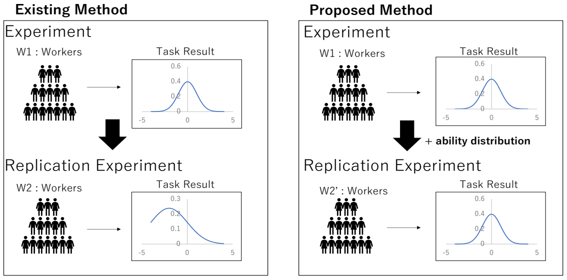

Figure 1: In this paper, we aim to improve the reproducibility of crowdsourcing experiments by focusing on the ability distribution of the workers.

Here we apply *item response theory* (IRT)(Baker 2004) to find this unknown ability, and therefore to show the ability distribution of the workers. IRT has been widely used in the context of standardized tests such as the Test of English as a Foreign Language (TOEFL). IRT is used to identify an unnamed factor, θ, which represents the worker's ability to obtain the high score of the test. IRT can not only find the ability of workers but also make good tests to measure workers' ability.

Here, we attempt to determine whether an ability distribution derived by IRT can improve the reproducibility of crowdsourcing experiments. In our experiment, the standard deviation (SD) of the results of our task was reduced to 60%, meaning that we could reduce the number repetitions to 32.9%, while still maintaining the same power level.

Preliminaries

Let θ be an *ability parameter* that represents the ability of each worker. Then, the probability of correctly answering task j (called an item in IRT) in the 2-parameter logistic model is expressed by the following equation:

$$P_j(\theta) = \frac{1}{1 + exp(-Da_j(\theta - b_j))}, -\infty < \theta < \infty,$$

where a_j and b_j are parameters representing the features

Presented in the Work in Progress and Demo track, HCOMP 2019.

of each item, called “item discrimination” and “item difficulty”, respectively. The a_j, b_j and θ are computed by IRT. The item discrimination, a_j, is a parameter that represents the power to distinguish workers with and without the ability. The item difficulty, b_j, is a parameter that represents how high the difficulty of the item is. D is called the “scale factor”, and is generally set to 1.7 (Baker 2004).

Method

We propose a method to improve the reproducibility of crowdsourcing experiments, based on the workers' ability distribution (Figure 1). In existing methods, the workers ($W1$) who participated in the experiment and the workers ($W2$) who participated in the re-experiment are different sets in terms of θ. The task results will likely therefore have different distributions, making it difficult to replicate the re-experiment. In our proposed method, however, the experimenter can report both the task results and the workers' ability distribution as metadata. When another researcher attempts to reproduce the experiment, it would therefore be possible to improve the reproducibility of the task results by selecting a worker $W2' \subseteq W2$ that has a distribution similar to $W1$, by measuring their ability using a test.

We construct our test as follows: First, we compute the parameters (a_j, b_j). Then, we choose *good* tasks for the test from all the tasks, based on the parameters. We choose tasks with higher a_j values, whereas b_j are scattered so that the test scores are proportional to θ values of the workers.

Preliminary Experiment

We conducted a preliminary experiment to determine the effect of the workers' ability distribution on the reproducibility of the experiment. To simulate the replication experiments, we generate $W2$ and $W2'$ by computing subsets of $W1$ in different ways.

Dataset. In this study, we used a set of microtask results obtained from real-world workers (details are in (Kobayashi et al. 2018)). The task is a 4-choice task, in which the 84 workers are asked to identify the painter of 96 pictures.

Procedure. (1) We used IRT to find a_j and b_j for each task, and selected the top 12 tasks in terms of their a_j values. Since the variance of b_j of all tasks is small, we included all the questions used the 12 tasks in the qualification test. (2) We computed the ability θ of all workers from test results. (3) We created two subsets from the 84 workers 10,000 times. [`Selected`] Twenty workers were extracted from workers in such a way as to have the same distribution; the deviation should be within 0.01 from the mean value of $W1$, and 0.05 from the SD of its ability distribution. [`Random`] Twenty workers were randomly extracted independent of the θ value. (4) For each pair of worker sets (`Random` and `Selected`), we computed the task results, and computed the difference from the original result produced by $W1$ in terms of the average of the accuracy. Then we compute SD of the value. (5) According to interval estimations from the SD values, we compared the sample sizes required to reproduce the experiment.

Table 1: Result of Standard Deviation (SD)

	`Selected`	`Random`
SD	0.57	1.00

Result. The result of our preliminary experiment is shown in Table 1. The SD for the `Selected` and `Random` groups are 0.57 and 1.00, respectively.

Assuming that the result of the original experiment is reliable, in the sense that it is close enough to be that with appropriate statistical population, the sample size required for the result of another experiment to be close enough to the original one can be computed, if we have an acceptable error, SD and a confidence level. Given a sample size n, a standard deviation σ, a confidence level 95%, and an error δ, then $1.96 \times \frac{\sigma}{\sqrt{n}} = \delta$ holds for the interval estimation:

Then, the ratio of sample size is (δ disappears):

$$ratio = \frac{n_{selected}}{n_{random}} = (\frac{\sigma_{selected}}{\sigma_{random}})^2$$

As a result, the sample size required in the `Selected` group is about 32.9% of that of the `Random` group. Thus, we can reduce the number of tasks (and thus the number of attempts) into 1/3, compared to the experiment with workers in the Random group, which implies that we need only 1/3 monetary and time cost for the replicate experiment.

Discussion & Limitation

Table 1 shows an improvement in reproducibility, because the SD of the `Selected` group is lower than that of the `Random` group. This result shows that using the workers' ability distribution as metadata of crowdsouricng experiments, obtained via IRT, is a promising approach to improve reproducibility.

Our method is effective when the difference among all workers in their results is large. In that sense, the task we used in the preliminary experiment is not favorable for our method, since the difference was not large (SD in the accuracy was only 0.09). We will apply the method to various tasks in the future.

IRT assumes all the tasks to be performed by the same set of workers.Therefore, in the future, we are considering applying collaborative filtering as a method of filling missing values.

Conclusion

We found that the reproducibility of crowdsourcing experiments can be improved by determining the workers' ability distributions. Since our method reduced the sample size required to replicate the result into 1/3 even in a case that is not favorable to the method, we believe that this approach is promising, although there are a lot of research issues to increase the applicability of the proposed method.

Acknowledgments

This work was partially supported by JST CREST Grant Number JPMJCR16E3 including AIP challenge program and JSPS KAKENHI Grant Number JP19K11978, Japan.

References

Baker, F. (Ed.), K. S. E. 2004. *Item Response Theory: Parameter Estimation Techniques*. Boca Raton: CRC Press.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.* 51(1):7:1–7:40.

Jiang, R., and Wang, J. 2016. Reprowd: Crowdsourced data processing made reproducible. *CoRR* abs/1609.00791.

Kobayashi, M.; Morita, H.; Matsubara, M.; Shimizu, N.; and Morishima, A. 2018. An empirical study on short-and long-term effects of self-correction in crowdsourced microtasks. In *AAAI HCOMP 2018*.

Paritosh, P. 2012. Human computation must be reproducible. In *WWW 2012, Lyon.*