Spammer Detection Based on Task Completion Time Variation in Crowdsourcing

Ayato Watanabe
Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501 Japan
Email: watanabe.a@dl.soc.i.kyoto-u.ac.jp

Keishi Tajima
Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501 Japan
Email: tajima@i.kyoto-u.ac.jp

***Abstract*—In many existing spammer detection methods, a worker choosing answers independently from the true answers is regarded as a spammer. These methods may regard a diligent but low-skilled worker as a spammer. Our method uses workers' task completion time instead of answers. If it is independent from the difficulty of the tasks, we regard the worker as a spammer. Our experimental result suggests that this approach is potentially useful, and the selection of tasks seems a key for success.**

***Index Terms*—human computation, worker quality**

I. Introduction

Spammer detection is one of the most important issues in crowdsourcing. In many existing spammer detection methods, a worker choosing answers independently from the true answers is regarded as a spammer. In this approach, workers whose answers are not useful for the requester are regarded as spammers, and it is reasonable if we only think of the benefit of the requester. However, if the worker is diligent but the answers are not useful because of the lack of the skill or knowledge, it may not be ethically desirable to determine such a worker as a spammer and refuse to pay the reward.

We propose a new approach that can avoid it. We want to reward the effort taken by the workers even if the answers are not accurate. However, a simple method that pays reward based on how much time spent by workers is vulnerable to cheating. Instead of simply using the length of task completion time, our method uses its variation. If a worker spent more time for difficult tasks and less time for easy tasks, our method regards the worker as diligent. In other words, our method examines whether task completion time is independent from the difficulty of tasks, while many existing methods examine whether task answers are independent from the true answers.

II. Related Work

In this section, we explain three types of related work: the standard approaches to spammer detection in crowdsourcing, spammer detection based on task completion time, and analysis of task completion time for other purposes. We only focus on spammer detection for classification tasks.

A. Standard Approaches to Spammer Detection

The simplest method for spammer detection in crowdsourcing is to inject some items with known ground truth, and reject workers who chose wrong answers for many of them.

When we do not have items with known ground truth, we can adopt a classical method proposed by Dawid and Skene [1], which jointly infer the true answers and the quality of workers by using the EM algorithm. In their method, workers whose answers are consistent with the majority of the workers are regarded as good workers. Consistency does not necessarily mean a good match with the majority. For example, a worker always answering the opposite to the majority in a binary classification is consistent. Based on this method, Ipeirotis et al. [2] defined the *cost* of each worker, which can be used as the metric to detect spammers.

The methods above uses consistency of workers' answer with the ground truth or estimated truth. On the other hand, Raykar and Yu [3] proposed a method of calculating the spammer score of a worker based on whether how much the worker's answers depend on the true answer. If the probability distribution of the answers of a worker is independent from the true answers, the worker is a spammer choosing the answers without looking at the given data.

These methods, which use the answers of the workers, may reject diligent workers without enough skill. Our method avoid it by focusing on the time the worker spent for tasks.

B. Spammer Detection using Task Completion Time

Kazai et al. [4] has proposed a method for classifying crowd workers into several types based on their behavioral characteristics including task completion time. However, they only use the average completion time of the worker, and their experiment shows that the average completion time is not very useful for detecting spammers.

Chen [5] has shown that the average task completion time of spammers is shorter than that of the ordinary workers, and its variation is also small. That is, spammers take short constant time for every task. However, their spammer detection method only uses the deviation of the average task completion time of a worker from the average over all the workers.

C. Other Analysis of Task Completion Time

Cheng et al. [6] proposed a method of measuring task difficulty by the time needed for achieving various error rates. Yang et al. [7] asked the workers who have completed tasks to report the complexity of each task. The result shows that the complexity of a task perceived by workers show good agreement. Based on these studies, we expect that task completion time has a strong correlation with task complexity, and task complexity perceived by different users are quite consistent, unless a worker is a spammer.

III. Proposed Method

As explained in Sec. I, our method examines whether task completion time of a worker is independent from the difficulty of the tasks. We use Pearson correlation coefficient for measuring the dependency.

Because we cannot know the difficulty of a task, we approximate it by the time spent by the other workers for the task. That is, we calculate correlation between task completion time of a worker and that of the other workers. If it has a positive correlation, we regard the worker as a diligent worker.

IV. Experiments

We conducted a preliminary experiment for validating our approach. We report its result in this section.

A. Data Set

We collected data by posting a image classification task on Amazon Mechanical Turk. Workers were asked to classify 70 images into the following seven categories: Samoyed, German Shepherd, Siberian Husky, Alaskan Malamute, Gray Wolf, Coyote, Dhole. The 70 images include 10 images in each category. Images are shown to the user one by one: a image is shown after the worker has chosen the answer for the previous image. We recorded the task completion time for each image. The order of images were shuffled after every 18 workers. Radio buttons for seven categories were shown in the alphabetical order. By removing workers who did not completed all 70 tasks, we collected data of 199 workers.

Fig. 1 shows the accuracy of answers by 199 workers for each image. Each box plot shows the distribution of accuracy for 10 images in each category. We can see that Samoyed is the easiest category, followed by German Shepherd.

B. Evaluation

We compute the spammer score by the method proposed by Raykar and Yu [3], sort the 199 workers by it, regard worst workers as spammers, and examine the correlation between task completion time of good workers and spammers.

C. Result

We first show the result of the method that uses the task completion time for 10 images in each category. Table I shows the average correlation between a pair of workers for each category. We can summarize these seven values as follows:

$$\text{Dhole} \gg \text{Samoyed} \approx \text{Shepherd} > \text{Husky} > \text{Wolf} \\ > \text{Malamute} \gg \text{Coyote}$$

Fig. 2 shows heat maps representing Pearson correlation coefficient for all worker pairs. The x-axis and y-axis are sorted by the spammer score.

In the heat map for Samoyed, which is the easiest category, blue, yellow, and red colors are distributed randomly. This is probably because Samoyed images are very easy to distinguish, without much variance in the difficulty, and both good workers and spammers take very short time for all 10 images.

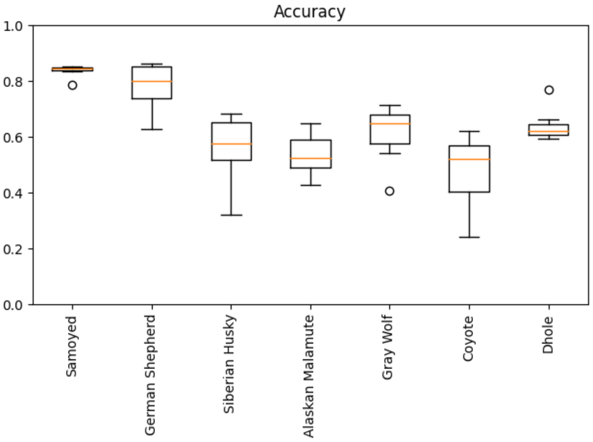


Fig. 1. Distribution of accuracy (proportion of correct answers in answers by 199 workers) for 10 images in each category.

TABLE I
Average Correlation between Workers for Each Category

Category	Average Worker Correlation
Samoyed	0.0670
German Shepherd	0.0643
Siberian Husky	0.0412
Alaskan Malamute	0.0278
Gray Wolf	0.0311
Coyote	0.0087
Dhole	0.1794

On the other hand, in the heat map for German Shepherd, we can see more green dots in the top-left quarter, which means good workers are more likely to have positive correlation. It may be because German Shepherd images are easy to distinguish but have a larger variance in difficulty than Samoyed, and diligent workers take more time for less easy ones.

The heat map for Dhole shows higher correlation. We found out that this is because there was a huge image in this category, and it took time for every worker to load this image through the network. This example shows that we need to carefully eliminate these undesirable factors when we adopt spammer detection methods based on task completion time.

In the heat map for Dhole, there are several red/yellow belts and they are more likely to appear near the bottom and right edges. It suggests that some spammers are less likely to have positive correlation with others.

The other four heat maps (Husky, Malamute, Wolf, Coyote) are more red. Colors are randomly distributed as in the heat map for Samoyed, and we cannot find clear patterns.

These observations suggest that correlation of task completion time is potentially useful, but how to use it is not straightforward, and selection of tasks may be a key to success.

Based on this hypothesis, we chose 5 images with the highest accuracy (the proportion of correct answers in answers

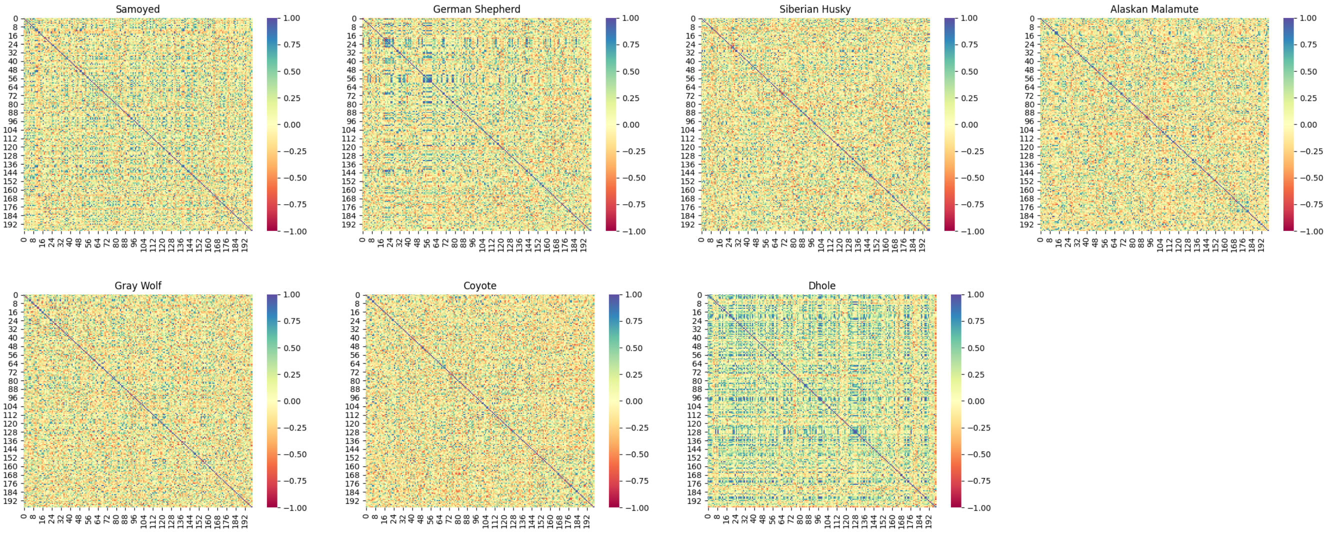


Fig. 2. Pearson correlation coefficient between workers based on their task completion time for 10 images in each category.

by 199 workers) and 5 images with the lowest accuracy, and calculate the correlation of the task completion time for those 10 images. Fig. 3 (a) shows the heat map of the correlation between worker pairs. The average correlation between a pair of workers was 0.0439. Similarly, we chose 5 images with the largest variance of task completion time by 199 workers (excluding the large Dhole image explained before) and 5 images with the smallest variance. Fig. 3 (b) shows the heat map. The average correlation was 0.1088.

Because task difficulty has correlation with image category, we also calculate a worker's average task completion time for 10 images in each category (thus, we obtain 7 values for each worker). Similarly, we calculate a worker's average task completion time for images classified into each category by the worker. Fig. 3 (c) and (d) shows the heat maps for them, and the average correlation was 0.0354 and 0.0614, respectively.

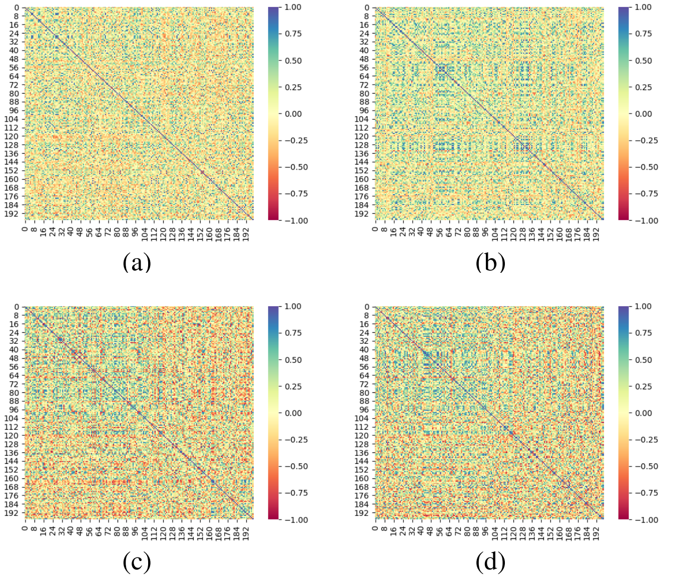
Against our expectations, we cannot make spammers distinctive in these heat maps. However, compared with the heat maps for Husky, Malamute, Wolf, and Coyote, the top-left quarter is more green and the areas near the bottom-left and top-right corners are more red. This result validates our hypothesis that task selection is a key, but we need to improve it further in order to make our method effective.

V. Conclusion

In this paper, we propose to use the correlation between task completion time of a pair of workers in order to detect spammers. The result of our experiment suggests that it is potentially useful, but how to utilize it is not straightforward, and the selection of the tasks may be one of the keys to success. We will investigate what tasks we should choose for spammer detection based on task completion time.

Acknowledgment

This work was supported by JST CREST Grant Number JPMJCR16E3, Japan.



(a)

(b)

(c)

(d)

Fig. 3. Correlation based on: (a) 5 most and 5 least difficult images, (b) 5 with the largest variance of task completion time and 5 with the smallest, (c) average time for each category, and (d) average time for each label.

References

[1] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[2] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *SIGKDD HCOMP workshop*, 2010, pp. 64–67.

[3] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *JMLR*, vol.13, no.1, pp. 491–518, 2012.

[4] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *CIKM*, 2011, p. 1941–1944.

[5] X. Chen, "A real time anti-spamming system in crowdsourcing platform," in *ICSESS*, 2016, pp. 981–984.

[6] J. Cheng, J. Teevan, and M. S. Bernstein, "Measuring crowdsourcing effort with error-time curves," in *CHI*, 2015, p. 1365–1374.

[7] J. Yang, J. Redi, G. Demartini, and A. Bozzon, "Modeling task complexity in crowdsourcing," *AAAI HComp*, pp. 249–258, 2016.