Expert Systems with Applications 41 (2014) 2678–2687

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces ☆

Yukino Baba [a,*], Hisashi Kashima [a], Kei Kinoshita [b], Goushi Yamaguchi [b], Yosuke Akiyoshi [b]

[a] *The University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656, Japan*
[b] *Lancers Inc., 3-10-13 Shibuya Shibuya-ku, Tokyo 150-0002, Japan*

ARTICLE INFO

Keywords:
Crowdsourcing
Machine learning
Improper task detection
Quality control

ABSTRACT

Controlling the quality of tasks, i.e., propriety of posted jobs, is a major challenge in crowdsourcing marketplaces. Most existing crowdsourcing services prohibit requesters from posting illegal or objectionable tasks. Operators in marketplaces have to monitor tasks continuously to find such improper ones; however, it is very expensive to manually investigate each task. In this paper, we present the results of our trial study on automatic detection of improper tasks to support the monitoring of activities by marketplace operators. We performed experiments using real task data from a commercial crowdsourcing marketplace and showed that the classifier trained by the operators' judgments achieves a high performance in detecting improper tasks. By analyzing the estimated classifier, we observed several effective features for detecting improper tasks, such as the words appeared in the task information, the amount of money that each worker will receive for the task, and the type of worker qualification option set for a task.

In addition, to reduce the annotation costs of the operators and improve classification performance, we considered the use of crowdsourcing for task annotation. We hired a group of crowdsourcing (non-expert) workers to monitor posted tasks and use their judgments to train the classifier. We were able to confirm that applying quality control techniques is beneficial for handling the variability in worker reliability and that it improved the performance of the classifier.

Finally, our results showed that the use of non-expert judgments of crowdsourcing workers in combination with expert judgments improves the performance of detecting improper crowdsourcing tasks, and that the use of crowdsourced labels allows a reduction in the required number of expert judgments by 25% while maintaining the level of detection performance.

1. Introduction

Crowdsourcing platforms provide online marketplaces for outsourcing various kinds of tasks to a large group of people. With the recent expansion of crowdsourcing platforms such as Amazon Mechanical Turk (MTurk)[1] and CrowdFlower,[2] the concept of crowdsourcing has been successfully applied in various areas of computer science research, including computer vision (Sorokin and Forsyth, 2008) and natural language processing (Snow et al., 2008). Business organizations also make use of crowdsourcing for processing large numbers of tedious tasks such as transcription and product categorization (Ipeirotis, 2010a).

One of the biggest challenges in crowdsourcing is ensuring the quality of the results submitted by crowdsourcing workers, because there is no guarantee that all workers are sufficiently competent needed to complete the offered tasks. Some faithless workers also try to get paid as easily as possible by submitting low-quality responses. Several approaches geared toward efficient quality control have been applied; for example, MTurk provides a pre-qualification system to assess the skill level of a prospective worker, and CrowdFlower enables requesters to inject gold standard data, i.e., a collection of tasks with known correct answers, into their tasks to measure a worker's performance automatically. Another promising approach is to introduce redundancy, which involves asking multiple workers to work on each task, and then aggregating their results to obtain a more reliable result by applying majority voting or other sophisticated statistical techniques (Dawid and Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010). In addition to

☆ Part of this work has been published in preliminary form in the proceedings of the Twenty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-13), 2013.

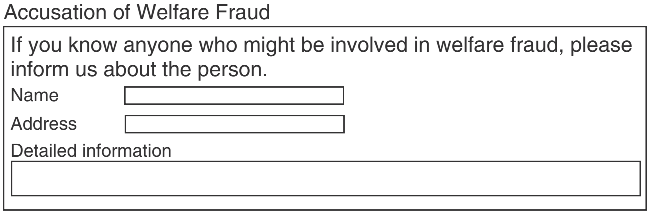
* Corresponding author.

E-mail addresses: yukino_baba@mist.i.u-tokyo.ac.jp (Y. Baba), kashima@mist.i.u-tokyo.ac.jp (H. Kashima), kinoshita.kei@lancers.co.jp (K. Kinoshita), yamaguchi.goushi@lancers.co.jp (G. Yamaguchi), akiyoshi.yosuke@lancers.co.jp (Y. Akiyoshi).

[1] https://www.mturk.com.

[2] http://crowdflower.com.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.11.011



Accusation of Welfare Fraud

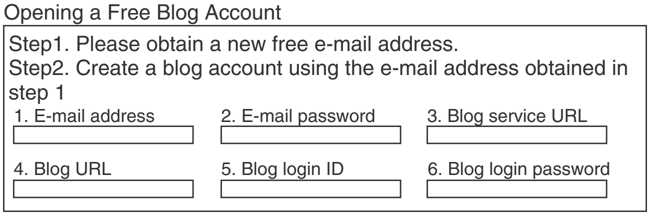
If you know anyone who might be involved in welfare fraud, please inform us about the person.

Name

Address

Detailed information

Fig. 1. Example of an improper task (requiring disclosure of another person's identity).



Opening a Free Blog Account

Step1. Please obtain a new free e-mail address.
Step2. Create a blog account using the e-mail address obtained in step 1

1. E-mail address
2. E-mail password
3. Blog service URL
4. Blog URL
5. Blog login ID
6. Blog login password

Fig. 2. Example of an improper task (requiring registration at another web service).

worker performance, it was pointed out that badly designed tasks may carry the low quality results in crowdsourcing (Kittur et al., 2013).

Controlling the *quality of tasks*, i.e., propriety of posted jobs, is another great challenge in crowdsourcing marketplaces. For maintaining the safety and integrity of marketplaces, their administrators in the marketplaces have to prevent requesters from posting illegal or objectionable tasks, and remove improper tasks immediately to prevent workers from working on them. Most existing crowdsourcing marketplaces prohibit specific kinds of tasks, for example, those entailing illegal or antisocial activities, those collecting personally identifiable information of workers (Fig. 1), or those requiring workers to register for a particular service (Fig. 2). Several existing studies have explored the popularity and characteristics of such improper tasks in crowdsourcing marketplaces (Motoyama et al., 2011; Wang et al., 2012). For instance, a study in 2012 showed that 12% of tasks on MTurk were related to fake account registration or fake posting on social networking services (Wang et al., 2012). Operators in crowdsourcing marketplaces have to monitor the tasks continuously to find such improper ones; however, the manual investigation of each task is very expensive.

In this paper, we present the results of our trial study on the automatic detection of improper tasks, which we conducted on Lancers,[3] a popular crowdsourcing marketplace in Japan. To support the monitoring of activities by marketplace operators, we use machine learning. Our proposed procedure to reduce the monitoring burden comprises three steps. (1) Operators annotate a portion of tasks to indicate whether each task is proper, and train a classifier by applying a supervised machine learning method to the annotated tasks. (2) When a new task is posted, the trained classifier determines whether the task is proper. (3) If the classifier finds potentially improper tasks, the monitoring system reports them to operators for a manual judgment.

To reduce the annotation costs further and improve the classification performance, we consider the use of crowdsourcing for task annotation. We hired a set of crowdsourcing workers to monitor posted tasks, and incorporated their judgments into classifier training. Since crowdsourcing workers are not experts in judging a task's impropriety, the quality of worker judgments is often lower than that of operators, and, the reliability of the judgment varies significantly from worker to worker. This variability motivates us to resort to applying quality control techniques to create accurate classifiers.

We conducted a feasibility study of our approach by using real task data from Lancers. We first show that the classifier trained by expert judgments achieves high performance (0.950 averaged area under the ROC curve (AUC)) in detecting improper tasks. We further collect judgments from the non-expert crowdsourcing workers of Lancers, and train a classifier using the judgments of both expert operators and non-expert crowdsourcing workers. Our results show that incorporating the judgments of non-experts achieves a statistically significant improvement (0.962 averaged AUC), and the use of crowdsourced labels allows a reduction in the number of expert judgments by 25% while maintaining the level of detection performance.

In summary, this paper makes three main contributions:

1. To the best of our knowledge, our work is the first to utilize the real operational data inside a commercial crowdsourcing marketplace to build an application for controlling the quality of tasks in crowdsourcing.
2. We use a machine learning approach to the task quality control problem and show that this approach is highly effective in detecting improper tasks in a real crowdsourcing marketplace (Section 3).
3. We show that the use of non-expert judgments by crowdsourcing workers in combination with expert judgments by operators in crowdsourcing marketplaces improves the performance of detecting improper crowdsourcing tasks (Sections 4 and 5).

2. Improper task detection in crowdsourcing marketplaces

2.1. Improper task detection problem

Our goal is to construct a classifier for detecting improper tasks. We formulate this problem as a supervised machine learning problem. Let us assume that there are N crowdsourcing tasks, and each task is represented as a D-dimensional real-valued feature vector denoted by $\boldsymbol{x}_i \in \mathbb{R}^D$. Crowdsourcing marketplaces have their individual definitions of improper tasks, and operators (i.e., domain experts) provide judgments for tasks. Let us denote the expert judgments for a task i by $y_{i,0} \in \{0,1\}$, where a label 1 indicates that a task is improper, and 0 indicates otherwise. In addition to the domain experts, J non-expert crowdsourcing workers are requested to annotate tasks, and we denote a set of workers who give judgments on task i by $\mathcal{J}_i \subseteq \{1,2,\ldots,J\}$. Note that a worker is not required to annotate all the tasks. Each worker is asked to answer K yes–no questions (e.g., "Does the task require workers to give personal information?"). Let $\boldsymbol{y}_{i,j} \in \{0,1\}^K$ be the responses on task i by worker j.

Our goal is to estimate an accurate binary classifier $f:\mathbb{R}^D \to \{0,1\}$ given the annotated dataset $\left(\{\boldsymbol{x}_i\}_{i\in\{1,2,\ldots,N\}}, \{\boldsymbol{y}_{i,j}\}_{i\in\{1,2,\ldots,N\},j\in\mathcal{J}_i}, \{y_{i,0}\}_{i\in\{1,2,\ldots,N\}}\right)$ as a training dataset.

2.2. Dataset

2.2.1. Task and expert judgments

We collected task data posted on a commercial crowdsourcing marketplace, Lancers, from June to November 2012, and created a dataset consisting of 96 improper tasks (judged by operators) and 2,904 randomly selected proper tasks. The dataset included 417 unique requesters. The improper tasks in our dataset can be divided into several categories, as summarized in Table 1.

[3] http://www.lancers.jp.