Quality-Control Mechanism utilizing Worker's Confidence for Crowdsourced Tasks

(Extended Abstract)

Yuko Sakurai[1], Tenda Okimoto[2], Masaaki Oka[3], Masato Shinoda[4], and Makoto Yokoo[3]
1: Kyushu University and JST PRESTO, Fukuoka, 819-0395, Japan, ysakurai@inf.kyushu-u.ac.jp
2: Transdisciplinary Research Integration Center, Tokyo, 101-8430, Japan, tenda@nii.ac.jp
3: Kyushu University, Fukuoka, 819-0395, Japan
{oka@agent., yokoo@}inf.kyushu-u.ac.jp
4: Nara Women's University, Nara, 630-8506, Japan, shinoda@cc.nara-wu.ac.jp

ABSTRACT

We propose a quality control mechanism that utilizes workers' self-reported confidences in crowdsourced labeling tasks. Generally, a worker has confidence in the correctness of her answers, and asking about it is useful for estimating the probability of correctness. However, we need to overcome two main obstacles in order to to use confidence for inferring correct answers. First, a worker is not always well-calibrated. Since she is sometimes over/underconfident, her level of confidence does not always accurately reflect the probability of correctness. In addition, she does not always truthfully report her actual confidence. Therefore, we design an indirect mechanism that enables a worker to declare her confidence by choosing a desirable reward plan from the set of plans that correspond to different confidence intervals. Our mechanism ensures that choosing the plan matching the worker's true confidence maximizes her expected utility.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*; K.4.4 [**Computers and Society**]: Electronic Commerce

General Terms

Algorithms, Economics, Theory, Experimentation, Human Factors

Keywords

Crowdsourcing, Mechanism Design, Human Computation

1. INTRODUCTION

One of the most interesting services recently introduced on the Web is crowdsourcing such as Amazon Mechanical Turk. Such a service is based on the idea of the *wisdom of crowds*, and it solves a problem by combining the forces of many people [4]. An advantage of crowdsourcing is having a large workforce available at relatively low cost, but the quality of the results is sometimes problematic. One straightforward way to infer accurate labels in crowdsourcing is to ask multiple workers to label the same data and accept the majority vote given by the workers. This corresponds to treating the quality of the labels given by different workers equally and simply considering the labels that receive the largest number of votes as the true ones. In crowdsourcing, however, since workers' abilities are not even, treating all labels given by different workers equally is not always a good way to infer true labels.

We consider a quality-control mechanism where a requester asks workers not only for labels on the data but also for their confidence in them. Although studies on designs for quality-control mechanisms for crowdsourced tasks have been advanced by AI and multi-agent system researchers [1, 2], no work has introduced a worker's confidence for in the quality control of general crowdsourced tasks. If the confidence declared by a worker affects her payment, she may over or under-declare her confidence, regardless of her actual confidence in her answers. For example, if we adopt a mechanism that pays more reward to more confident workers, they will obviously be tempted to over-declare their confidence. Therefore, we have to develop a mechanism that is robust against strategic manipulations by workers regarding their own confidence.

In this paper, we design a mechanism that is robust against a worker's over/under-declarations of confidence and tolerant to over/underestimations of it. In our mechanism, a requester offers a set of reward plans that correspond to different confidence intervals, and workers declare their confidence by choosing one of them. One reason why we chose such an *indirect* approach to confidence elicitation is that it is troublesome for a worker to numerically report her confidence. In psychometrics, even if examinees are asked to numerically give their confidence scores (probabilities of correctness), many reported them as if they were binary (0% or 100%) [3]. By showing several reward plans corresponding to discretized confidence levels, a requester can entice workers to consider their confidence at required levels of detail. Furthermore, our mechanism ensures that a worker's expected utility is maximized if she truthfully selects a reward plan that corresponds to her probability of correctness.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. PRELIMINARIES

For simplicity, we consider a task for which the answer is given as a binary label $\{0, 1\}$, such as a yes/no decision problem or image labeling. Let $l \in \{0, 1\}$ denote the true label (answer) for the task. The requester specifies the number of workers, denoted as n, to solve the problem when he posts a task in crowdsourcing. The label given by worker $i \in N$ is denoted as $l_i \in \{0, 1\}$.

The *accuracy of worker* i, that is, the probability that worker i correctly assigns the label, is defined as $a_i = P(l_i = l)$. *Confidence* $x_i \in [0.5, 1]$ stands for worker i's subjective probability of her answer's correctness. If the worker is well-calibrated, x_i is identical to a_i. If the worker is overconfident, $x_i > a_i$ holds. If the worker is underconfident, $x_i < a_i$ holds. Also, we assume that a worker declares her confidence as y_i, which is not always equal to x_i.

The requester sets two *reward functions* f and g when she posts a task in crowdsourcing. The worker's reward is $f(y_i)$ if the requester concludes that her label l_i is true, and it is $g(y_i)$ if her label l_i is false. For any confidence score of y_i, it is natural to assume that $f(y_i) \geq g(y_i)$. Without assuming this condition, a worker may have an incentive to declare the opposite label. Based on this definition, we assume that each worker considers her declared label l_i to be true.

When worker i with her true confidence x_i declares y_i, her *expected utility* is defined as $u(x_i, y_i) = x_i f(y_i) + (1 - x_i)g(y_i)$. We assume that each worker believes that the requester will make a correct decision. In crowdsourcing, it is difficult for each worker to know the other workers' abilities as common knowledge, since the workers are gathered via the network and do not know each other. Also, when the number of workers is reasonably large, no single worker has a decisive power to reverse the decision of the requester. Consequently, we assume that a worker thinks the requester will judge her answer correct with probability x_i.

3. TRUTHFUL M-PLAN MECHANISM

We focus on the difficulty for a worker to estimate her confidence and propose an indirect mechanism that enables a worker to declare her confidence by choosing a desirable reward plan from the set of plans that correspond to different confidence intervals. By showing several reward plans, a requester can entice each worker to consider her confidence at the required level of detail. Our mechanism ensures that a worker's expected utility is maximized if she truthfully selects a reward plan that corresponds to her confidence.

The method for determining the rewards for each plan is as follows: First, we assume that the requester divides the range of confidence scores into m intervals. Let $\mathbf{s} = (s_0, \ldots, s_{m-1}, s_m)$ be the list of threshold confidences, where $s_0 = 0.5$, $s_m = 1$, and $s_{j-1} \leq s_j$ for every j $(1 \leq j \leq m)$. Plan j means the interval $[s_{j-1}, s_j]$. The special case of $m = 1$ corresponds to majority voting. On the other hand, when we use small enough intervals and set m to $+\infty$, this planning is equivalent to the direct revelation mechanism.

Plan j has two different reward amounts (α_j, β_j): α_j applies when a reported label is correct and β_j applies when it is incorrect. Reward functions f and g consist of $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_m$. Thus, we define

$$f(y_i) = \sum_{1 \leq j \leq m} \alpha_j I_{[s_{j-1}, s_j]}(y_i),$$

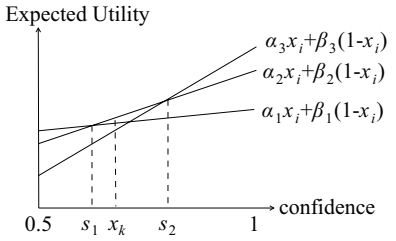

Figure 1: Expected utilities for m-plan mechanism

$$g(y_i) = \sum_{1 \leq j \leq m} \beta_j I_{[s_{j-1}, s_j]}(y_i)$$

, where $I_{[a,b]}(y_i)$ means that $I_{[a,b]}(y_i) = 1$ if $y_i \in [a, b]$ and $I_{[a,b]}(y_i) = 0$ if $y_i \notin [a, b]$.

We propose a method of constructing the reward plans to derive truthful reports from the workers as follows:

- For any plan j, the reward for correct labels should be higher than the reward for incorrect labels: $\alpha_j \geq \beta_j$.
- The rewards for correct labels increase with respect to j: $\alpha_1 < \alpha_2 < \ldots < \alpha_m$.
- The rewards for incorrect labels decrease with respect to j: $\beta_1 > \beta_2 > \ldots > \beta_m$.
- The expected utility of plan j is the same as that of plan $j+1$ at s_j: $\alpha_j s_j + \beta_j(1 - s_j) = \alpha_{j+1} s_j + \beta_{j+1}(1 - s_j)$.

EXAMPLE 1. *Consider a task with a set of three reward plans. We assume that the confidence of worker k, called x_k, is found in $[s_1, s_2]$. Then the maximum expected utility is determined by the function of $\alpha_2 x_i + \beta_2(1 - x_i)$. As shown in Fig. 1, worker k can maximize her expected utility by selecting plan* 2.

Furthermore, we evaluated the 2-plan mechanism for image labeling on AMT. When the threshold confidence is set to 0.75, we categorized the abilities of workers into two groups very effectively and the obtained accuracy of our mechanism exceeded the results of majority voting.

4. ACKNOWLEDGMENTS

This work is supported by a research grant from PRESTO, Japan Science and Technology Corporation (JST).

5. REFERENCES

[1] D. F. Bacon, Y. Chen, I. Kash, D. C. Parkes, M. Rao, and M. Sridharan. Predicting your own effort. In *AAMAS 2012*, pages 695–702, 2012.

[2] R. Cavallo and S. Jain. Efficient crowdsourcing contests. In *AAMAS 2012*, pages 677–686, 2012.

[3] K. Kato and Y. Zhang. An item response model for probability testing. In *International Meeting of the Psychometric Society*, 2010.

[4] E. Law and L. V. Ahn. *Human Computation*. Morgan & Claypool Publishers, 2011.