SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2015, Vol. 58 xxxxxx:1–xxxxxx:12
doi: xxxxxxxxxxxxxx

Crowdsourcing label quality: a theoretical analysis

WANG Wei & ZHOU Zhi-Hua*

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*

Received xxxxxxxx xx, xxxx; accepted xxxxxxxx xx, xxxx

Abstract Crowdsourcing has been an effective and efficient paradigm for providing labels for large-scale unlabeled data. In the past few years, many methods have been developed for inferring labels from the crowd, but few theoretical analyses have been presented to support this popular human-machine interaction process. In this paper, we theoretically study the quality of labels inferred from crowd workers by majority voting and provide an analysis of label quality that shows that the label error rate decreases exponentially with the number of workers selected for each task. We also study the problem of eliminating low-quality workers from the crowd, and provide a conservative condition for eliminating low-quality workers without eliminating any non-low-quality worker with high probability. We also provide an aggressive condition for eliminating all low-quality workers with high probability.

Keywords unlabeled data, crowdsourcing, majority voting, label quality, worker selection

Citation Wang W, Zhou Z-H. Crowdsourcing label quality: a theoretical analysis. Sci China Inf Sci, 2015, 58: xxxxxx(12), doi: xxxxxxxxxxxxxx

1 Introduction

In recent years, unlabeled data can often be obtained abundantly and cheaply for more and more machine learning applications. Generally, learning algorithms need labeled data to train a model for making predictions on future data. This practice has been well established in machine learning. Providing labels for large amounts of unlabeled data has always been a challenge because labeling the data is expensive and time-consuming. Fortunately, crowdsourcing [1,2] has been an effective and efficient paradigm that can provide labels for large-scale unlabeled data in applications across widespread domains such as image annotation, natural language processing, objection recognition and recommendation. The most famous crowdsourcing system is Amazon Mechanical Turk (AMT), a market where users (known as Taskmasters) submit their “microtasks” that can be completed by workers in exchange for small monetary payments. In AMT, users can post any Human Intelligence Tasks (HITs) that computers are currently unable to perform, e.g., annotating images of trees versus images of non-trees.

Crowdsourcing makes it possible for not just one but many independent, relatively inexpensive workers (experts and non-experts) to offer their opinions (labels) on the HITs and determine a solution by aggregating these crowd opinions. The workers usually come from a large range of society and each of them is presented with multiple tasks. The worker has to answer the question about the task presented to

[*] Corresponding author (email: zhouzh@nju.edu.cn)

© Science China Press and Springer-Verlag Berlin Heidelberg 2015

her/him (e.g., whether the image contains trees or not) and provides a label based on her/his own opinion. Among these workers, some may be more reliable than others. Sometimes, there may exist "spammers" who assign random labels in the tasks (e.g., a robot pretending to be a human for the monetary payment) or "adversaries" who give wrong answers deliberately. To improve quality and reliability, common wisdom is to add redundancy into the labels, i.e., each task is presented with multiple workers. In this way, multiple labels are collected for each task and the ground-truth label can be inferred from the crowd. There have been many experimental results that show that this label redundancy could improve the label quality significantly [3–5].

1.1 Prior work

Most popular methods for inferring the labels from the crowd in the past few years build probabilistic models for the crowdsourcing process and derive the labels using algorithms based on Expectation Maximization (EM) [6] and other inference tools. Raykar et al. [7,8] used a two-coin model to measure the performance of each worker in term of sensitivity (true positive rate) and specificity (1-false positive rate) with respect to the unknown ground-truth label. After imposing prior knowledge on sensitivity and specificity, they iteratively estimated the two terms by an EM algorithm with two assumptions: that the performance of each worker is independent of the specific task and that the workers in the crowd are conditionally independent to each other given the ground-truth label. Whitehill et al. [9] formulated a probabilistic model of worker quality and task difficulty in the labeling process with the conditional independence assumption and applied an EM algorithm to infer the most probable label for each task. Subsequently, Welinder et al. [10] introduced worker bias and generalized the probabilistic model in [9] with a high-dimensional formulation of variables representing task difficulty, worker quality, and worker bias. In the circumstance where workers are dominated by spammers, Raykar and Yu [11] proposed an empirical Bayesian algorithm based on EM to iteratively estimate the ground-truth label and eliminate spammers. Liu et al. [12] transformed the crowdsourcing problem into a variational inference problem in graphical models and inferred the labels with variational inference tools including Belief Propagation and Mean Field. As they noted, the performance of their method critically depends on the choice of prior knowledge about the worker's reliability, and the MF-form of their method is closely related to the common methods based on EM algorithms. The minimax entropy principle has also been used to infer crowdsourced labels. Zhou et al. [13] proposed a minimax entropy method with the assumption that the labels are generated by a probabilistic distribution over workers and tasks. Task difficulty and worker quality can be induced by maximizing the entropy of this probabilistic distribution and the ground-truth label can be inferred by minimizing the entropy of this probabilistic distribution. Yan et al. [14] studied learning from the crowd in an active learning setting and employed a probabilistic model to provide criterion for selecting both the task and the worker from which to query the label. Wauthier and Jordan [15] realized that inferring the labels and learning from the inferred labels are often treated separately in crowdsourcing. They proposed a Bayesian framework named the Bayesian Bias Mitigation for Crowdsourcing to unify them.

Because each task is presented to multiple workers and each worker is also presented with multiple tasks, the crowdsourcing system must decide how to choose which tasks are assigned to each worker. Karger et al. [16] proposed a task assignment algorithm based on a random regular bipartite graph and proved the optimality of their algorithm under the assumption that the workers in the crowd are conditionally independent to each other given the class label. However, they only focused on the situation where all tasks are homogeneous, i.e., the label error of a worker does not depend on the specific task and all tasks are equally difficult to her/him. In real-world applications, a worker may have different qualities for different tasks, i.e., her/his performance is not consistent across different tasks. A natural intuition is to apply the exploration-exploitation method that has been studied in the multi-armed bandit problem [17]: estimate the performance of each worker and assign a worker to the tasks that she/he is good at. Several methods have been developed following this approach. Ho and Vaughan [18] proposed an algorithm for assigning heterogeneous tasks to workers with different qualities based on the online primal-dual technique [19]. However, their method assumes that the system can evaluate the worker's performance immediately. Subsequently, Ho and Vaughan [20] utilized tasks with ground-truth labels to estimate the performance of new-coming workers and proposed a provably near-optimal assignment algorithm for heterogeneous tasks based on the online primal-dual technique. Furthermore, Chen et al. [21] formulated a finite-horizon Markov Decision Process in a Bayesian setting for heterogeneous task assignment and characterized the optimality using dynamic programming, which was solved by an optimistic knowledge gradient method.

As for the theoretical analysis of inferring the crowdsourced labels, there have been several results, most of which take an exploration-exploitation approach. Dekel and Shamir [22] studied the problem of pruning low-quality workers in the setting where each task is only labeled by one worker with a two-step process: first identify the workers that significantly deviate from the hypothesis trained on the entire unfiltered data as low-quality workers, then filter the labels labeled by these low-quality workers and retrain the model on the cleaned data. They also presented a theoretical generalization analysis for this two-step process. Tran-Thanh et al. [23] introduced the Multi-Armed Bandit (MAB) model to the study of crowdsourcing. They used the first ϵ budget to estimate the worker's quality and maximized the utility of the remaining $(1-\epsilon)$ budget based on these estimations. An upper bound on the regret for their Bounded ϵ-first algorithm was derived. Abraham et al. [24] formalized a model for worker selection called the Bandit Survey problem, which is similar to but technically different from the MAB problem, because the MAB problem collects rewards as feedback while the Bandit Survey problem collects the opinions of workers. They also presented the algorithms and a theoretical analysis for this Bandit Survey model.

1.2 Our focus and contributions

In prior work, existing well-designed methods for inferring the labels from the crowd have their restrictions: the methods based on probabilistic models and inference tools heavily depend on prior knowledge of probabilistic distribution and the initial starting point. If the prior knowledge is different from the distribution generating the ground-truth, the performance is dramatically poor. Sometimes, the solution may be locally optimal even if the prior knowledge is exactly the same as the distribution generating the ground-truth. The methods for task assignment either need tasks with ground-truth labels to estimate the quality of the worker or heavily depend on prior knowledge. Existing theoretical studies on crowdsourcing either focus on the setting where each task is only labeled by one worker or formalize the crowdsourcing process as a bandit problem where the feedback relies on the immediate evaluation of the worker's label, which cannot be applied to the general crowdsourcing problem.

In real-world crowdsourcing applications, the simplest and most general method for inferring labels is to use majority voting over multiple labels. It is a good error-pruning strategy and there have been many reported experimental results on real-world crowdsourcing data sets [3–5] the showed that majority voting performs significantly well at improving label quality. Inspired by these observations, we present a theoretical analysis of label quality for majority voting in crowdsourcing that shows that the label error rate decreases exponentially with the number of workers selected for each task. We also provide a conservative condition for eliminating low-quality workers without eliminating any non-low-quality worker with high probability. We also present an aggressive condition for eliminating all low-quality workers with high probability.

The rest of this paper is organized as follows. After introducing some preliminaries in Section 2, we present the theoretical analysis on label quality in Section 3 and study the problem of eliminating low-quality workers in Section 4. Finally, we draw our conclusions in Section 5.

2 Preliminaries

The taskmaster has a set of m tasks $\{t_1, \ldots, t_m\}$ over $\mathcal{X}$, each task t_i corresponds to a binary classification example x_i with an unobserved ground-truth label $y_i \in \{0, 1\}$, $1 \leqslant i \leqslant m$ (e.g., annotating whether an image includes trees or not). The taskmaster assigns these tasks to the workers in the crowd, where W denotes the set of all workers in the crowd. To improve quality and reliability, an example x_i is generally presented to N workers denoted by $\{w_1, \dots, w_N\}$, $w_j \in W$, for instance, by randomizing the order of the workers and selecting the first N. Each worker w_j makes a prediction on x_i and creates a label $y_i^j = w_j(x_i) \in \{0, 1\}$. A final label is then inferred for x_i based on the labels $\{y_i^1 \dots, y_i^N\}$ provided by the N workers. In this paper, we focus on the majority voting strategy that uses the majority label as the inferred label $\widehat{y}_i$ for the binary classification problem, i.e.,

$$\widehat{y}_i = \begin{cases} 1, & \text{if } \frac{1}{N}\sum_{j=1}^{N} y_i^j > \frac{1}{2}, \\ \text{random guess}, & \text{if } \frac{1}{N}\sum_{j=1}^{N} y_i^j = \frac{1}{2}, \\ 0, & \text{if } \frac{1}{N}\sum_{j=1}^{N} y_i^j < \frac{1}{2}. \end{cases} \tag{1}$$

This is a good error-pruning strategy when no information is known about the worker's quality. Here, we utilize the two-coin model introduced by Raykar et al. [8] to characterize the quality of each worker. For a random example (x, y), worker $w_j \in W$ provides label y^j on x based on two biased coins. If the ground-truth label y is 1, the worker flips a coin with bias α_j (sensitivity) and provides a correct label $y^j = y$ with probability α_j; if y is 0, the worker flips a coin with bias β_j (specificity) and provides a correct label $y^j = y$ with probability β_j. For $y = 1$, the sensitivity for w_j is defined as

$$\alpha_j = P(y^j = 1|y = 1);$$

for $y = 0$, the specificity for w_j is defined as

$$\beta_j = P(y^j = 0|y = 0).$$

Actually, a similar one-coin model has been considered for task assignment in Karger et al. [16] and Ho et al. [20]. Karger et al. [16] assumed that w_j is characterized by reliability parameter p_j that generates error randomly for each example, i.e., w_j provides label y^j for x such that $y^j = y$ with probability p_j and $y^j \neq y$ with probability $1 - p_j$. This reliability parameter p_j does not depend on the specific example. This setting discussed by Karger et al. [16] is called a homogeneous setting because each worker has equal quality on all examples. Ho et al. [20] generalized this to the heterogeneous setting, where the tasks can be divided into several types and each worker performs consistently on tasks of the same type. The reason why we follow the two-coin model is that the worker may have different qualities for different task classes.

3 Theoretical analysis on label quality

3.1 Uniform distribution with bounded workers

In this section, we start with a simple model for analyzing the quality of inferred labels. There are some kind of tasks that require domain knowledge to complete, so the taskmaster hopes to attract the people she/he needs and select an appropriate crowd. In this situation, she/he may assign tasks to the community that consists of people who have knowledge about the tasks and will complete them honestly. Once the taskmaster has selected the appropriate crowd, it is reasonable to assume that the performance of the worker in the crowd is no worse than predicting all tasks as positive or negative because they have knowledge about the tasks. Without loss of generality, we assume that the positive class is the minority class and $A(w_j) \geqslant P(y = 1)$ for all workers $w_j \in W$, where $A(\cdot)$ is the accuracy defined as

$$A(w_j) = P_{(x \in \mathcal{X}, y)}\big(w_j(x) = y\big).$$

Hence for worker w_j we obtain

$$A(w_j) = P\big(w_j(x) = 1, y = 1\big) + P\big(w_j(x) = 0, y = 0\big) = \alpha_j P(y = 1) + \beta_j P(y = 0).$$

Here, we assume that α_j and β_j do not depend on the specific example, i.e., w_j has equal quality on all examples, as in Raykar et al. [8]. We discuss the setting where w_j has different qualities on different examples in Subsection 3.2.

Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the random variables corresponding to α_j and β_j over different workers, respectively. Considering that the crowd may include workers with all possible sensitivities and specificities and that each worker is selected randomly from the crowd, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be thought of as uniformly distributed over some domain. With $A(w_j) \geqslant P(y=1)$ for $w_j \in W$, we get that $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ should satisfy the following constraint:

$$P(y=1) \cdot \boldsymbol{\alpha} + P(y=0) \cdot \boldsymbol{\beta} \geqslant P(y=1).$$

Thus, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is uniformly distributed over the domain $0 \leqslant \boldsymbol{\alpha} \leqslant 1$, $0 \leqslant \boldsymbol{\beta} \leqslant 1$ and $P(y=1) \cdot \boldsymbol{\alpha} + P(y=0) \cdot \boldsymbol{\beta} \geqslant P(y=1)$. A similar uniform setting where the workers range from very bad to very good was discussed by Dekel and Shamir [22], but for a different theoretical analysis. We now present the following result on label quality:

Theorem 1. Suppose that the sensitivities and specificities of the workers in the crowd are uniformly distributed over the above domain and each task is presented with N workers selected randomly from the crowd. The inferred labels generated by majority voting then satisfy the following bound:

$$P\big(\widehat{y} \neq y\big) \leqslant 2\exp\left(-\frac{C^2N}{18(2-C)^2}\right)P\left(y=1\right) + 2\exp\left(-\frac{(3C-2C^2)^2N}{18(2-C)^2}\right)P\big(y=0\big).$$

Here, $C = \frac{P(y=1)}{P(y=0)} \leqslant 1$.

Proof. Letting $C = \frac{P(y=1)}{P(y=0)} \leqslant 1$, the constraint $P(y=1) \cdot \boldsymbol{\alpha} + P(y=0) \cdot \boldsymbol{\beta} \geqslant P(y=1)$ can be simplified to $C \cdot \boldsymbol{\alpha} + \boldsymbol{\beta} \geqslant C$. Because $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is uniformly distributed over the domain $0 \leqslant \boldsymbol{\alpha} \leqslant 1$, $0 \leqslant \boldsymbol{\beta} \leqslant 1$ and $C \cdot \boldsymbol{\alpha} + \boldsymbol{\beta} \geqslant C$, the probability density function of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{2}{2-C}$. We may then calculate the marginal probability density functions $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\beta})$. First,

$$p(\boldsymbol{\alpha}) = \int_{C-C\cdot\boldsymbol{\alpha}}^{1} \frac{2}{2-C}\mathrm{d}\boldsymbol{\beta} = \frac{2(1-C+C\cdot\boldsymbol{\alpha})}{2-C}.$$

Second, if $0 \leqslant \boldsymbol{\beta} < C$, we get

$$p(\boldsymbol{\beta}) = \int_{1-\frac{\boldsymbol{\beta}}{C}}^{1} \frac{2}{2-C}\mathrm{d}\boldsymbol{\alpha} = \frac{2\boldsymbol{\beta}}{(2-C)C}.$$

If $C \leqslant \boldsymbol{\beta} \leqslant 1$, we get

$$p(\boldsymbol{\beta}) = \int_{0}^{1} \frac{2}{2-C}\mathrm{d}\boldsymbol{\alpha} = \frac{2}{2-C}.$$

We now calculate the expectations $\mathbb{E}(\boldsymbol{\alpha})$ and $\mathbb{E}(\boldsymbol{\beta})$ using the marginal probability density functions.

$$\mathbb{E}(\boldsymbol{\alpha}) = \int_{0}^{1} \boldsymbol{\alpha} \cdot p(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\alpha} = \frac{3-C}{6-3C},$$
$$\mathbb{E}(\boldsymbol{\beta}) = \int_{0}^{C} \boldsymbol{\beta} \cdot p(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\beta} + \int_{C}^{1} \boldsymbol{\beta} \cdot p(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\beta} = \frac{3-C^2}{6-3C}.$$

For example (x, y), N workers $w_1, \ldots, w_N$ are selected randomly from the crowd and each of them provides a label y^j on x based on her/his sensitivity and specificity. The inferred label $\widehat{y}$ is generated by majority voting according to (1). If the ground-truth label $y=1$, let $S_{\boldsymbol{\alpha}} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}\big(y^j = 1 | y = 1\big)$. From Lemma 1 we obtain

$$\begin{aligned} P\big(\widehat{y} \neq y | y=1\big) &= P\big(\widehat{y} = 0 | y = 1\big) = P\left(S_{\boldsymbol{\alpha}} < \frac{1}{2}\right) \\ &\leqslant P\left(\big|\mathbb{E}(S_{\boldsymbol{\alpha}}) - S_{\boldsymbol{\alpha}}\big| > \mathbb{E}(S_{\boldsymbol{\alpha}}) - \frac{1}{2}\right) \\ &\leqslant 2\exp\left(-\frac{C^2N}{18(2-C)^2}\right). \end{aligned}$$

If the ground-truth label $y=0$, let $S_{\boldsymbol{\beta}} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}\big(y^j=0|y=0\big)$. From Lemma 1, we obtain

$$\begin{aligned}P\big(\widehat{y}\neq y|y=0\big) &= P\big(\widehat{y}=1|y=0\big) = P\left(S_{\boldsymbol{\beta}} < \frac{1}{2}\right)\\ &\leqslant P\left(\big|\mathbb{E}(S_{\boldsymbol{\beta}}) - S_{\boldsymbol{\beta}}\big| > \mathbb{E}(S_{\boldsymbol{\beta}}) - \frac{1}{2}\right)\\ &\leqslant 2\exp\left(-\frac{(3C-2C^2)^2N}{18(2-C)^2}\right).\end{aligned}$$

Considering that $P\big(\widehat{y}\neq y\big) = P\big(\widehat{y}\neq y|y=1\big)P(y=1) + P\big(\widehat{y}\neq y|y=0\big)P(y=0)$, Theorem 1 is proved.

Theorem 1 states that the label error rate decreases exponentially with the number of workers selected for each task when the quality of each worker is bounded and each worker is selected randomly from the crowd.

3.2 Analog-Gaussian distribution with unbounded workers

In Subsection 3.1, we assume that each worker has equal quality for all examples. We now generalize to the setting where the sensitivity and specificity of worker w_j depend on the task x_i, i.e.,

$$\alpha_{i,j} = P(y_i^j = 1|y_i = 1), \quad \beta_{i,j} = P(y_i^j = 0|y_i = 0).$$

In AMT, once the tasks are posted by the taskmaster, thousands of people have internet access to them. Sometimes, there may be adversaries in the crowd who give the incorrect answer for the task deliberately ($\alpha_{i,j}$ or $\beta_{i,j}=0$). The accuracy of an adversary is 0, so the qualities of workers in the crowd are no longer bounded. In real-world crowdsourcing, there may exist some workers who have the same sensitivity and specificity. Thus, although each worker is selected randomly from the crowd, the distribution of sensitivity and specificity over all workers is no longer uniform. Intuitively, normal workers are much more than experts, spammers, and adversaries. For example x_i, the distribution of sensitivity and specificity for the workers is more like a Gaussian distribution, i.e.,

$$p(\boldsymbol{\alpha_i}) = \frac{1}{Z_{\boldsymbol{\alpha_i}}\sqrt{2\pi}}\exp\left(-\frac{(\boldsymbol{\alpha_i}-\mu_i)^2}{2}\right), \quad \boldsymbol{\alpha_i}\in[0,1], \tag{2}$$

$$p(\boldsymbol{\beta_i}) = \frac{1}{Z_{\boldsymbol{\beta_i}}\sqrt{2\pi}}\exp\left(-\frac{(\boldsymbol{\beta_i}-\nu_i)^2}{2}\right), \quad \boldsymbol{\beta_i}\in[0,1]. \tag{3}$$

Here, $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$ denote the random variables corresponding to $\alpha_{i,j}$ and $\beta_{i,j}$ on x_i over the workers, respectively. Terms $Z_{\boldsymbol{\alpha_i}}$ and $Z_{\boldsymbol{\beta_i}}$ are normalization terms such that

$$\int_0^1 p(\boldsymbol{\alpha_i})\mathrm{d}\boldsymbol{\alpha_i} = 1 \quad \text{and} \quad \int_0^1 p(\boldsymbol{\beta_i})\mathrm{d}\boldsymbol{\beta_i} = 1, \tag{4}$$

and $0\leqslant \mu_i,\nu_i\leqslant 1$. The distribution parameters μ_i and ν_i represent the opinions of the most commonly occurring workers in the crowd on x_i. We now present the following result on label quality:

Theorem 2. For example (x_i,y_i), suppose that the sensitivity and specificity on x_i over the workers are distributed as (2) and (3) and the task x_i is presented to N workers selected randomly from the crowd. If $\mu_i > \frac{1}{2}$ and $\nu_i > \frac{1}{2}$, the inferred label $\widehat{y_i}$ generated by majority voting satisfies the following bound:

$$P\big(\widehat{y_i}\neq y_i\big) \leqslant 2\exp\big(-2N(\varphi(\mu_i)-1/2)^2\big)\mathbb{I}\big(y_i=1\big) + 2\exp\big(-2N(\varphi(\nu_i)-1/2)^2\big)\mathbb{I}\big(y_i=0\big).$$

Here, $\varphi(\tau) = 0.827\big(\phi(\tau)-\phi(1-\tau)\big)+\tau$ and $\phi(\tau)=\exp(-\frac{\tau^2}{2})$.

Proof. Let $X = \boldsymbol{\alpha_i} - \mu_i$ and $Y = \frac{X}{\sqrt{2}}$. We obtain

$$\mathbb{E}(\boldsymbol{\alpha_i}) = \int_0^1 \boldsymbol{\alpha_i}\cdot p(\boldsymbol{\alpha_i})\mathrm{d}\boldsymbol{\alpha_i}$$

$$
\begin{aligned}
&= \frac{1}{Z_{\boldsymbol{\alpha_i}}\sqrt{2\pi}}\int_{-\mu_i}^{1-\mu_i} X\cdot\exp\left(-\frac{X^2}{2}\right)\mathrm{d}X + \frac{\mu_i}{Z_{\boldsymbol{\alpha_i}}\sqrt{2\pi}}\int_{-\mu_i}^{1-\mu_i}\exp\left(-\frac{X^2}{2}\right)\mathrm{d}X\\
&= \frac{1}{Z_{\boldsymbol{\alpha_i}}\sqrt{2\pi}}\int_{-\mu_i/\sqrt{2}}^{(1-\mu_i)/\sqrt{2}}\sqrt{2}Y\cdot\exp(-Y^2)\mathrm{d}Y+\mu_i. \qquad (5)
\end{aligned}
$$

Considering the integral formulation

$$\int Y\cdot\exp(-Y^2)\mathrm{d}Y = -\frac{1}{2}\exp(-Y^2),$$

from (5), we get

$$\mathbb{E}(\boldsymbol{\alpha_i}) = \frac{1}{2Z_{\boldsymbol{\alpha_i}}\sqrt{\pi}}\left(\phi(\mu_i)-\phi(1-\mu_i)\right)+\mu_i. \qquad (6)$$

With (4), we have

$$Z_{\boldsymbol{\alpha_i}} = \frac{1}{\sqrt{2\pi}}\int_0^1\exp\left(-\frac{(\boldsymbol{\alpha_i}-\mu_i)^2}{2}\right)\mathrm{d}\boldsymbol{\alpha_i} = \frac{1}{\sqrt{2\pi}}\int_{-\mu_i}^{1-\mu_i}\phi(X)\mathrm{d}X.$$

For some $\Delta > 0$ and $1/2 < \mu_i < \mu_i+\Delta \leqslant 1$, we get

$$\int_{-(\mu_i+\Delta)}^{1-(\mu_i+\Delta)}\phi(X)\mathrm{d}X - \int_{-\mu_i}^{1-\mu_i}\phi(X)\mathrm{d}X = \int_{\mu_i}^{\mu_i+\Delta}\phi(X)\mathrm{d}X - \int_{1-(\mu_i+\Delta)}^{1-\mu_i}\phi(X)\mathrm{d}X < 0.$$

Hence, $Z_{\boldsymbol{\alpha_i}}$ is a monotonically decreasing function of μ_i for $\frac{1}{2} < \mu_i \leqslant 1$. Let

$$\Phi(\lambda) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\lambda}\phi(X)\mathrm{d}X, \qquad (7)$$

we have

$$\min_{\mu_i\in(\frac{1}{2},1]} Z_{\boldsymbol{\alpha_i}} = Z_{\boldsymbol{\alpha_i}}\big|_{\mu_i=1} = \frac{1}{\sqrt{2\pi}}\int_0^1\phi(X)\mathrm{d}X = \Phi(1)-\Phi(0) = 0.8413-0.5 = 0.3413.$$

For $\frac{1}{2} < \mu_i \leqslant 1$, from (6), we have

$$\mathbb{E}(\boldsymbol{\alpha_i}) \geqslant 0.827\left(\phi(\mu_i)-\phi(1-\mu_i)\right)+\mu_i. \qquad (8)$$

Let $\varphi(\mu_i) = 0.827\left(\phi(\mu_i)-\phi(1-\mu_i)\right)+\mu_i$. We calculate its derivative, which is

$$\varphi'(\mu_i) = 0.827\left(-\mu_i\phi(\mu_i)-(1-\mu_i)\phi(1-\mu_i)\right)+1. \qquad (9)$$

Because $1/2 < \mu_i \leqslant 1$, from (9), we obtain

$$\varphi'(\mu_i) > 1-0.827\phi(1-\mu_i) > 0. \qquad (10)$$

This means that $\varphi(\mu_i)$ is a monotonically increasing function of μ_i for $\frac{1}{2} < \mu_i \leqslant 1$. Hence, we get

$$\varphi(\mu_i) > \varphi(\mu_i)\Big|_{\mu_i=1/2} = 1/2.$$

With (8), we obtain $\mathbb{E}(\boldsymbol{\alpha_i}) \geqslant \varphi(\mu_i) > 1/2$. If $y_i = 1$, let $S_{\boldsymbol{\alpha_i}} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(y_i^j = 1|y_i = 1)$. From Lemma 1, we get

$$
\begin{aligned}
P\left(\widehat{y_i}\neq y_i|y_i=1\right) &= P\left(S_{\boldsymbol{\alpha_i}} < 1/2\right)\\
&\leqslant P\left(|\mathbb{E}(\boldsymbol{\alpha_i}) - S_{\boldsymbol{\alpha_i}}| > \mathbb{E}(\boldsymbol{\alpha_i}) - 1/2\right)\\
&\leqslant 2\exp\left(-2N(\varphi(\mu_i)-1/2)^2\right).
\end{aligned}
$$

Similarly, we obtain

$$\mathbb{E}(\boldsymbol{\beta_i}) = \frac{1}{2Z_{\boldsymbol{\beta_i}}\sqrt{\pi}}\left(\phi(\nu_i) - \phi(1-\nu_i)\right) + \nu_i > 1/2.$$

If $y_i = 0$, let $S_{\boldsymbol{\beta_i}} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}\left(y_i^j = 0|y_i = 0\right)$. We then have

$$\begin{aligned} P\left(\widehat{y_i} \neq y_i|y_i = 0\right) &= P\left(S_{\boldsymbol{\beta_i}} < 1/2\right) \\ &\leqslant P\left(|\mathbb{E}(\boldsymbol{\beta_i}) - S_{\boldsymbol{\beta_i}}| > \mathbb{E}(\boldsymbol{\beta_i}) - 1/2\right) \\ &\leqslant 2\exp\left(-2N(\varphi(\nu_i) - 1/2)^2\right). \end{aligned}$$

Thus, Theorem 2 is proved.

Theorem 2 states that when $\mu_i, \nu_i > \frac{1}{2}$, the probability of inferring the incorrect label for x_i decreases exponentially with the number of workers selected for it. When $\mu_i, \nu_i > \frac{1}{2}$, the most commonly occurring workers in the crowd will make correct predictions on x_i with high probability, i.e., the crowd is capable of completing the task. When $\mu_i, \nu_i < \frac{1}{2}$, we obtain the following result:

Corollary 1. For example (x_i, y_i), suppose that the sensitivity and specificity for x_i over the workers are distributed as (2) and (3) and the task x_i is presented to N workers selected randomly from the crowd. If $\mu_i < \frac{1}{2}$ and $\nu_i < \frac{1}{2}$, the inferred label $\widehat{y_i}$ generated by majority voting satisfies the following bound:

$$P\left(\widehat{y_i} = y_i\right) \leqslant 2\exp\left(-2N(\varphi(\mu_i) - 1/2)^2\right)\mathbb{I}\left(y_i = 1\right) + 2\exp\left(-2N(\varphi(\nu_i) - 1/2)^2\right)\mathbb{I}\left(y_i = 0\right).$$

Here, $\varphi(\tau)$ are given in Theorem 2.

Corollary 1 states that when $\mu_i, \nu_i < \frac{1}{2}$, the probability of inferring the correct label for x_i decreases exponentially with the number of workers selected for it. When $\mu_i, \nu_i < \frac{1}{2}$, the most commonly occurring workers in the crowd will make incorrect predictions on x_i with high probability, i.e., the crowd is almost dominated by low-quality workers and is not capable of completing the current task. For such extreme crowd dominated by low-quality workers, we may never achieve good label quality. In the following sections, we mainly focus on the situation when $\mu_i, \nu_i > \frac{1}{2}$.

Theorem 2 provides the error bound on x_i. We now derive the error bound over the example space $\mathcal{X}$. If there exists some $\xi_0 > \frac{1}{2}$ such that $\xi_0 \leqslant \mu_i$ and $\xi_0 \leqslant \nu_i$ for all $x_i \in \mathcal{X}$, we obtain the following result:

Theorem 3. Suppose there exists some $\xi_0 > \frac{1}{2}$ such that $\xi_0 \leqslant \mu_i$ and $\xi_0 \leqslant \nu_i$ for all $x_i \in \mathcal{X}$. The inferred labels generated by majority voting then satisfy the following bound:

$$P_{(x,y)}\left(\widehat{y} \neq y\right) \leqslant 2\exp\left(-2N(\varphi(\xi_0) - 1/2)^2\right).$$

Here, $\varphi(\tau)$ are given in Theorem 2.

Proof. Because $\mu_i \geqslant \xi_0$, $\nu_i \geqslant \xi_0$ and $\varphi(\tau)$ is a monotonically increasing function of τ for $1/2 < \tau \leqslant 1$, we get $\varphi(\xi_0) \geqslant \varphi(\mu_i)$ and $\varphi(\xi_0) \geqslant \varphi(\mu_i)$. Considering that

$$P_{(x,y)}\left(\widehat{y} \neq y\right) = \int_{x_i \in \mathcal{X}} P\left(\widehat{y_i} \neq y_i\right) p(x_i)\mathrm{d}x_i,$$

by Theorem 2, Theorem 3 is proved.

For the heterogeneous setting discussed in Ho et al. [20] where the tasks are divided into several types and each worker performs consistently on tasks of the same type, the condition in Theorem 3 can be met. This is easy to understand: support there are l types, and each type of tasks corresponds to a pair (μ_b, ν_b), where $1 \leqslant b \leqslant l$, $\mu_b > \frac{1}{2}$ and $\nu_b > \frac{1}{2}$. For these l types, there exists some $\xi_0 > \frac{1}{2}$ such that $\xi_0 \leqslant \mu_b$ and $\xi_0 \leqslant \nu_b$.

Unfortunately, sometimes the condition in Theorem 3 may not be met, i.e., for any small $\xi > 0$, there exists some $x_k \in \mathcal{X}$ such that $\mu_k - \frac{1}{2} < \xi$ or $\nu_k - \frac{1}{2} < \xi$. For this task x_k, its μ_k or ν_k is very close to 1/2. This means that the most commonly occurring workers in the crowd will make incorrect predictions on x_k with a probability that is very close to 1/2, i.e., x_k is a very difficult task for the crowd to complete.

Generally, this difficulty may be caused by various feature noise. For the tasks in which the features are corrupted badly, workers can only guess their labels. We use the Tsybakov [25] condition, which is usually used to characterize difficult examples in learning problem, to model these tasks that are difficult for the crowd, i.e., for some $C_T > 0$, $\gamma > 0$ and all $0 < \xi \leqslant 1/2$,

$$P\left(x_i \in \mathcal{X} : \mu_i - 1/2 \leqslant \xi \vee \nu_i - 1/2 \leqslant \xi\right) \leqslant C_T \xi^{\gamma}. \tag{11}$$

Based on the Tsybakov condition, we give the following result on label quality:

Theorem 4. Suppose that the tasks in $\mathcal{X}$ meet the Tsybakov condition in (11). The inferred labels generated by majority voting satisfy the following bound for any $0 < \xi \leqslant 1/2$:

$$P_{(x,y)}\left(\widehat{y} \neq y\right) \leqslant 2\exp\left(-2N(\varphi(\xi) - 1/2)^2\right) + C_T \xi^{\gamma}.$$

Here, $\varphi(\tau)$ are given in Theorem 2.

Proof. Considering that $\mathcal{X} = \{x_i : \mu_i - 1/2 \leqslant \xi \vee \nu_i - 1/2 \leqslant \xi\} \cup \{x_i : \mu_i - 1/2 > \xi \wedge \nu_i - 1/2 > \xi\}$, From Theorem 3 and (11), Theorem 4 is proved.

Theorem 4 states that the error bound on the inferred labels depends on the number of workers selected for each task and difficult tasks, as characterized by the Tsybakov condition. The first term of the right hand side in the bound indicates the probability of inferring the incorrect labels for non-difficult tasks (measured by ξ) and the second term of the right hand side in the bound indicates the quantity of difficult tasks. For these difficult tasks, it is hard to infer their ground-truth labels. Furthermore, the Tsybakov condition shown in (11) can also be generalized to the case where there are extreme examples x_s for which $\mu_s, \nu_s < \frac{1}{2}$ (discussed in Corollary 1). These extreme examples are also included in the Tsybakov condition.

The distribution in (2) and (3) can be generalized as the following distribution:

$$p(\boldsymbol{\alpha_i}) = \frac{1}{Z_{\boldsymbol{\alpha_i}}\sqrt{2\pi}\sigma}\exp\left(-\frac{(\boldsymbol{\alpha_i} - \mu_i)^2}{2\sigma^2}\right), \tag{12}$$

$$p(\boldsymbol{\beta_i}) = \frac{1}{Z_{\boldsymbol{\beta_i}}\sqrt{2\pi}\sigma}\exp\left(-\frac{(\boldsymbol{\beta_i} - \nu_i)^2}{2\sigma^2}\right). \tag{13}$$

This distribution is not difficult to understand: σ denotes the variance and has no influence on the expectation in the Gaussian distribution, hence we can derive the error bounds. However, the parameters $\varphi(\mu_i)$, $\varphi(\nu_i)$, $\varphi(\xi)$ and $\varphi(\xi_0)$ depend on σ.

4 Eliminating low-quality workers

We call a worker w_j as a low-quality worker if she/he provides correct predictions for tasks with a probability that is no larger than $\frac{1}{2}$, i.e., $A(w_j) \leqslant \frac{1}{2}$. Obviously, these low-quality workers will degenerate the quality of the crowd. If some of them can be removed, the label quality will be improved. As the crowdsourcing process continues, the workers complete more and more tasks and the final labels for these tasks are inferred by majority voting. The theoretical results in Section 3 show that these inferred labels are good approximations of the ground-truth labels. Intuitively, we can count the number of labels that are different from the inferred labels for each worker, and the workers who have large numbers of such labels will be the low-quality ones with a high probability. Let $M_{\boldsymbol{\alpha}}$ denote the number of tasks with an inferred label of 1 completed by w_j, and let $M_{\boldsymbol{\beta}}$ denote the number of tasks with an inferred label of 0 completed by w_j[1]. Further, we define the following pseudo-sensitivity and pseudo-specificity of w_j to be

$$\widehat{\alpha}(w_j) = \frac{1}{M_\alpha}\sum_{i=1}^{M_\alpha}\mathbb{I}\left(y_i^j = 1 | \widehat{y}_i = 1\right), \quad \widehat{\beta}(w_j) = \frac{1}{M_\beta}\sum_{i=1}^{M_\beta}\mathbb{I}\left(y_i^j = 0 | \widehat{y}_i = 0\right).$$

[1] In AMT, each worker has a registered ID to get the monetary payments for completing the tasks, hence the system can track the tasks completed by each worker. In this paper, we ignore the situation where several workers share one ID.

Workers with small pseudo-sensitivity and small pseudo-specificity may be low-quality workers. Once the probable low-quality workers are identified, we can ignore them or decrease the probability that they are selected for future tasks. Because the inferred labels are not ground-truth, we must do this carefully, otherwise we may eliminate some non-low-quality workers.

Theorem 5. Let $\eta_1 \geqslant P(\widehat{y} \neq 1|y = 1)$ denote the upper bound on the error rate of inferred labels for the positive class and let $\eta_0 \geqslant P(\widehat{y} \neq 0|y = 0)$ denote the upper bound on the error rate of inferred labels for the negative class. For any $\delta \in (0, 1)$, if $\widehat{\alpha}(w_j) \leqslant (1/2-\eta_1)C_1 - \epsilon_1$ and $\widehat{\beta}(w_j) \leqslant (1/2-\eta_0)C_0 - \epsilon_0$, then $A(w_j) \leqslant 1/2$ holds with a probability of at least $1 - \delta$. Here, $\epsilon_1 = \sqrt{\frac{\ln(4/\delta)}{2M_{\boldsymbol{\alpha}}}}$, $\epsilon_0 = \sqrt{\frac{\ln(4/\delta)}{2M_{\boldsymbol{\beta}}}}$, $C_1 = \frac{P(y=1)}{P(\widehat{y}=1)}$, and $C_0 = \frac{P(y=0)}{P(\widehat{y}=0)}$.

Proof. Using Lemma 2, we obtain that $P(y^j = 1|\widehat{y} = 1) \leqslant \widehat{\alpha}(w_j) + \epsilon_1$ holds with a probability of at least $1 - \delta/2$. If $\widehat{\alpha}(w_j) \leqslant (1/2 - \eta_1)C_1 - \epsilon_1$, we get

$$\begin{aligned}
P(y^j = 1|y = 1) &= \frac{P(y^j = 1, \widehat{y} = 1, y = 1)}{P(y = 1)} + \frac{P(y^j = 1, \widehat{y} \neq 1, y = 1)}{P(y = 1)} \\
&\leqslant \frac{P(y^j = 1, \widehat{y} = 1)}{P(y = 1)} + \frac{P(\widehat{y} \neq 1, y = 1)}{P(y = 1)} \\
&= P(y^j = 1|\widehat{y} = 1)/C_1 + P(\widehat{y} \neq 1|y = 1) \leqslant 1/2.
\end{aligned}$$

Similarly, if $\widehat{\beta}(w_j) \leqslant (1/2 - \eta_0)C_0 - \epsilon_0$, we get that $P(y^j = 0|y = 0) \leqslant 1/2$ holds with a probability of at least $1 - \delta/2$. Thus, Theorem 5 is proved.

Theorem 5 indicates that we can identify the low-quality workers by estimating their pseudo-sensitivity and pseudo-specificity using the inferred labels. This provides a conservative condition for eliminating the low-quality workers because it will never remove any non-low-quality worker ($A(w_j) > \frac{1}{2}$) with a probability of at least $1 - \delta$, which is a suitable strategy for the setting where there are not abundant workers in the crowd. The upper bounds η_1 and η_0 have been discussed in the proofs of the theorems in Section 3; they decrease exponentially with the number of workers selected for each task. Further, $C_1, C_0 \approx 1$, so $\widehat{\alpha}(w_j)$ and $\widehat{\beta}(w_j)$ are close to and smaller than 1/2.

Theorem 6. Let $\eta_1 \geqslant P(\widehat{y} \neq 1|y = 1)$ denote the upper bound on the error rate of the inferred labels for the positive class and let $\eta_0 \geqslant P(\widehat{y} \neq 0|y = 0)$ denote the upper bound on the error rate of the inferred labels for the negative class. For any $\delta \in (0, 1)$, if $A(w_j) \leqslant 1/2$, then $\widehat{\alpha}(w_j) \leqslant 1 - (\frac{1}{2} - \eta_1)C_1 + \epsilon_1$ or $\widehat{\beta}(w_j) \leqslant 1 - (\frac{1}{2} - \eta_0)C_0 + \epsilon_0$ holds with a probability of at least $1 - \delta$. Here, ϵ_1, ϵ_0, C_1, and C_0 are given in Theorem 5.

Proof. Because $A(w_j) \leqslant 1/2$, we obtain that either $P(y^j = 1|y = 1) \leqslant 1/2$ or $P(y^j = 0|y = 0) \leqslant 1/2$ holds.

If $P(y^j = 1|y = 1) \leqslant 1/2$, we get

$$\begin{aligned}
P(y^j = 1|\widehat{y} = 1) &= \frac{P(y^j = 1, y = 1, \widehat{y} = 1)}{P(\widehat{y} = 1)} + \frac{P(y^j = 1, y \neq 1, \widehat{y} = 1)}{P(\widehat{y} = 1)} \\
&\leqslant \frac{P(y^j = 1, y = 1)}{P(\widehat{y} = 1)} + \frac{P(y \neq 1, \widehat{y} = 1)}{P(\widehat{y} = 1)} \\
&= P(y^j = 1|y = 1) \cdot C_1 + \frac{P(y \neq 1, \widehat{y} = 1)}{P(\widehat{y} = 1)}. \qquad (14)
\end{aligned}$$

Considering that

$$\begin{aligned}
\frac{P(y \neq 1, \widehat{y} = 1)}{P(\widehat{y} = 1)} &= \frac{P(\widehat{y} = 1) - (P(y = 1) - P(y = 1, \widehat{y} \neq 1))}{P(\widehat{y} = 1)} \\
&= 1 - C_1 + P(\widehat{y} \neq 1|y = 1) \cdot C_1 \\
&\leqslant 1 - C_1 + \eta_1 \cdot C_1, \qquad (15)
\end{aligned}$$

from (14), we obtain $P(y^j = 1|\widehat{y} = 1) \leqslant 1-(\frac{1}{2}-\eta_1)C_1$. From Lemma 2, we obtain that $P(y^j = 1|\widehat{y} = 1) \geqslant \widehat{\alpha}(w_j) - \epsilon_1$ holds with a probability of at least $1 - \delta/2$. Hence, we get that $\widehat{\alpha}(w_j) \leqslant 1 - (\frac{1}{2} - \eta_1)C_1 + \epsilon_1$

holds with a probability of at least $1-\delta/2$. If $P(y^j=0|y=0)\leqslant 1/2$, similarly we obtain that $\widehat{\beta}(w_j)\leqslant 1-(\frac{1}{2}-\eta_0)C_0+\epsilon_0$ holds with a probability of at least $1-\delta/2$. Thus, Theorem 6 is proved.

Theorem 6 indicates that the pseudo-sensitivity and pseudo-specificity of any low-quality worker will meet the condition in the theorem with high probability, and further that $\widehat{\alpha}(w_j)$ and $\widehat{\beta}(w_j)$ are both close to and larger than 1/2. It provides an aggressive condition for eliminating all low-quality workers, which is applicable to the setting where there are abundant workers but the low-quality ones will significantly affect the crowd. Furthermore, Theorems 5 and 6 can also be used to eliminate low-sensitivity or low-specificity workers, which is a suitable strategy for the setting where correctly classifying the positive or negative class is more important.

5 Conclusion

The majority voting strategy is widely used in crowdsourcing as it is a good error-pruning method and many reported experimental results on real-world crowdsourcing data sets [3–5] have shown that it performs significantly well at improving the label quality. Our theoretical study is inspired by these real-world observations and provides an analysis of label quality that shows that the label error rate decreases exponentially with the number of workers selected for each task. We also study the problem of eliminating low-quality workers from the crowd. We provide a conservative condition for eliminating low-quality workers without eliminating any non-low-quality worker with high probability as well as an aggressive condition for eliminating all low-quality workers with high probability. The conditions may inspire the development of new algorithms for task assignment and worker selection problems.

The distribution can be generalized to any computational distribution if further prior knowledge about the crowd is known. Many methods have been developed for inferring labels from the crowd, but few theoretical analyses have been presented to support this popular paradigm. Most prior algorithmic and theoretical results on crowdsourcing are based on the assumption that the workers in the crowd are conditionally independent given the class label, e.g., the work in [7–9,11,12,14,16,20,24]. To study crowdsourcing while taking into account the relationships among workers will be an interesting direction for future research.

References

1 Howe J. The rise of crowdsourcing. Wired, 2006, 14.06

2 Howe J. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business New York: Crown Publishing Group, 2008

3 Sheng V S, Provost F J, Ipeirotis P G. Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 614–622

4 Snow R, O'Connor B, Jurafsky D, et al. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, 2008. 254–263

5 Sorokin A, Forsyth D. Utility data annotation with Amazon Mechanical Turk. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop on Internet Vision, Anchorage, 2008. 1–8

6 Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc Ser B, 1977, 39: 1–38

7 Raykar V C, Yu S, Zhao L H, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual International Conference on Machine Learning, Quebec, 2009

8 Raykar V C, Yu S, Zhao L H, et al. Learning from crowds. J Mach Learn Res, 2010, 11: 1297–1322

9 Whitehill J, Ruvolo P, Wu T, et al. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Bengio Y, Schuurmans D, Lafferty J D, et al., eds. Advances in Neural Information Processing Systems 22. Cambridge: MIT Press, 2009. 2035–2043

10 Welinder P, Branson S, Belongie S, et al. The multidimensional wisdom of crowds. In: Lafferty J D, Williams C K I, Shawe-Taylor J, et al., eds. Advances in Neural Information Processing Systems 23. Cambridge: MIT Press, 2010. 2424–2432

11 Raykar V C, Yu S. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. J Mach Learn Res, 2012, 13: 491–518

12 Liu Q, Peng J, Ihler A T. Variational inference for crowdsourcing. In: Bartlett P L, Pereira F C N, Burges C J C, et al., eds. Advances in Neural Information Processing Systems 25. Cambridge: MIT Press, 2012. 701–709

13 Zhou D, Platt J C, Basu S, et al. Learning from the wisdom of crowds by minimax entropy. In: Bartlett P L, Pereira F C N, Burges C J C, et al., eds. Advances in Neural Information Processing Systems 25. Cambridge: MIT Press, 2012. 2204–2212

14 Yan Y, Rosales R, Fung G, et al. Active learning from crowds. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, 2011. 1161–1168

15 Wauthier F L, Jordan M I. Bayesian bias mitigation for crowdsourcing. In: Shawe-Taylor J, Zemel R S, Bartlett P L, et al., eds. Advances in Neural Information Processing Systems 24. Cambridge: MIT Press, 2011. 1800–1808

16 Karger D R, Oh S, Shah D. Iterative learning for reliable crowdsourcing systems. In: Shawe-Taylor J, Zemel R S, Bartlett P L, et al., eds. Advances in Neural Information Processing Systems 24. Cambridge: MIT Press, 2011. 1953–1961

17 Auer P, Cesa-Bianchi N, Freund Y, et al. The nonstochastic multiarmed bandit problem. SIAM J Comput, 2003, 32: 48–77

18 Ho C-J, Vaughan J W. Online task assignment in crowdsourcing markets. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, 2012

19 Buchbinder N, Naor J. Online primal-dual algorithms for covering and packing problems. In: Proceedings of the 13th Annual European Symposium, Palma de Mallorca, 2005. 689–701

20 Ho C-J, Jabbari S, Vaughan J W. Adaptive task assignment for crowdsourced classification. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, 2013. 534–542

21 Chen X, Lin Q, Zhou D. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, 2013. 64–72

22 Dekel O, Shamir O. Vox populi: Collecting high-quality labels from a crowd. In: Proceedings of the 22nd Conference on Learning Theory, Quebec, 2009

23 Tran-Thanh L, Stein S, Rogers A, et al. Efficient crowdsourcing of unknown experts using multi-armed bandits. In: Proceedings of the 20th European Conference on Artificial Intelligence, Montpellier, 2012. 768–773

24 Abraham I, Alonso O, Kandylas V, et al. Adaptive crowdsourcing algorithms for the bandit survey problem. In: Proceedings of the 26th Annual Conference on Learning Theory, Princeton, 2013. 882–910

25 Tsybakov A. Optimal aggregation of classifiers in statistical learning. Ann Stat, 2004, 32: 135–166

Appendix A

Lemma 1. Hoeffding[2] Bound. Let $X_1, \ldots, X_N$ be independent random variables and define the empirical mean of these variables as $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Assume that X_i is bounded, i.e., for $1 \leqslant i \leqslant N$ such that $X_i \in [a_i, b_i]$. The following inequality then holds:

$$P\left(\left|\mathbb{E}(\overline{X}) - \overline{X}\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right).$$

Lemma 2. Let $X_1, \ldots, X_N$ be independent random variables and $X_i \in \{0, 1\}$ for $1 \leqslant i \leqslant N$. Define the empirical mean of these variables as $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then for any $\epsilon, \delta \in (0, 1)$, if $N \geqslant \frac{\ln(2/\delta)}{2\epsilon^2}$, $\left|\mathbb{E}(\overline{X}) - \overline{X}\right| \leqslant \epsilon$ holds with a probability of at least $1 - \delta$.

Proof. Let $\delta = 2\exp(-2N\epsilon^2)$. From Lemma 1, Lemma 2 is proved.

[2] Hoeffding W. Probability inequalities for sums of bounded random variables. J AM Stat Assoc, 1963, 58: 13–30.