Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores

Satoshi Oyama
Hokkaido University
oyama@ist.hokudai.ac.jp

Yukino Baba
The University of Tokyo
yukino_baba@mist.i.u-tokyo.ac.jp

Yuko Sakurai
Kyushu University
ysakurai@inf.kyushu-u.ac.jp

Hisashi Kashima
The University of Tokyo
kashima@mist.i.u-tokyo.ac.jp

Abstract

We have developed a method for using confidence scores to integrate labels provided by crowdsourcing workers. Although confidence scores can be useful information for estimating the quality of the provided labels, a way to effectively incorporate them into the integration process has not been established. Moreover, some workers are overconfident about the quality of their labels while others are underconfident, and some workers are quite accurate in judging the quality of their labels. This differing reliability of the confidence scores among workers means that the probability distributions for the reported confidence scores differ among workers. To address this problem, we extended the Dawid-Skene model and created two probabilistic models in which the values of unobserved true labels are inferred from the observed provided labels and reported confidence scores by using the expectation-maximization algorithm. Results of experiments using actual crowdsourced data for image labeling and binary question answering tasks showed that incorporating workers' confidence scores can improve the accuracy of integrated crowdsourced labels.

1 Introduction

Crowdsourcing on the Web is a promising approach to solving problems that are difficult for computers (but relatively easy for humans). It has thus been extensively studied in various computer science disciplines such as information retrieval, database management, data mining, and machine learning. Typically, a group of people, i.e., a "crowd," is asked to make a judgment regarding given data. The judgments are usually in the form of a binary or multi-class label, a real value, or a short text. Such human judgments, i.e., "annotations," are indispensable for many Web search and data mining tasks such as ranking search results, classifying images, and resolving Web entities. The collection of data annotations through crowdsourcing services, as represented by Amazon Mechanical Turk[1], is becoming a pervasive strategy since a large number of annotations can be collected at relatively low cost.

An inherent problem in applying crowdsourcing is *quality control*. In contrast with well-controlled cases with reliable, screened workers, labels provided by crowdsourcing workers tend to contain many errors due to their varied abilities and dedication levels. Moreover, some crowdsourcing workers, i.e., "spam workers," simply produce random annotations without actually looking at the data in order to earn easy money.

The most straightforward way to make crowdsourced annotations reliable is to obtain multiple annotations from different workers for each data item and then use a simple majority vote to infer the true ones. The implicit assumption with this approach is that all workers have the same probability of making an error. However, in actual crowdsourcing, the probability of making an error varies among workers, so treating the labels given by different workers equally is not an effective approach.

Several methods have been proposed for inferring true labels from worker provided labels that consider the differences in the abilities of workers to provide true labels. In the most well-known method, proposed by Dawid and Skene [1979], each worker is assumed to have a distinct conditional probability of producing his/her label given a (an unknown) true label. They estimated the true labels and the model parameter by using the expectation-maximization (EM) algorithm. Several other methods also consider the difficulty of the task as well as the ability of the workers in inferring the true labels [Whitehill *et al.*, 2009; Welinder *et al.*, 2010a].

The studies mentioned above took a machine-based approach: the label or worker quality is automatically estimated using a statistical inference or machine learning technique. In contrast, we use a human-based approach to determining label quality: the workers are directly asked to report their level of confidence in the labels they provide. Since a worker can easily judge the difficulty of a task and his/her ability to perform it, he/she is the person best suited to evaluate the quality of the label given. Therefore, asking a worker to re-

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

Figure 1: Human intelligence task in which workers are asked to assign confidence scores

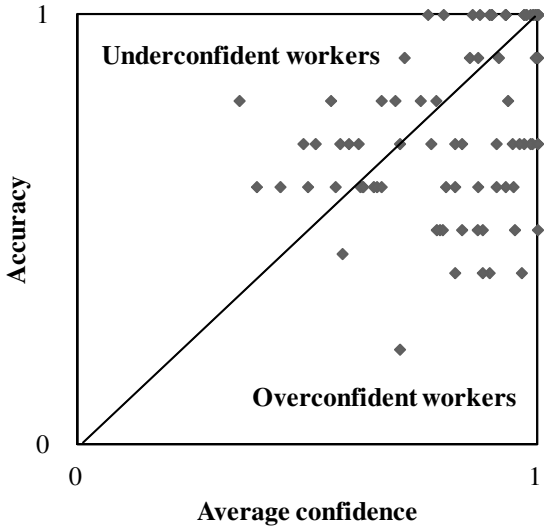

Figure 2: Correlation between accuracy and average confidence of each worker

port the quality of the labels given by the worker is reasonable in the framework of human computation. If the evaluation is done at the time of task completion, the additional burden is not onerous. Workers could be asked to assign a numerical confidence score ranging from 0 to 100 or, more simply, to give a binary response, e.g., "confident" or "not confident." Figure 1 shows an example human intelligence task (HIT) in which the worker assigns a confidence score as a percentage ranging from 0 to 100%.

The possibility of using these confidence scores to improve the quality of crowdsourced labels was investigated by Ipeirotis [2009]. He conducted experiments in which workers were asked to report the task difficulty[2]. He showed that the reported difficulty was correlated with the probability of a correct answer. From this finding, he suggested that asking workers about the difficulty of a labeling task might be a promising alternative to estimating it using a sophisticated algorithm. A similar study was conducted by Kazai [2011].

In line with the direction taken by Ipeirotis [2009], we make use of the confidence scores to improve annotation quality. Although confidence scores given by workers should be useful information for inferring the true labels, a way to effectively incorporate them in an inference algorithm has not been established. In addition, the quality of the reported scores varies among workers just as the label quality does. Some workers may be overconfident and report a high level of confidence even though their labels are actually incorrect, while other workers may be underconfident and report a low level of confidence even though their labels are actually correct. Some workers may be quite accurate in judging their actual abilities, i.e., they are "well-calibrated." And other workers may report a level of confidence in a random manner or without due consideration. Figure 2 shows that there was a positive correlation (≈ 0.455) between worker confidence and accuracy[3]. The data points falling on or near the diagonal line represent "well-calibrated" workers. There are many data points representing "overconfident" or "underconfident" workers, i.e., workers whose confidence scores were not consistent with the actual accuracy. This means that the confidence scores should not be treated equally; instead, the characteristics of the individual workers should be taken into consideration.

In this work, we assume that each worker has a distinct conditional distribution for the confidence scores given the true label and his/her labels. This enables us to model each worker's particular tendency in giving confidence scores, such as an overconfident worker who gives a high confidence score with high probability even when the true label and his/her label are different or an underconfident worker who gives a low confidence score with high probability even when the true label and his/her label are the same.

We propose using two generative probabilistic models in which the crowdsourced label depends on both the true label and the workers' confidence scores. The unobserved true labels are estimated from the observed workers' labels and confidence scores by using the EM algorithm, which alternatingly estimates the true labels and the model parameters.

The organization of the paper is as follows. In Section 2, we describe our problem setting and our proposed probabilistic models, which incorporate worker confidence scores as observable variables. Section 3 describes the method used to infer the true labels given the worker labels and confidence scores. Experimental results for image labeling and binary question answering tasks using real crowdsourced data are given in Section 4. Section 5 discusses related work on inferring true labels or worker abilities in crowdsourced labeling tasks. We summarize the key points and discuss possible future work in Section 6.

[2] A worker's level of confidence in the label is basically opposite the worker's subjective evaluation of the task's difficulty.

[3] This point is discussed in more detail in Section 4.

2 Problem Setting and Proposed Models

2.1 Problem Setting

The problem setting is similar to that of Dawid and Skene [1979]. There are N data items and J crowdsourcing workers (each worker does not necessarily label all items). Let $\mathcal{J}_i \subseteq \{1, \ldots, J\}$ be the subset of workers who labeled item i. $t_i \in \{0, 1\}$ $(i \in \{1, \ldots, N\})$ is the true label for data item i, and $y_{ij} \in \{0, 1\}$ $(j \in \mathcal{J}_i)$ is the label for data i given by worker j. In contrast to the setting of Dawid and Skene [1979], we collect additional information from workers as well as the label estimates. Each worker is asked to assign a confidence score to his/her labels. The level of confidence of worker j in his/her label for item i is given by $c_{ij} \in \{0, 1\}$ $(j \in \mathcal{J}_i)$. If the worker is confident, $c_{ij} = 1$; otherwise, $c_{ij} = 0$. The confidence score is given as a binary variable for simplicity, but the model can be easily extended to enable the use of more general confidence scores, such as multi-level scales and numerical scores.

Our goal is to infer the set of true labels $\{t_i\}(i \in \{1, \ldots, N\})$ given the set of workers' labels $\{y_{ij}\}$ and the set of confidence scores $\{c_{ij}\}$ $(i \in 1, \ldots, N, j \in \mathcal{J}_i)$.

2.2 Proposed Models

We propose using probabilistic generative models of the confidence scores as well as the labels given by crowdsourcing workers. With these models, we can use workers' confidence scores as well as their labels to infer the value of the true labels. For example, if a worker's confidence about his/her label for an item is high, the likelihood that his/her label coincides with the true label is high.

Our models are given as a factorization of the joint distribution:

$$p(\{t_i\}, \{y_{ij}\}, \{c_{ij}\}) = \prod_{i \in \{1,\ldots,N\}} \prod_{j \in \mathcal{J}_i} p(c_{ij}|y_{ij}, t_i) p(y_{ij}|t_i) p(t_i).$$

The value of a true label for item i takes 1 with probability p_i and 0 with probability $1 - p_i$; that is, it is sampled from a Bernoulli distribution,

$$p(t_i) = p_i^{t_i}(1 - p_i)^{(1-t_i)},$$

with parameter p_i.

In the original Dawid-Skene model, the prior probability of the true label is common among different items. This is a reasonable assumption for such areas as medical diagnosis, where the typical question (which is common to all items) is whether a person has had a certain disease and we can assume the prior probability of the disease occurring in the population. In crowdsourcing, however, a task can consist of different kinds of questions, such as, "Is Mount Everest the highest mountain in the world?" and "Is the Nile River longer than the Amazon River?" It is difficult to consider a common prior among the answers to these questions. In addition, the choice of class labels is sometimes arbitrary; for example, asking "Is the Amazon River longer than the Nile River?" instead of the latter question changes the correct answer from "yes" to "no." We therefore introduce parameter p_i, which can be estimated by the rate of workers giving label 1 to item i.

Worker labels $\{y_{ij}|j \in \mathcal{J}_i\}$ for item i are conditionally independent given true label t_i. The $\boldsymbol{\alpha}^{(j)} = \{\alpha_0^{(j)}, \alpha_1^{(j)}\}$ in Figure 3 represents the set of parameters for worker j, where $\alpha_0^{(j)}$ is the probability of worker j giving label 1 if the true label is 0, and $\alpha_1^{(j)}$ is the probability of worker j giving label 1 if the true label is 1. Therefore, when $t_i = 1$, label y_{ij} given by worker j for item i is sampled from a Bernoulli distribution,

$$p(y_{ij}|t_i = 1) = (\alpha_1^{(j)})^{y_{ij}}(1 - \alpha_1^{(j)})^{(1-y_{ij})},$$

with parameter $\alpha_1^{(j)}$. Similarly, when $t_i = 0$, label y_{ij} given by worker j for item i is sampled from a Bernoulli distribution,

$$p(y_{ij}|t_i = 0) = (\alpha_0^{(j)})^{y_{ij}}(1 - \alpha_0^{(j)})^{(1-y_{ij})},$$

with parameter $\alpha_0^{(j)}$.

Worker j's confidence score c_{ij} for his/her label for item i depends on the true label t_i and his/her label y_{ij}, and it is also sampled from a Bernoulli distribution. Our two proposed models, a worker-independent model and a worker dependent model, are variants of the confidence generating model.

In the worker dependent model, $\boldsymbol{\beta}^{(j)} = \{\beta_{00}^{(j)}, \beta_{01}^{(j)}, \beta_{10}^{(j)}, \beta_{11}^{(j)}\}$ is the set of parameters specific to worker j. Here, for example, $\beta_{00}^{(j)}$ is the probability that worker j's confidence $c_{ij} = 1$ when true label $t_i = 0$ and worker j's label $y_{ij} = 0$. In this case, the confidence is sampled from the following distribution.

$$p(c_{ij}|t_i = 0, y_{ij} = 0) = (\beta_{00}^{(j)})^{c_{ij}}(1 - \beta_{00}^{(j)})^{(1-c_{ij})}.$$

When $t_i = 0$ and $y_{ij} = 1$, the confidence is sampled from the following distribution.

$$p(c_{ij}|t_i = 0, y_{ij} = 1) = (\beta_{01}^{(j)})^{c_{ij}}(1 - \beta_{01}^{(j)})^{(1-c_{ij})}.$$

The conditional distributions for the other two cases, $p(c_{ij}|t_i = 1, y_{ij} = 0)$ and $p(c_{ij}|t_i = 1, y_{ij} = 1)$ are similarly defined.

In the worker independent model, all workers are assumed to share the identical parameters $\boldsymbol{\beta} = \{\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}\}(= \boldsymbol{\beta}^{(j)})$. For example, when true label $t_i = 0$ and worker j's label $y_{ij} = 0$, the confidence is sampled from the following distribution (common to all workers).

$$p(c_{ij}|t_i = 0, y_{ij} = 0) = (\beta_{00})^{c_{ij}}(1 - \beta_{00})^{(1-c_{ij})}.$$

The conditional distributions for the remaining cases, $p(c_{ij}|t_i = 0, y_{ij} = 1)$, $p(c_{ij}|t_i = 1, y_{ij} = 0)$, and $p(c_{ij}|t_i = 1, y_{ij} = 1)$, are defined similarly.

The worker dependent model is not based on such an assumption. Introducing worker specific distributions for the confidence values enables more flexible worker modeling, so the model captures the different tendencies among workers in reporting their confidence, i.e., overconfident workers reporting high confidence even when their labels are incorrect and underconfident workers reporting low confidence even when their labels are correct.

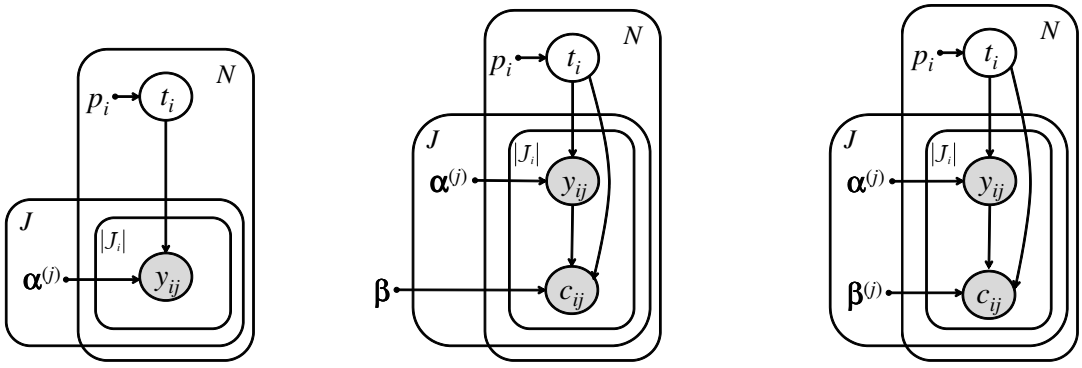

Figure 3: Graphical models for label integration: (a) Dawid-Skene model, (b) worker independent confidence model, and (c) worker dependent confidence model

The differences among the Dawid-Skene model, our worker independent confidence model, and our worker dependent confidence model are illustrated in the graphical models given in Figure 3. We can see that our models extend the Dawid-Skene model by introducing worker confidence scores as probabilistic variables.

3 Inference Algorithm

Given labels $\{y_{ij}\}$ and confidence scores $\{c_{ij}\}$ obtained from the workers, we want to estimate true labels $\{t_i\}$. Similar to the approach of Dawid and Skene [1979], we use the EM algorithm to obtain the maximum likelihood estimate of model parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$, with true labels $\{t_i\}$ as latent variables.

We first give the EM algorithm for the worker dependent confidence model since it also gives that for the worker independent confidence model as a special case. The EM algorithm for the worker dependent model alternately performs two steps until convergence.

E-step: Estimate the expected values of unobserved variables $\{t_i\}$ by using the current estimates of parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$.

M-step: Estimate parameters $\{\boldsymbol{\alpha}^{(j)}\}$ and $\{\boldsymbol{\beta}^{(j)}\}$ by using the current expectations of $\{t_i\}$.

In the E-step, the expectation of t_i is represented as

$$\begin{aligned} E[t_i] &= p(t_i = 1|\{y_{ij}\}, \{c_{ij}\}) \\ &= \frac{p_i}{z_i} \prod_{j \in \mathcal{J}_i} \Big\{ (\alpha_1^{(j)})^{y_{ij}} (1 - \alpha_1^{(j)})^{(1-y_{ij})} \\ &\times (\beta_{11}^{(j)})^{y_{ij} c_{ij}} (1 - \beta_{11}^{(j)})^{y_{ij}(1-c_{ij})} \\ &\times (\beta_{10}^{(j)})^{(1-y_{ij}) c_{ij}} (1 - \beta_{10}^{(j)})^{(1-y_{ij})(1-c_{ij})} \Big\}, \end{aligned} \tag{1}$$

where z_i is an unknown normalization constant. Computing the expectation also requires evaluating

$$\begin{aligned} 1 - E[t_i] &= p(t_i = 0|\{y_{ij}\}, \{c_{ij}\}) \\ &= \frac{1 - p_i}{z_i} \prod_{j \in \mathcal{J}_i} \Big\{ (\alpha_0^{(j)})^{y_{ij}} (1 - \alpha_0^{(j)})^{(1-y_{ij})} \\ &\times (\beta_{01}^{(j)})^{y_{ij} c_{ij}} (1 - \beta_{01}^{(j)})^{y_{ij}(1-c_{ij})} \\ &\times (\beta_{00}^{(j)})^{(1-y_{ij}) c_{ij}} (1 - \beta_{00}^{(j)})^{(1-y_{ij})(1-c_{ij})} \Big\}. \end{aligned} \tag{2}$$

Solving these two equations (given that $E[t_i] + (1 - E[t_i]) = 1$) to obtain the value of z_i, we obtain the value of $E[t_i]$.

In the M-step, the maximum likelihood estimates of $\alpha_\mu^{(j)} (\mu \in \{0, 1\})$ are respectively computed by using

$$\begin{aligned} \hat{\alpha}_0^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i]) y_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i])} \\ \hat{\alpha}_1^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i] y_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i]} \quad . \end{aligned} \tag{3}$$

$\boldsymbol{\beta}^{(j)}$ is estimated by using

$$\begin{aligned} \hat{\beta}_{00}^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i])(1 - y_{ij}) c_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i])(1 - y_{ij})} \\ \hat{\beta}_{01}^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i]) y_{ij} c_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} (1 - E[t_i]) y_{ij}} \\ \hat{\beta}_{10}^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i](1 - y_{ij}) c_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i](1 - y_{ij})} \\ \hat{\beta}_{11}^{(j)} &= \frac{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i] y_{ij} c_{ij}}{\sum_{\{i:j \in \mathcal{J}_i\}} E[t_i] y_{ij}}. \end{aligned} \tag{4}$$

In practice, we often face the "zero frequency problem." For example, if we have no data such that worker j gives label 1 with confidence 0 to an instance whose true label is 1, the maximum likelihood estimation gives 0 probability to the event. Therefore, we use the Laplace smoothing technique to

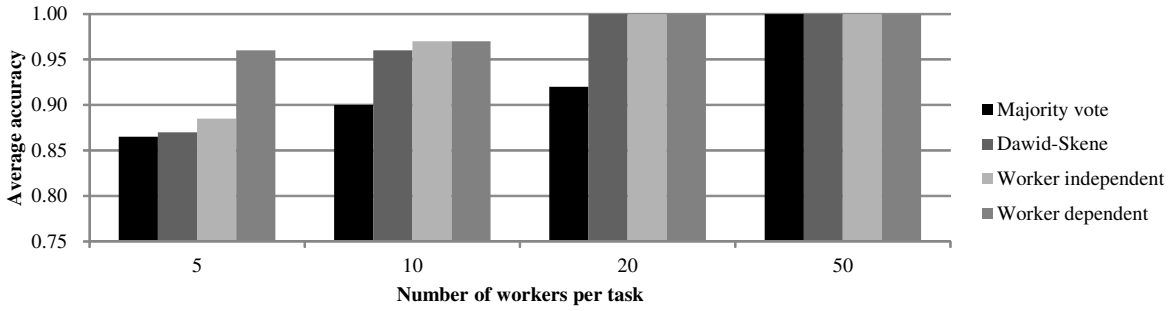

Figure 4: Results for image labeling

give virtual occurrence counts to such unobserved events:

$$\hat{\beta}_{00}^{(j)} = \frac{\sum_{\{i:j\in\mathcal{J}_i\}}(1-E[t_i])(1-y_{ij})c_{ij}+1}{\sum_{\{i:j\in\mathcal{J}_i\}}(1-E[t_i])(1-y_{ij})+2}.$$

The EM algorithm for the worker independent model is given by simply changing $\boldsymbol{\beta}^{(j)}$ to $\boldsymbol{\beta}$ in Eqs. (1), (2), and (4) and by changing the domains of summation from $\{i : j \in \mathcal{J}_i\}$ to $\{i, j : 1 \leq i \leq N, 1 \leq j \leq J\}$ in Eq. (4).

4 Evaluation

4.1 Image Labeling

To evaluate the effectiveness of using confidence scores in inferring true labels, we conducted experiments using Amazon Mechanical Turk. We chose ten images from the Caltech-UCSD Birds 200 dataset [Welinder *et al.*, 2010b] and asked crowdsourcing workers to choose one of two bird names as the label for each image. We also asked them to report their level of confidence in each choice. To determine the effect of the number of workers on the accuracy of the inferred labels, we asked 100 workers to label the same ten images. Although each worker labeled all ten images, there were some missing values due to input errors. The workers were instructed to report their confidence level for each answer by entering a numeric value ranging from 0 to 100.

Since our current model is based on the assumption that the confidence score is in binary form, we had to convert the confidence scores into binary form. We first found the median score for each worker and then converted his/her confidence scores greater than the median to 1 and the ones smaller than the median to 0. The confidence scores equal to the median were converted to either 1 or 0 so that the number of confidence scores of 1 and of 0 became better balanced.

To see the effect of the number of workers per task on accuracy, we split the workers into groups of equal size, inferred the true labels from the worker labels and confidence scores within each group, and averaged the accuracies of the true labels obtained from each group. We conducted experiments with four different group sizes: 5, 10, 20, and 50. The average accuracies for each group size were obtained with majority vote, the Dawid-Skene model, the worker-independent confidence model, and the worker-dependent confidence model.

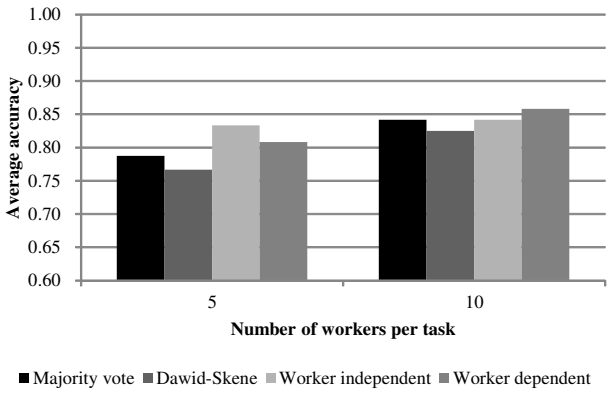

Figure 5: Results for binary question answering

As shown in Figure 4, with 50 workers, all three models and even a simple majority vote provided sufficient accuracy due to the high level of redundancy. In practice, however, the number of workers that can be used for a task is limited due to cost. Ten workers at most for each task would be a reasonable number. As shown in Figure 4, when the number of workers was 5 or 10, the two models using the confidence score achieved better accuracy than majority vote and the Dawid-Skene model. In particular, for the case of 5 workers per task, the worker-dependent confidence model greatly outperformed the other models.

4.2 Binary Question Answering

We conducted experiments using another dataset, one containing 120 binary questions on general knowledge, e.g. "Is Mount Everest the highest mountain in the world?" We used Lancers crowdsourcing service[4] to collect answers for this dataset. Each question was answered by ten crowdsourcing workers. Again, along with the answers, we also asked them to assign confidence scores. The confidence scores were processed in exactly the same way as in the image labeling experiments. We collected the confidence scores, which ranged from 0 and 100, and converted them into binary form.

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

In this case, totally 42 workers involved in the tasks. As shown in Figure 5, when we used all ten of the workers' labels, the accuracy of simple majority vote was similar to that of the two models using the confidence score. As we saw in the results for image labeling, majority vote can achieve sufficient accuracy when the number of labels per item is relatively large. In contrast, when we used the labels of half the workers (average number of labels per item is five), the worker-independent confidence model achieved the best accuracy, followed by the worker-dependent confidence model. The better performance of the worker-independent model is attributed to the fact that the variations in the confidence scores given by the workers were smaller than those for the image labeling experiments. The standard deviation of the average scores for image labeling was 0.22 while that for binary question answering was 0.16.

5 Related Work

5.1 Quality Control in Crowdsourcing

One of the fundamental challenges in crowdsourcing is controlling the quality of the obtained data. Crowdsourcing workers are rarely trained and do not necessarily have adequate ability to complete their assigned tasks accurately [Snow *et al.*, 2008]. There are also great differences in the skill levels of such workers. A particular problem is malicious behavior by spam workers [Eickhoff and de Vries, 2011]. These workers are motivated by financial reward and thus complete the task as quickly as possible with minimum effort, resulting in worthless submissions.

Promising approaches for quality control can be categorized into task design [Kittur *et al.*, 2008], worker filtering, and inter-agreement metrics for multiple submissions. A widely used approach is to obtain multiple submissions from different workers and aggregate them by applying a majority vote [Sheng *et al.*, 2008] or other rules. Dawid and Skene [1979] addressed the problem of aggregating medical diagnoses from multiple doctors to improve decision accuracy. Smyth *et al*. [1995] applied this method to the problem of inferring true labels for images from multiple noisy labels. Whitehill *et al*. [2009] explicitly modeled the difficulty of each task, and Welinder *et al*. [2010a] introduced the idea of evaluating the difficulty of each task differently for each worker.

A number of researchers in the machine learning and data mining communities have addressed the problem of supervised learning from multiple labels obtained from crowdsourcing workers [Sheng *et al.*, 2008]. Raykar *et al*. [2010] extended the Dawid-Skene model to enable inferring both the true labels and predictive models simultaneously. Yan *et al*. [2010] presented a model in which the error rate of the workers is assumed to depend on the task. Kajino *et al*. [2012] proposed a convex optimization formulation for learning from crowds. Other research has contributed to labeler selection in the contexts of repeated trials [Donmez *et al.*, 2009], active learning [Yan *et al.*, 2011], and clustering [Gomes *et al.*, 2011].

5.2 Use of Confidence Scores in Quality Control

We assume that confidence scores are useful information for estimating the reliability of labels given by workers; however, this assumption is not especially novel. Ipeirotis [2009] conducted an experiment to examine the correlation between the self-reported difficulty of tasks and the probability of correct answers. Kazai [2011] investigated the relationship between worker confidence and label quality in the context of document relevance assessment. Branson *et al*. [2010] collected three-level confidence scores for visual recognition tasks. However, these efforts did not include the use of confidence scores for quality control purpose, and they did not investigate the usefulness of the scores.

In Kazai's study, each worker was asked to rate his/her familiarity with the given topic and the task difficulty. The results showed that workers who rated the task easier had higher accuracy. However, the workers who claimed to be an expert had less accuracy than the ones who did not. Karzai postulated that veteran workers would be better at measuring their expertise than amateur workers and that the less confident workers would be more likely to take more time completing the tasks.

While Ipeirotis and Kazai investigated the correlation between worker confidence and the reliability of the output, to the best of our knowledge, we are the first to propose a method for utilizing the level of confidence to improve the quality of crowdsourced labels. Studies using item response theory [de Ayala, 2009] have used tests consisting of multiple choice questions for which each examinee was asked to express his/her degree of confidence that the response was correct [de Finetti, 1965]. The assumptions made in item response theory differ significantly from those in crowdsourcing; the true answers are known to the test provider in item response theory while they are unknown in crowdsourcing.

6 Conclusion

We propose utilizing crowdsourcing workers' confidence scores to integrate their answers and to infer the unobserved true labels. We extended the Dawid-Skene model to incorporate confidence scores into the label generation process; that is, workers' confidence levels scores depend on the unobserved true labels and workers' labels. We devised an EM-based algorithm for estimating the model parameters and true labels. The experimental results showed that incorporating workers' confidence scores can improve the accuracy of integrated crowdsourced labels, especially when the number of workers that can be used for a task is limited.

One possible future direction is to design effective ways of asking workers to assign confidence scores. In item response theory [de Ayala, 2009], examinees are sometimes asked to assign a confidence distribution to the response they give to each question [de Finetti, 1965]. Kato and Zhang [2010] reported that more than 60% of the confidence scores were either '0%' or '100%' in their experiments and suggested adding a pre-training phase for teaching examinees how to better represent their confidence. How to ask workers to assign confidence scores is an important research question. Although assigning a confidence score for each response is rel-

atively easy, it is still an additional burden. Replacing the use of confidence scores with another metric that is automatically measurable (such as the time needed for completing a task) is another possible direction worth studying.

Acknowledgments

SO was supported by JSPS KAKENHI 24650061. YB and HK were supported by the FIRST program. YS was supported by the JST PRESTO program.

References

[Branson *et al.*, 2010] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual Recognition with Humans in the Loop. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, 2010.

[Dawid and Skene, 1979] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statics)*, 28(1):20–28, 1979.

[de Ayala, 2009] R. J. de Ayala. *The Theory and Practice of Item Response Theory*. The Guilford Press, 2009.

[de Finetti, 1965] Bruno de Finetti. Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item. *British Journal of Mathematical and Statistical Psychology*, 18(1):87–123, 1965.

[Donmez *et al.*, 2009] Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[Eickhoff and de Vries, 2011] Carsten Eickhoff and Arjen de Vries. How Crowdsourcable is Your Task. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, 2011.

[Gomes *et al.*, 2011] Ryan Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In *Advances in Neural Information Processing 24*, 2011.

[Ipeirotis, 2009] Panos Ipeirotis. How good are you, Turker?, 2009. http://www.behind-the-enemy-lines.com/2009/01/how-good-are-you-turker.html.

[Kajino *et al.*, 2012] Hiroshi Kajino, Yuta Tsuboi, and Hisashi Kashima. A Convex Formulation for Learning from Crowds. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*, 2012.

[Kato and Zhang, 2010] Kentaro Kato and Yiping Zhang. An Item Response Model for Probability Testing. In *International Meeting of the Psychometric Society*, 2010.

[Kazai, 2011] Gabriella Kazai. In Search of Quality in Crowdsourcing for Search Engine Evaluation. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR)*, 2011.

[Kittur *et al.*, 2008] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the 26th International Conference on Human Factors in Computing Systems (CHI)*, 2008.

[Raykar *et al.*, 2010] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from Crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[Sheng *et al.*, 2008] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

[Smyth *et al.*, 1995] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring Ground Truth from Subjective Labelling of Venus Images. In *Advances in Neural Information Processing Systems 7*, 1995.

[Snow *et al.*, 2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 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 (EMNLP)*, 2008.

[Welinder *et al.*, 2010a] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems 23*, 2010.

[Welinder *et al.*, 2010b] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22*, 2009.

[Yan *et al.*, 2010] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer G. Dy. Modeling Annotator Expertise: Learning When Everybody Knows a Bit of Something. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[Yan *et al.*, 2011] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer G. Dy. Active Learning from Crowds. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.