Interactive Learning with Proactive Cognition Enhancement for Crowd Workers

Jing Zhang,[1]* **Huihui Wang,**[1] **Shunmei Meng,**[1] **Victor S. Sheng**[2]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
200 Xiaolingwei Street, Nanjing 210094, China
[2]Department of Computer Science, Texas Tech University
2500 Broadway, Lubbock, TX 79409, U.S.A.
{jzhang, huihuiwang, mengshunmei}@njust.edu.cn, victor.sheng@ttu.edu

Abstract

Learning from crowds often performs in an active learning paradigm, aiming to improve learning performance quickly as well as to reduce labeling cost by selecting proper workers to (re)label critical instances. Previous active learning methods for learning from crowds do not have any proactive mechanism to effectively improve the reliability of workers, which prevents to obtain steadily rising learning curves. To help workers improve their reliability while performing tasks, this paper proposes a novel Interactive Learning framework with Proactive Cognitive Enhancement (ILPCE) for crowd workers. The ILPCE framework includes an interactive learning mechanism: When crowd workers perform labeling tasks in active learning, their cognitive ability to the specific domain can be enhanced through learning the exemplars selected by a psychological model-based machine teaching method. A novel probabilistic truth inference model and an interactive labeling scheme are proposed to ensure the effectiveness of the interactive learning mechanism and the performance of learning models can be simultaneously improved through a fast and low-cost way. Experimental results on three real-world learning tasks demonstrate that our ILPCE significantly outperforms five representative state-of-the-art methods.

Introduction

Crowdsourcing has become a popular solution to acquire labels for machine learning because it is convenient, fast, and costs less. However, one non-negligible flaw of this manner is that the quality of labels collected from crowds is not always high enough to train good models. Due to the varying levels of expertise of workers, incorrect labels (i.e., label noises) are spread throughout the dataset. Therefore, quality control becomes one of the most critical issues in crowdsourcing studies (Daniel et al. 2018). Crowdsourcing platforms usually would regulate the behaviors of workers via various mechanisms (such as user interface design (Retelny et al. 2014), reputation inspection and ranking (Irani and Silberman 2013), and incentives mechanism (Yang et al. 2016)) so that they are willing or might be forced to provide high-quality answers. However, these methods usually only have a moderate effect, and neither can be adjusted to specific tasks nor ensure that errors dismiss in the results.

Another train of thought is to utilize some data integration methods to improve data quality, which has been widely adopted in machine learning related communities. These methods employ a redundancy mechanism to increase the quality of labels, namely, *repeated labeling* scheme (Ipeirotis et al. 2014). It allows an instance to be labeled by different crowd workers to obtain multiple noisy labels and then infers the true labels for these instances. Over the past decade, researchers proposed numerous truth inference algorithms (Zheng et al. 2017), demonstrated as one of the most important branches of crowdsourcing study. In addition to obtaining true labels, an inference algorithm may also estimate the other aspects of the labeling process, such as the difficulty of instances and the reliability of workers. Having obtained the integrated (inferred) labels of instances, we can use the dataset to train learning models, forming a two-stage (i.e., "inference plus model training") learning paradigm.

Although the repeated labeling scheme significantly improves the quality of labels, it still faces two difficulties: (**1**) the labeling cost sharply increases for multiple queries on each instance; and (**2**) it cannot guarantee the stability of the quality of answers provided by the crowd workers. For the first difficulty, we can resort to active learning (Settles 2009), which can reduce labeling cost through the design of sampling strategies that select the instances potentially contributing the most to current learning models. Compared with the traditional instance selection strategies (Fu, Zhu, and Li 2013), active learning strategies in crowdsourcing may also consider the distributions of noisy labels and some inferred information such as the reliability of workers (Sheng, Provost, and Ipeirotis 2008; Yan et al. 2011; Rodrigues, Pereira, and Ribeiro 2014).

However, we have noticed that, compared with the traditional active learning, active learning from crowds can hardly maintain a smooth rising learning curve, which was observed from the experimental results in many previous studies (Sheng, Provost, and Ipeirotis 2008; Yan et al. 2011; Rodrigues, Pereira, and Ribeiro 2014; Zhang, Wu, and Shengs 2015). This phenomenon not only causes the performance of learning models can hardly being improved but

*Jing Zhang is the corresponding author of the paper.

also requires more budget to acquire noisy labels. One of the reasons is that the workers probably make mistakes repeatedly because of lacking expertise. Some efforts have been made to address this issue by training the crowd workers with domain knowledge in advance (Singla et al. 2014; Amir et al. 2016; Mac Aodha et al. 2018). These *pre-training* methods are usually conducted before the task begins so that their effects are possibly undesirable because the workers may quickly forget the critical knowledge that they have learned, or the pre-trained workers may even refuse to attend the crowdsourcing tasks.

To systematically address the above issues, this paper proposes a novel interactive learning framework that proactively enhances the cognitive ability of crowd workers while they are performing labeling tasks under the active learning process. The contributions of this paper are three-fold:

- We propose a novel interactive learning mechanism. Based on the classic Generalized Context Model in psychology, we use a machine learning method to select a set of exemplars from an exemplar pool for workers to learn when performing the labeling tasks. The mechanism can enhance the cognition of workers to a specific domain, resulting in the improvement of their reliability.
- We propose a novel Bayesian inference model, which infers the true labels and difficulties of instances as well as the reliability of workers. All the inferred information will be used to form the above interactive learning mechanism.
- We also design a novel active learning process (including structures, learning strategies, and interactive scheme) to implement the proposed interactive learning mechanism. Experimental results on three learning tasks demonstrate the advantages of the techniques proposed in the paper.

This paper demonstrates that the machine intelligence and human intelligence can promote each other and develop together through ingenious design.

Related Work

Active learning from crowds The open, dynamic, and budget-limited features of crowdsourcing make it a natural choice to adopt an active learning paradigm. In crowdsourcing settings, active learning usually involves relabeling the instances that were previously labeled (Sheng, Provost, and Ipeirotis 2008; Yan et al. 2011; Rodrigues, Pereira, and Ribeiro 2014; Lin, Mausam, and Weld 2016) because the learning models are sensitive to incorrect labels. Furthermore, active learning strategies are more complicated than traditional ones (Settles 2009). They need to comprehensively consider three aspects of noisy labels, instances, and workers when forming the instance and worker selection strategies that can optimize the learning performance. This paper only considers simple strategies, especially ignoring worker selection, since one primary goal is to investigate the cognitive ability enhancement of crowd workers in general.

Truth inference To estimate the true labels of training instances, a truth inference algorithm is applied to their collected noisy labels. Besides this main function, an inference algorithm may model some other critical features of labeling

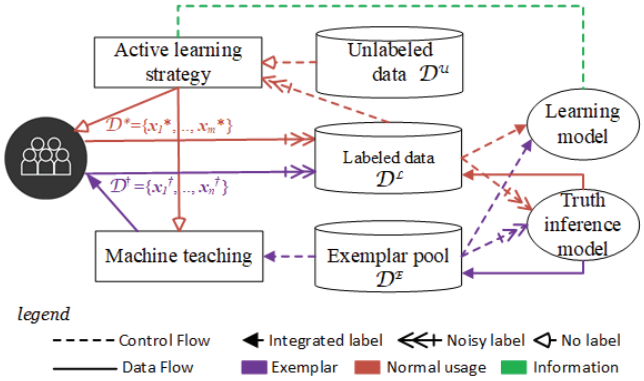

Figure 1: The proposed interactive learning framework

systems. For example, RY (Raykar et al. 2010) models the specificity and sensitivity of labeling bias of workers. GLAD (Whitehill et al. 2009) models the reliability of workers and the difficulties of tasks. Bi et al. (2014) added the dedication of workers into the inference model, and further Kurve, Miller, and Kesidis (2015) added the intention of workers, which can distinguish malicious workers from normal ones. All this information may be utilized in the design of active learning strategies. Like many other mainstream probabilistic models, the inference algorithm proposed in this paper models the difficulties of tasks using real numbers as well as the reliability of workers using confusion matrix.

Teaching crowd workers Teaching crowd workers with domain knowledge is a straightforward way of improving their reliability (Singla et al. 2014; Servajean et al. 2016; Amir et al. 2016; Mac Aodha et al. 2018; Zhou, Nelakurthi, and He 2018). None of the above studies perform a two-directional optimization on the performance of learning models and teaching models in an interactive environment. Furthermore, because human learning involves a complex cognitive psychological process, if we have modeled the cognition of crowd workers, we can seek approaches to use the minimum exemplars to train their cognition models, which is one of the typical application scenarios of machine teaching (Zhu 2015). For example, Patil et al. (2014) demonstrated their first attempt to use the Generalized Context Model (Nosofsky 2011) in machine teaching, which improves human learners. This paper unifies learning model training and cognitive modeling-based machine teaching into an interactive crowdsourcing learning framework.

The Proposed Method

This section first presents our interactive learning framework and then describes each technical detail in the framework.

Interactive Learning Framework

The proposed novel interactive learning framework is shown in Figure 1. There are three data sources in the framework: unlabeled data ($\mathcal{D}^{\mathcal{U}}$), labeled data ($\mathcal{D}^{\mathcal{L}}$), and an exemplar pool ($\mathcal{D}^{\mathcal{E}}$). Instances in both $\mathcal{D}^{\mathcal{L}}$ and $\mathcal{D}^{\mathcal{E}}$ have obtained multiple noisy labels. The difference is that the instances in $\mathcal{D}^{\mathcal{E}}$ serve as the exemplars for machine teaching.

Thus, their integrated labels have a high probability of being correct and their contents are easy to understand for humans. Different from the previous studies (Singla et al. 2014; Mac Aodha et al. 2018) that used ground truth to teach humans, our solution is completely agnostic, which generates exemplars for human learners with the help of the ground truth inference algorithms.

The truth inference has two functions. First, it estimates the true labels from the crowdsourced labels so that each instance will be assigned an integrated label. Second, it models the reliability of workers and the difficulties of instances, which is used to identify the instances that can be exemplars. After inference, the instances with integrated labels, denoted by $\mathcal{D}^{\mathcal{O}} = \mathcal{D}^{\mathcal{E}} \cup \mathcal{D}^{\mathcal{L}}$, is used to train learning models. Taking the current learning model into account, the active learning strategy selects m instances, denoted by $\mathcal{D}^* = \{\boldsymbol{x}_i^*\}_{i=1}^m$, based on their representativeness and uncertainty. Instances $\mathcal{D}^*$ are pushed forward to crowd workers to acquire labels and are also fed into a machine teaching algorithm.

When the machine teach algorithm receives $\mathcal{D}^*$, it selects n exemplars with integrated labels from the exemplar pool, denoted by $\mathcal{D}^\dagger = \{< \boldsymbol{x}_i^\dagger, \hat{y}_i^\dagger >\}_{i=1}^n$. The exemplars may belong to different classes and are considered to be the most helpful to workers' cognition. On the interface of human intelligence tasks (HITs), both $\mathcal{D}^*$ and $\mathcal{D}^\dagger$ are shown simultaneously. Workers label $\mathcal{D}^*$ and *optionally relabel* some items in $\mathcal{D}^\dagger$ if they think that the integrated labels of those items are incorrect. When the newly noisy labeled data are collected, we perform the truth inference again and then update the exemplar pool and the current learning model. Thus, another novelty of our interactive learning is that exemplars are constantly changing as labeling tasks progresses.

Bayesian Truth Inference

We propose a novel Bayesian truth inference for multi-class annotation, which can models the difficulty of instances and the reliability of workers.

Problem statement The dataset with crowdsourced labels is denoted by $\mathcal{D}^{\mathcal{O}} = \{< \boldsymbol{x}_i, y_i, \boldsymbol{l}_i >\}_{i=1}^I$, where $\boldsymbol{x}_i$, y_i and $\boldsymbol{l}_i$ are the feature portion, unknown class label, and noisy label set of instance $\boldsymbol{x}_i$, respectively. Suppose totally J workers label the instances. That is, we have $\boldsymbol{l}_i = \{l_{ij}\}_{j=0}^J$, where $l_{ij} \in \{0, 1, ..., K\}$. Here, $l_{ij} = k (1 \leq k \leq K)$ means that worker j labels instance $\boldsymbol{x}_i$ as class k and $l_{ij} = 0$ means that the worker does not provide any label. All crowdsourced labels form a matrix $L^{I \times J}$. The truth inference aims to assign each instance $\boldsymbol{x}_i$ an integrated label $\hat{y}_i$ that is inferred from the crowdsourced labels and minimize the empirical errors:

$$err = \min \left\{ \frac{1}{I} \sum\nolimits_{i=1}^{I} \mathbb{I}(\hat{y}_i \neq y_i) \right\}, \text{given } L, \qquad (1)$$

where $\mathbb{I}$ is an indicator function.

Bayesian Inference Model (BIM) The probabilistic graphic representation of the proposed Bayesian inference model is shown in Figure 2.

(**1**) *Modeling reliability of crowd workers*. We model the reliability of worker j with a confusion matrix $\Pi^{(j)} = \{\pi_{kl}^{(j)}\}, (1 \leq k, l \leq K)$, where $\pi_{kl}^{(j)}$ represents that the probability of worker j labeling (true) class k as class l. Thus, the reliability of worker j can be defined using the trace of the matrix as follows:

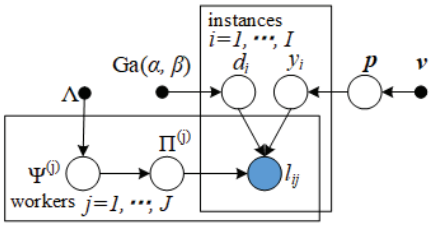

Figure 2: The Bayesian inference model

$$r^{(j)} = \text{Tr}(\Pi^{(j)})/K. \qquad (2)$$

Under a Bayesian probabilistic framework, a row of confusion matrix $\boldsymbol{\pi}_k^{(j)}$ of worker j is generated from a Dirichlet distribution with parameters $\boldsymbol{\psi}_k^{(j)} = [\psi_{k1}^{(j)}, \psi_{k2}^{(j)}, ..., \psi_{kK}^{(j)}]$:

$$p(\boldsymbol{\pi}_k^{(j)} | \boldsymbol{\psi}_k^{(j)}) = \frac{\Gamma(\sum_{l=1}^K \psi_{kl}^{(j)})}{\prod_{l=1}^K \Gamma(\psi_{kl}^{(j)})} \prod_{l=1}^K (\pi_{kl}^{(j)})^{\psi_{kl}^{(j)} - 1}, \qquad (3)$$

where Γ is the gamma function. The prior distribution of variable $\psi_{kl}^{(j)}$ is modeled by an exponential distribution with hyper-parameters λ_{kl} as follows:

$$p(\psi_{kl}^{(j)} | \lambda_{kl}) = \frac{1}{\lambda_{kl}} \exp\Big(- \frac{\psi_{kl}^{(j)}}{\lambda_{kl}} \Big). \qquad (4)$$

Then, we can calculate the joint probability of confusion matrices $\boldsymbol{\Pi} = \{\Pi^{(j)}\}_{j=1}^J$ and their parameters $\boldsymbol{\Psi} = \{\Psi^{(j)}\}_{j=1}^J$ with hyper-parameters $\boldsymbol{\Lambda} = \{\lambda_{kl}\}_{k,l=1}^K$ as follows:

$$p(\boldsymbol{\Pi} | \boldsymbol{\Psi}) = \prod_{j=1}^J p(\Pi^{(j)} | \Psi^{(j)}) = \prod_{j=1}^J \prod_{k=1}^K p(\boldsymbol{\pi}_k^{(j)} | \boldsymbol{\psi}_k^{(j)}), \quad (5)$$

$$p(\boldsymbol{\Psi} | \boldsymbol{\Lambda}) = \prod_{j=1}^J p(\Psi^{(j)} | \boldsymbol{\Lambda}) = \prod_{j=1}^J \prod_{k=1}^K \prod_{l=1}^K p(\psi_{kl}^{(j)} | \lambda_{kl}). \quad (6)$$

(**2**) *Modeling true labels*. We assume that the true label y_i of instance $\boldsymbol{x}_i$ is generated from a Dirichlet distribution with parameters $\boldsymbol{p} = [p_1, p_2, ..., p_K]$. Also, the prior for these class proportions can be set to a Dirichlet distribution with hyper-parameters $\boldsymbol{\nu} = [\nu_1, \nu_2, ..., \nu_K]$ as follows:

$$p(\boldsymbol{p} | \boldsymbol{\nu}) = \frac{\Gamma(\sum_{k=1}^K \nu_k)}{\prod_{k=1}^K \Gamma(\nu_k)} \prod_{k=1}^K p_k^{\nu_k - 1}. \qquad (7)$$

(**3**) *Modeling difficulties of instances*. We model the difficulty of instance $\boldsymbol{x}_i$ using a real number as follows:

$$DF(\boldsymbol{x}_i) = 1/d_i, \qquad (8)$$

where real number d_i is generated from a Gamma distribution with hyper-parameters (α, β). We assume that $\mathbf{d} = \{d_1, d_2, ..., d_I\}$ is i.i.d. Then, we have

$$p(\boldsymbol{d} | \alpha, \beta) = \prod_{i=1}^I p(d_i | \alpha, \beta) = \prod_{i=1}^I \frac{d_i^{(\alpha-1)} \beta^\alpha e^{-x\beta}}{\Gamma(\alpha)}. \qquad (9)$$

The difficulty $1/d_i$ of instance $\boldsymbol{x}_i$ can be applied to the reliability of worker j as the form of $(\pi_{kl}^{(j)})^{d_i}$. Then, for this instance $\boldsymbol{x}_i$ with true label y_i, the probability of worker j providing label l_{ij} will change to $(\pi_{y_i l_{ij}}^{(j)})^{d_i}/\sum_{q=1}^{K}(\pi_{y_i q}^{(j)})^{d_i}$.

(**4**) *Overall posterior probability density and Gibbs sampling solution*. Given the observed labels L, the posterior probability density of all random variables is

$$\begin{aligned}
&p(\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{p}, \boldsymbol{\Pi}, \boldsymbol{\Psi} | L, \boldsymbol{\nu}, \boldsymbol{\Lambda}, \alpha, \beta) \\
&\propto p(L|\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{p}, \boldsymbol{\Pi}, \boldsymbol{\Psi}) p(\boldsymbol{y}|\boldsymbol{p}) p(\boldsymbol{p}|\boldsymbol{\nu}) p(\boldsymbol{d}|\alpha, \beta) p(\boldsymbol{\Pi}|\boldsymbol{\Psi}) p(\boldsymbol{\Psi}|\boldsymbol{\Lambda}) \\
&\propto p(\boldsymbol{p}|\boldsymbol{\nu}) p(\boldsymbol{\Psi}|\boldsymbol{\Lambda}) \prod_{i=1}^{I} \left\{ p_{y_i} \frac{d_i^{(\alpha-1)} \beta^{\alpha} e^{-x\beta}}{\Gamma(\alpha)} \prod_{j=1}^{J} \frac{(\pi_{y_i l_{ij}}^{(j)})^{d_i}}{\sum_{q=1}^{K} (\pi_{y_i q}^{(j)})^{d_i}} \right\}.
\end{aligned} \tag{10}$$

Variables in Eq. (10) can be solved by Gibbs sampling as long as we perform the sampling operations on the density functions below until it converges:

$$p(\boldsymbol{p}|rest) \propto \prod_{k=1}^{K} p_k^{\{\sum_{i=1}^{I} \mathbb{I}(y_i=k)\} + \nu_k - 1}, \tag{11}$$

$$p(y_i = k|rest) \propto p_k \prod_{j=1}^{J} \frac{(\pi_{y_i l_{ij}}^{(j)})^{d_i}}{\sum_{q=1}^{K} (\pi_{y_i q}^{(j)})^{d_i}}, \tag{12}$$

$$p(\boldsymbol{\pi}_k^{(j)}|rest) \propto \prod_{l=1}^{K} (\pi_{kl}^{(j)})^{\{\sum_{i=1}^{I} \mathbb{I}(y_i=k, l_{ij}=l)\} + \psi_{kl}^{(j)} - 1}, \tag{13}$$

$$p(\boldsymbol{\psi}_k^{(j)}|rest) \propto \frac{\Gamma(\sum_{q=1}^{K} \psi_{kq}^{(j)})}{\Gamma(\psi_{kl}^{(j)})} (\pi_{kl}^{(j)}) \exp(-\frac{\psi_{kl}^{(j)}}{\lambda_{kl}^{(j)}}), \tag{14}$$

$$p(d_i|rest) \propto p_{y_i} \frac{d_i^{(\alpha-1)} \beta^{\alpha} e^{-x\beta}}{\Gamma(\alpha)} \prod_{j=1}^{J} \frac{(\pi_{y_i l_{ij}}^{(j)})^{d_i}}{\sum_{q=1}^{K} (\pi_{y_i q}^{(j)})^{d_i}}. \tag{15}$$

Machine Teaching for Crowd Workers

Machine teaching (Zhu 2015) is an inverse problem of machine learning. Given a learner and a test set, machine teaching seeks a small teaching set $\mathcal{D}^{\dagger}$ such that the learner trained on $\mathcal{D}^{\dagger}$ has the smallest test error. In this study, the test set is the instances selected by the active learning strategy, i.e., $\mathcal{D}^{*} = \{\boldsymbol{x}_i^{*}\}_{i=1}^{m}$ in Figure 1. The machine teaching framework poses an optimization problem:

$$\min_{\mathcal{D}^{\dagger} \in \mathcal{D}^{\mathcal{E}}} \text{loss}(\mathcal{D}^{\dagger}) + \text{effort}(\mathcal{D}^{\dagger}). \tag{16}$$

The search space is the exemplar pool $\mathcal{D}^{\mathcal{E}}$ and the fixed-size teaching set $\mathcal{D}^{\dagger}$ contains exemplars $\{< \boldsymbol{x}_i^{\dagger}, \hat{y}_i^{\dagger} >\}_{i=1}^{n}$. The effort() function usually links with the size of the teaching set. Since this study only considers the fixed-size teach set, we simply let $\text{effort}(\mathcal{D}^{\dagger}) = 0$. We define the teaching loss function as the generalization error:

$$\text{loss}(\mathcal{D}^{\dagger}) = \mathbb{E}_{(\boldsymbol{x}^{*}, y^{*}) \sim p(\boldsymbol{x}^{*}, y^{*})} \mathbb{E}_{\hat{y}^{*} \sim \hat{p}(y^{*}|\boldsymbol{x}^{*}, \mathcal{D}^{\dagger})} \mathbb{I}(y* \neq \hat{y}^{*}). \tag{17}$$

The outer expectation is with respect to the test distribution and the inner expectation is with respect to the predictions that the learner makes. In this study, the given learner is based on human cognition models.

There are many cognition models for human learning (Love 2013), among which the exemplar-based models have a strong connection with machine learning. The exemplar-based models assume that when making decisions, people often retrieve a limited set of items from memory. These items (i.e., exemplars) provide evidence for competing options. People have limited capacity in memory for learning exemplars, which is coincide with the settings of machine teaching. Furthermore, the exemplar retrieval process in memory works similarly to the calculation of the similarity between the stored exemplars and the items to be judged (Giguère and Love 2013). To model the human cognition (i.e., the given learner in machine teaching), we employ the classic Generalized Context Model (GCM) (Nosofsky 2011). We extend GCM to the multi-class decision. Given teaching set $\mathcal{D}^{\dagger} = \{< \boldsymbol{x}_i^{\dagger}, \hat{y}_i^{\dagger} >\}_{i=1}^{n}$ and a test item $\boldsymbol{x}^{*}$, GCM estimates the label probability as:

$$\hat{p}(y^{*} = k|\boldsymbol{x}^{*}, \mathcal{D}^{\dagger}) = \frac{\left(b + \sum_{i \in \mathcal{D}^{\dagger}: \hat{y}_i^{\dagger} = k} e^{-cdst(\boldsymbol{x}^{*}, \boldsymbol{x}_i^{\dagger})}\right)^{\tau}}{\sum_{k=1}^{K} \left(b + \sum_{i \in \mathcal{D}^{\dagger}: \hat{y}_i^{\dagger} = k} e^{-cdst(\boldsymbol{x}^{*}, \boldsymbol{x}_i^{\dagger})}\right)^{\tau}}, \tag{18}$$

where $dst()$ is the normalized distance, c is a scaling parameter that specifies the rate of similarity decreasing with distance, b is background similarity, and τ is the response scaling parameter. Parameters $\{b, c, \tau\}$ can be viewed as constants determined by previous psychological experiments (Nosofsky and Palmeri 1997; Giguère and Love 2013).

Compared with a similar model in (Patil et al. 2014), our model is more complicated: first, our model is a multi-class decision model; and second, the true labels of the test set $\mathcal{D}^{*}$ are unknown, which means we must enumerate the classes of labels of a test item when minimizing the loss function. Plugging Eq.(18) into Eq.(17), by searching the fixed-size teaching set $\mathcal{D}^{\dagger}$ in exemplar pool $\mathcal{D}^{\mathcal{E}}$, we minimize the loss function as follows:

$$\begin{aligned}
\text{loss}(\mathcal{D}^{\dagger}) &= \underset{\mathcal{D}^{\dagger} \in \mathcal{D}^{\mathcal{E}}}{\text{argmin}} \, \mathbb{E}_{\boldsymbol{x}^{*} \sim p(\boldsymbol{x}^{*})} \sum_{k=1}^{K} p(y^{*} = k) \hat{p}(y^{*} \neq k|\boldsymbol{x}^{*}, \mathcal{D}^{\dagger}) \\
&= \underset{\mathcal{D}^{\dagger} \in \mathcal{D}^{\mathcal{E}}}{\text{argmin}} \frac{1}{m} \sum_{j=1}^{m} \sum_{k=1}^{K} \Bigg(p(y_j^{*} = k|\boldsymbol{x}_j^{*}) \\
&\quad + \frac{1 - K p(y_j^{*} = k|\boldsymbol{x}_j^{*})}{1 + \left(\frac{b + \sum_{i: \hat{y}_i^{\dagger} \neq k} e^{-cdst\left(\boldsymbol{x}_j^{*}, \boldsymbol{x}_i^{\dagger}\right)}}{b + \sum_{i: \hat{y}_i^{\dagger} = k} e^{-cdst\left(\boldsymbol{x}_j^{*}, \boldsymbol{x}_i^{\dagger}\right)}} \right)} \Bigg).
\end{aligned} \tag{19}$$

In our settings, for any test item $\boldsymbol{x}_j^{*}$ (i.e., the item posted to workers for labeling), its true $p(y_j^{*} = k|\boldsymbol{x}_j^{*})$ is unknown. However, since we have already built a learning model h using data set $\mathcal{D}^{\mathcal{O}}$ with integrated labels, we can use $h(\boldsymbol{x}_j^{*})$ to estimate this value used in Eq.(19). The optimization of Eq.(19) seems challenging. However, in practice, have two simple solutions. As we know, m and n are usually small. Because a HIT cannot contain too much information we usually have $m, n \leq 10$. Besides, the exemplar pool usually at most contains several hundreds of items, which is the upper

limit of human learning ability. Therefore, if the size of exemplar pool $|\mathcal{D}^{\mathcal{E}}|$ is small (in most cases, it is.), we can enumerate $C^n_{|\mathcal{D}^{\mathcal{E}}|}$ combinations of n teaching exemplars. If $|\mathcal{D}^{\mathcal{E}}|$ is large, we can use a greedy forward searching method to obtain an approximate optimal solution. In our experiments, we only use the first exhaustive method.

Active Leaning Strategies and Algorithm

The active learning strategies solve the problems of how to generate exemplar pool $\mathcal{D}^{\mathcal{E}} = \{< \boldsymbol{x}_i^E, \hat{y}_i^E >\}_{i=1}^n$ and query set $\mathcal{D}^* = \{\boldsymbol{x}_i^*\}_{i=1}^m$, which are usually heuristic and can be adjusted according to application domains.

Generation of exemplar pool We define what kind of instances can serve as exemplars, which is based on two factors: label uncertainty of instances and difficulties of instances. The label uncertainty of an instance is defined based on the diversity of the collected labels and the reliability of the workers who labeled it. That is, each noisy label has a different weight. For instance $\boldsymbol{x}_i$, it obtains J labels in total and n_k labels of class k are obtained from different workers with reliability $\{r^{(k_i)}\}_{i=1}^{n_k}$ defined by Eq.(2), its entropy-formed label uncertainty can be defined as follows:

$$UL(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \frac{\sum_{i=1}^{n_k} r^{(k_i)}}{\sum_{j=1}^{J} r^{(j)}} \log \frac{\sum_{i=1}^{n_k} r^{(k_i)}}{\sum_{j=1}^{J} r^{(j)}}. \quad (20)$$

Strategy 1 *(Exemplar selection)*: *For each class k ($1 \le k \le K$), we select both the simplest and the most difficult instanceswith the minimum uncertainty as exemplars:*[1]

$$\boldsymbol{x}_i^E = \underset{\boldsymbol{x}_i \in \mathcal{D}^{\mathcal{O}}}{\operatorname{argmin}} \left(\overline{UL}(\boldsymbol{x}_i) + 0.5 - |0.5 - \overline{DF}(\boldsymbol{x}_i)|\right), \quad (21)$$

where $\overline{UL}$ and $\overline{DF}$ are normalized label uncertainty and instance difficulty over all inferred instances $\mathcal{D}^{\mathcal{O}}$, respectively.

Generation of query instances We design a comprehensive strategy for instance selection. Besides the label uncertainty, our strategy also considers the model uncertainty of instances measured by current learning model (h) and the representativeness of instances. The model uncertainty of instance $\boldsymbol{x}_i$ is defined as:

$$UH(\boldsymbol{x}_i) = -\sum_{k=1}^{K} h(\hat{y}_i = k) \log h(\hat{y}_i = k), \quad (22)$$

where $h(\hat{y}_i = k)$ is the probability of h predicting y_i as class k. The representativeness of instance $\boldsymbol{x}_i$ is defined as:

$$RP(\boldsymbol{x}_i) = \frac{1}{|\mathcal{D}^{\mathcal{U}}| + |\mathcal{D}^{\mathcal{L}}| - 1} \sum_{\boldsymbol{x}_j \in (\mathcal{D}^{\mathcal{U}} \cup \mathcal{D}^{\mathcal{L}}) \setminus \boldsymbol{x}_i} dst(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad (23)$$

which is measured by averaging the distances from $\boldsymbol{x}_i$ to all the other instances.

Strategy 2 *(Query instance selection)*: *We select the instances with the maximum label uncertainty, the maximum model uncertainty, and the maximum representativeness:*[2]

$$\boldsymbol{x}_i^* = \underset{\boldsymbol{x}_i \in (\mathcal{D}^{\mathcal{U}} \cup \mathcal{D}^{\mathcal{L}})}{\operatorname{argmax}} \left(\overline{UL}(\boldsymbol{x}_i)\overline{UH}(\boldsymbol{x}_i)RP(\boldsymbol{x}_i)\right)^{1/3}, \quad (24)$$

where $\overline{UH}$ is normalized model uncertainty over $\mathcal{D}^{\mathcal{U}} \cup \mathcal{D}^{\mathcal{L}}$.

Here, Eq.(24) provides a ranking mechanism for us to select m query instances.

Algorithm 1 ILPCE

Input: $\mathcal{D}^{\mathcal{U}}, \{\boldsymbol{\nu}, \boldsymbol{\Lambda}, \alpha, \beta\}, \{b, c, \tau\}, m, n$
Output: learning model $h(\mathbf{x})$

```
Initialization: a small portion (5%) of instances are randomly chosen to acquire values from crowd workers;
while h(x) can be improved & the budget is enough do
    while NOT convergence do
        Gibbs sampling by Eqs.(11)∼(15).
    Perform Strategies 1 & 2 to form D^E and D^*
    Learn h(x) from D^O = D^E ∪ D^L
    Optimize Eq.(19) to obtain D^†
    Workers (re)label D^* and D^† (opt.) while learn D^†
return h(x).
```

Algorithm ILPCE We summarize all key steps of our Interactive Learning with Proactive Cognition Enhancement in Algorithm 1. In the beginning, a small portion of unlabeled instances is selected to acquire labels to overcome the cold-start issue. Then, it goes into the proposed interactive learning process. The time complexity of the algorithm is $O(\frac{J|\mathcal{D}^{\mathcal{U}}|(t^{inf}+t^{mt}+t^{ml})}{m})$, where J is the number of workers, t^{inf}, t^{mt}, t^{ml} are the running time of truth inference, machine teaching, and learning model training, respectively.

Experiments

We recruited 328 workers from Figure-Eight.com to label three classification datasets. We developed a Web application that encapsulates the compared algorithms to show HITs on the platform. Each HIT contains five query items ($\mathcal{D}^*$) and at most ten teaching exemplars with integrated labels ($\mathcal{D}^{\dagger}$). The workers watched the teaching exemplars, answered five query items, and optionally provided additional labels to the exemplars if they thought that their current integrated labels were incorrect. To investigate the teaching performance, we gradually increased the payment per HIT according to the number of HITs that a worker had finished so that the worker was willing to do more HITs.

Datasets

We used three image classification datasets that are not easy for human experts in our experiments. (**1**) Dataset *Butterflies* in (Mac Aodha et al. 2018) includes 2224 images of five different species of butterflies captured in real-world situations with varying image quality. We randomly selected almost a half images from each class of the original dataset to form a 1000-image dataset. (**2**) *Birds* species classification on the

[1] For our humans, we usually start to learn concepts from simple exemplars and deepen our understanding with difficult exemplars.

[2] $UL(\boldsymbol{x}_i) = -\sum_{k=1}^{K} (1/K) \log(1/K)$, for all $\boldsymbol{x}_i \in \mathcal{D}^{\mathcal{U}}$.

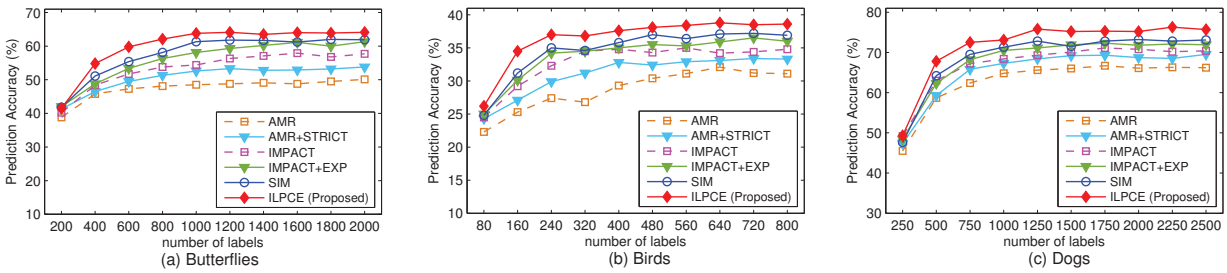

Figure 3: Comparison of the performance of learning models in prediction accuracy on three datasets

Caltech-UCSD Birds 200 dataset (Welinder et al. 2010) includes 6033 images with 200 species. We extracted 8 bird species, each of which contains 50 images. The dataset includes 400 images. (**3**) Dataset *Dogs* in (Bi et al. 2014) contains 10 dog species extracted from the ImageNet (Deng et al. 2009), each of which contains 142 images. We used all these 1420 images in our experiments.

For all the datasets, we held out 30% images from each class as the test sample for learning models. The remainder 70% were used for the (inter)active learning processes.

Experimental Settings

We carefully chose five state-of-the-art algorithms in comparison. The crowd workers did not know the existence of these algorithms. Each algorithm was assigned a group id. When a worker began the first HIT, s/he was randomly linked with a group id, which means that her/his outcomes would be evaluated with the corresponding algorithm. Each worker only belonged to one group.

Algorithms in comparison Our interactive learning framework includes three technical points: truth inference, machine teaching, and active learning. At each point, we may have multiple technical choices. However, due to the limit of budget, it impossible to evaluate too many combinations of these techniques. The principles of choosing the compared algorithms are as follow: First, we ignore the comparison of truth inference. We designed a novel truth inference BIM because we need simultaneously model the reliability of worker and difficulties of instances. However, the study focuses on how to perform machine teaching in active learning. Thus, to avoid the influence of truth inference, we always used our BIM in all experiments. Second, since this study is the first one that embeds machine teaching within active crowdsourcing learning. We must include some combinations of machine teaching and active learning techniques. Thus, the algorithms in comparison are:

(**1**) AMR (Zhao, Sukthankar, and Sukthankar 2011) selects the instances for crowd labeling based their uncertainty and inconsistency. (**2**) IMPACT (Lin, Mausam, and Weld 2016) selects the instances with the greatest impact on the current learning model. (**3**) AMR+STRICT is the combination of the AMR active learning and the STRICT (Singla et al. 2014) machine teaching. STRICT generates nearly-optimal minimum teaching set against a predefined hypothesis $\mathcal{H}$. In our experiments, we use the ground truth to train a set of predefined hypotheses. Since STRICT was proposed for binary labeling in the static environment, we used the one-versus-the-rest scheme to randomly sample instances for building hypotheses for each class in each active learning iteration. Ten teaching exemplar is generated from $\mathcal{D}^{\mathcal{E}} \cup \mathcal{D}^{\mathcal{L}}$. (**4**) IMPACT+EXP is the combination of the IMPACT active learning and the STRICT (Singla et al. 2014) machine teaching. EXP (Mac Aodha et al. 2018) is similar to STRICT but uses different objective function for multi-class labeling. The teaching process is the same as STRICT. (**5**) SIM is a simple version of the proposed ILPCE. For each $\boldsymbol{x}_i^* \in \mathcal{D}^*$, we selected two most similar teaching exemplar from $\mathcal{D}^{\mathcal{E}}$.

Parameter settings and evaluation metric The parameter settings of our method are as follows: Each element in hyper-parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Lambda}$ is set to $1/K$, where K is the number of classes. That is, we use uniform priors. The hyper-parameters (α, β) for the Gamma distribution is set to $(5.0, 1.0)$, making the shape of the probability density function as a Gaussian distribution. The parameters $\{b, c, \tau\}$ for cognition model are set to $\{5.07, 2.96, 4.80\}$, which are taken from the previous psychological experiment (Giguère and Love 2013). Finally, as mentioned before, we set $m = 5$ and $n = 10$. For the other algorithms used in the comparison, we used the same settings as they were in the original articles. However, their learning models are updated after five instances have obtained labels from crowd workers. All learning models are trained with SIFT (Lowe 2004) features using logistic regression with L2 regularization. We use the cosine distance to measure the similarity of instances.

Since in the datasets, the class distributions are nearly balanced. We simply use the overall accuracy to evaluate the performance of learning models and the reliability of crowd workers. The average values are reported below.

Experimental Results

We first show experimental results on the performance of learning models, which is the primary goal of active learning. Then, we show experimental results on the performance of crowd workers to verify the effectiveness of the proposed interactive learning mechanism.

Performance of Learning Models Figures 3(a), 3(b) and 3(c) show the prediction accuracy of the learning models trained by different algorithms on three datasets. From the results, we have some consistent observations as follows: (**1**) Machine teaching does have a positive impact

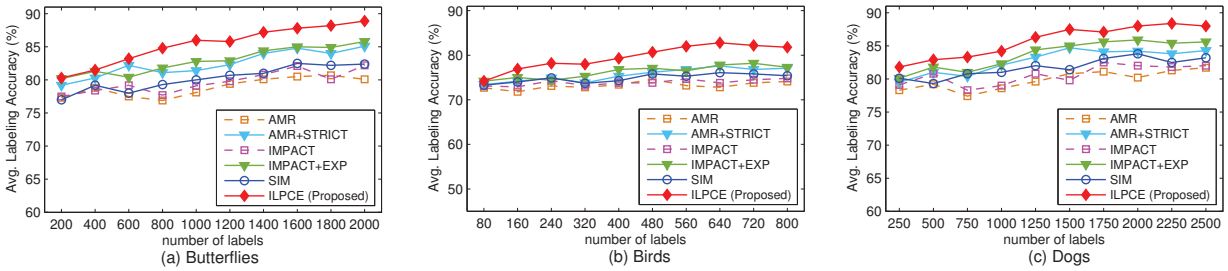

Figure 4: Comparison of the average labeling accuracy of crowd workers on three datasets

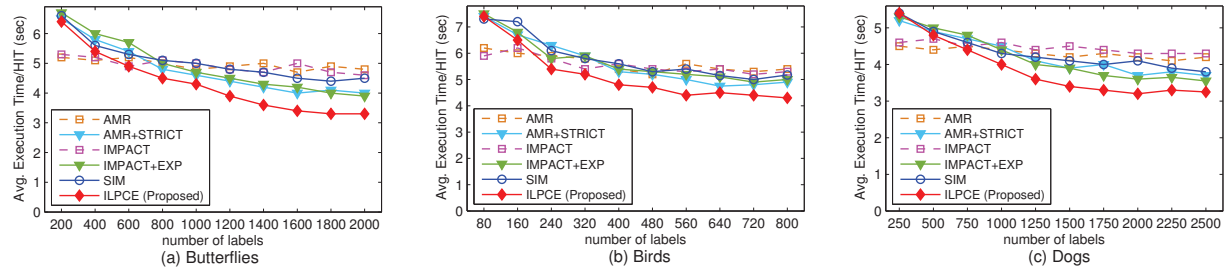

Figure 5: Comparison of the average execution time per HIT of crowd workers on three datasets

on the performance of learning models. On the three datasets, the algorithms without machine teaching (AMR and IMPACT) are consistently worse than their machine-teaching counterparts (AMR+STRICT and IMPACT+EXP). (**2**) Active learning strategy plays a critical role in learning model performance. The order of the performance of the models trained by different active learning strategies is SIM$\succ$IMPACT$\succ$AMR. Here, SIM can be treated as a weak teaching version of the proposed ILPCE. Although we add machine teaching (STRICT) into AMR, we still have AMR+STRICT$\prec$IMPACT. Similarly, after adding machine teaching (EXP) into IMPACT, IMPACT+EXP still can hardly surpass SIM. (**3**) The proposed ILPCE significantly outperforms all the other methods on the three datasets. The increment of the performance comes from two aspects: first, we have a better active learning strategy (SIM$\succ$IMPACT$\succ$AMR); and second, our machine teaching method demonstrates its effectiveness (ILPCE$\succ$SIM). (**4**) Machine teaching can achieve better results in moderately difficult tasks. From the perspective of either human common sense or the prediction performance in Figure 3, the difficulty order of three tasks is *Birds*$\succ$*Butterflies*$\succ$*Dogs*. Our ILPCE gains around 15 points increment than the worst AMR on the moderate difficult task *Butterflies*. On *Dogs* and *Birds*, the increments are 9 and 6 points, respectively.

Performance of Crowd Workers To further investigate the impact of the proposed method on worker reliability, we calculated the average labeling accuracy of the workers on each dataset when performing different algorithms, shown in Figure 4(a), 4(b) and 4(c). The experimental results consistently show: (**1**) Compared with the methods without machine teaching (AMR and IMPACT), their machine-teaching counterparts (AMR+STRICT and IMPACT+EXP) have higher labeling accuracy. (**2**) The proposed ILPCE not only has the highest labeling accuracy but also can continuously raise accuracy. (**3**) SIM only has a rather weak teaching effect. (**4**) The teaching effect is most evident on the moderately difficult dataset (*Butterflies*) with the greatest accuracy increment of 8.8 points (comparing ILPCE with AMR). On the hardest dataset *Birds* and easiest dataset *Dogs*, the increments are 7.1 and 6.3 points, respectively.

Figures 5(a), 5(b), and 5(c) show the execution time per HIT during the learning processes. The experimental results consistently show: (**1**) The machine teaching scheme (AMR+STRICT, IMPACT+EXP, SIM, and ILPCE) costs crowd work more time at the early stage of active learning because they spend time to learn domain knowledge. As the learning goes, the HIT execution speed is accelerated. (**2**) The proposed ILPCE has the best execution speed acceleration on all datasets. (**3**) On the moderately difficult dataset (*Butterflies*), the acceleration effect of machine teaching methods is most conspicuous.

Conclusion

We propose an interactive learning framework, which not only includes novel truth inference and active learning strategy but also provides a proactive mechanism that uses machine teaching to improve the cognition of crowd workers. Experiments on three real-world image annotation tasks show that the proposed novel active learning strategy and psychological model-based machine teaching together improve the performance of learning models. Particularly, our machine teaching method proactively enhances the reliability of worker and accelerates their task completion time.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grants 91846104, 61603186, 61806096, 61702264, and the Natural Science Foundation of Jiangsu Province, China, under grant BK20160843.

References

Amir, O.; Kamar, E.; Kolobov, A.; and Grosz, B. J. 2016. Interactive teaching strategies for agent training. In *IJCAI*, 804–811.

Bi, W.; Wang, L.; Kwok, J. T.; and Tu, Z. 2014. Learning to predict from crowdsourced data. In *UAI*, 82–91.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys* 51(1):7.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fu, Y.; Zhu, X.; and Li, B. 2013. A survey on instance selection for active learning. *Knowledge and Information Systems* 35(2):249–283.

Giguère, G., and Love, B. C. 2013. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences* 110(19):7613–7618.

Ipeirotis, P. G.; Provost, F.; Sheng, V. S.; and Wang, J. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28(2):402–441.

Irani, L., and Silberman, S. 2013. Interrupting worker invisibility in amazon mechanical turk. In *ACM SIGCHI*, 611–620.

Kurve, A.; Miller, D. J.; and Kesidis, G. 2015. Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention. *IEEE Trans. Knowledge and Data Engineering* 27(3):794–809.

Lin, C. H.; Mausam, M.; and Weld, D. S. 2016. Re-active learning: Active learning with relabeling. In *AAAI*, 1845–1852.

Love, B. C. 2013. Categorization. In *The Oxford Handbook of Cognitive Neuroscience*. 342–358.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Mac Aodha, O.; Su, S.; Chen, Y.; Perona, P.; and Yue, Y. 2018. Teaching categories to human learners with visual explanations. In *CVPR*, 3820–3828.

Nosofsky, R. M., and Palmeri, T. J. 1997. An exemplar-based random walk model of speeded classification. *Psychological Review* 104(2):266.

Nosofsky, R. M. 2011. The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization* 18–39.

Patil, K. R.; Zhu, J.; Kopeć, Ł.; and Love, B. C. 2014. Optimal teaching for limited-capacity human learners. In *NIPS*, 2465–2473.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11(Apr):1297–1322.

Retelny, D.; Robaszkiewicz, S.; To, A.; Lasecki, W. S.; Patel, J.; Rahmati, N.; Doshi, T.; Valentine, M.; and Bernstein, M. S. 2014. Expert crowdsourcing with flash teams. In *ACM UIST*, 75–85.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian process classification and active learning with multiple annotators. In *ICML*, 433–441.

Servajean, M.; Joly, A.; Shasha, D.; Champ, J.; and Pacitti, E. 2016. Theplantgame: Actively training human annotators for domain-specific crowdsourcing. In *ACM MM*, 720–721.

Settles, B. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *ACM SIGKDD*, 614–622.

Singla, A.; Bogunovic, I.; Bartók, G.; Karbasi, A.; and Krause, A. 2014. Near-optimally teaching the crowd to classify. In *ICML*, volume 32(2), 154–162.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, 2035–2043.

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. G. 2011. Active learning from crowds. In *ICML*, volume 11, 1161–1168.

Yang, D.; Xue, G.; Fang, X.; and Tang, J. 2016. Incentive mechanisms for crowdsensing: Crowdsourcing with smartphones. *IEEE/ACM Trans. Networking* 24(3):1732–1744.

Zhang, J.; Wu, X.; and Shengs, V. S. 2015. Active learning with imbalanced multiple noisy labeling. *IEEE Trans. Cybernetics* 45(5):1095–1107.

Zhao, L.; Sukthankar, G.; and Sukthankar, R. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *IEEE 3rd International Conference on Privacy, Security, Risk and Trust and IEEE 3rd International Conference on Social Computing*, 728–733.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10(5):541–552.

Zhou, Y.; Nelakurthi, A. R.; and He, J. 2018. Unlearn what you have learned: Adaptive crowd teaching with exponentially decayed memory learners. In *ACM SIGKDD*, 2817–2826.

Zhu, X. 2015. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 4083–4087.