Doubly Robust Crowdsourcing

Chong Liu and Yu-Xiang Wang
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, 93106 CA
chongliu@cs.ucsb.edu and yuxiangw@cs.ucsb.edu

Abstract

Large-scale *labeled* datasets are the indispensable fuel that ignites the AI revolution as we see today. Most such datasets are constructed using crowdsourcing services such as Amazon Mechanical Turk which provides noisy labels from non-experts at a fair price. The sheer size of such datasets mandates that it is only feasible to collect a few labels per data point. We formulate the problem of test-time label aggregation as a statistical estimation problem of inferring the expected voting score in an ideal world where all workers label all items. By imitating workers with supervised learners and using them in a doubly robust estimation framework, we prove that the variance of estimation can be substantially reduced, even if the learner is a poor approximation. Synthetic and real-world experiments show that by combining the doubly robust approach with adaptive worker/item selection, we often need as low as 0.1 labels per data point to achieve nearly the same accuracy as in the ideal world where all workers label all data points.

1 Introduction

The rise of machine learning approaches in artificial intelligence have enabled machines to perform well on many cognitive tasks that were previously thought of as what makes us humanly. In many specialized tasks, e.g.,wildlife recognition in images [7], conversational speech recognition [25], translating Chinese text into English [6], learning-based systems are shown to have reached and even surpassed human-level performances. These remarkable achievements could not have been possible without the many large-scale data sets that are made available by researchers over the past two decades. ImageNet, for instance, has long been regarded as what spawned the AI revolution that we are experiencing today. These labels do not come for free. ImageNet's 11 million images were labeled using Amazon Mechanical Turk (AMT) into more than 15,000 synsets (classes in an ontology). On average, each image required roughly $2-5$ independent human annotations, which are provided by 25,000 AMT workers over a period of three years [1]. We estimate that the cost of getting all these annotations to go well above a million dollars.

[1]The quoted statistics are from http://www.image-net.org/papers/ImageNet_2010.pdf.

As the deep learning models get larger and more powerful every day so as to tackle some of the more challenging AI tasks, their ferocious appetites for even larger labeled data set have grown tremendously as well. However, unlike the abundant unlabeled data, it is often difficult, expensive, or even impossible to consult expert opinions on large number of items. Here the items can be images, documents, voices, sentences, and so on. Services such as AMT have made it much easier to seek the wisdom of the crowd by having non-experts (called workers in the remainder of this paper) to provide many noisy annotations at a much lower cost. A large body of work have been devoted to finding a more scalable solution. These include a variety of label-aggregation methods[20, 24, 26, 27], end-to-end human-in-the-loop learning [12], online/adaptive worker selections [2, 21] and so on. At the heart of these approaches, are various ways to evaluate individual worker performances and quantify the uncertainty in their provided labels.

In this paper, we take a pre-trained crowdsourcing model with worker evaluation as a blackbox and consider the problem of true label inference for new data points. We formulate this problem as a statistical estimation problem and propose a number of ways to radically reduce the number of worker annotations. These include:

Worker imitation We propose to imitate each worker with a simple supervised learner that learns to predict the worker's label using the item feature.

Doubly robust crowdsourcing (DRC) By tapping into the literature on doubly robust estimation, we design algorithms that exploit the possibly unreliable imitation agents and significantly reduce the estimation variance (hence annotation cost!) while remaining unbiased.

Adaptive item/worker selection (AWS/AIS) We propose to bootstrap the imitation agents' confidence estimates to adaptively filter out high confidence items and selecting the most qualified workers for low-confidence item, without any additional cost.

Our results are summarized as follows.

1. We theoretically show that the doubly robust crowdsourcing technique can be used to generically improve any given crowdsourcing models using any nontrivial learned imitation agents.

2. Synthetic and real-world experiments show DRC improves the label accuracy over the standard probabilistic inference with Dawid-Skene model in almost all budget levels and all data sets.

3. Moreover, DRC with AIS and AWS often reduces the cost by orders of magnitudes, while keeping the same level of accuracy. In several data sets, the proposed technique can often get away with using only 0.1 annotations per item while achieving nearly the same accuracy that can be obtained by having all workers annotating all items.

1.1 Related Work

We briefly summarize the related work. Our study is motivated by the many trailblazing approaches in label-aggregation including the wisdom-of-crowds [24], Dawid-Skene model [4, 26], minimax entropy approach [27], permutation-based model [19], worker cluster model [9], crowdsourced regression model [15] and so on. Our contribution is complementary as we can take any of these models as blackboxes and hopefully improve their true-label inference.

Doubly robust techniques originates from the causal inference literature [17, 1] and the use of it for variance reduction had led to several breakthroughs in machine learning, e.g., [11, 22]. We drew our inspirations directly from the use of doubly robust techniques in the off-policy evaluation problem in bandits and reinforcement learning [5, 10, 23]. The variance analysis and weight-clipping are adapted from the calculations in [5, 23] with some minor differences. To the best of our knowledge, this is the first paper considering doubly robust techniques in crowdsourcing.

Our idea of adaptive item/worker selection is inspired by the recent work of [2, 21]. They propose an AI-aided approach that reduces the number of worker labels per item to be < 1 in an object detection task. The key idea is to train a computer vision algorithm to detect the bounding boxes using the aggregated labels that have been obtained thus far and if the algorithm achieves a high confidence on a new image, then the annotation provided by the algorithm is taken as is.

The differences of our work is twofold. First, our use of supervised learner is not to predict the true labels but rather to imitate workers. Second, our confidence measure is determined by supervised learners' approximation to what all workers would say about an item, rather than as a prior distribution added to model-based probabilistic inference.

2 Problem Setup

In this section, we introduce the notations and formulate the problem as a statistical estimation problem.

2.1 Symbols and Notations

Suppose we have n items, m workers, and k classes. We adopt the notation $[k] := \{1, 2, 3, ..., k\}$. Each item $j \in [n]$ is described as a d-dimensional feature vector $\boldsymbol{x}_j$, and the feature matrix is $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^\mathsf{T} \in \mathsf{R}^{n \quad d}$. Each item $j \in [n]$ also has a hidden true label $y_j \in [k]$ which indicates the correct class that item j belongs to.

Workers, such as those on AMT, are requested to classify items into one of the k classes. We denote the label that Worker $i \in [m]$ assigns to item j as $\ell_{ij} \in [k]$. It is important to distinguish the worker-produced labels ℓ_{ij} with the true label y_j, as the workers are considered non-experts and they make mistakes. From here onwards, we will refer to the potentially noisy and erroneous labels from workers as "annotations". Conveniently, we

also collect ℓ_{ij} into a matrix $L \in ([k] \cup \{?\})^{m \times n}$, where any entries in L that are $?$ are unobserved labels. We use $[m] \times [n]$, $\mathcal{M}_i \subseteq [n]$, $\mathcal{N}_j \subseteq [m]$ to denote the indices of the observed annotations, indices of all items worker i annotated and indices of all workers that annotated Item j respectively. For a generic item (x, y), $\mathcal{M}_x$ collects the indices of workers who annotated the item and the corresponding annotation is denoted by ℓ_i for each $i \in \mathcal{M}_x$.

2.2 Problem Statement

The goal of the paper is related to but different from the standard crowdsourcing problem which aims at learning a model that one can use to infer the true labels $y_1, \ldots, y_n$ using noisy annotations $L[\,]$ (and sometimes item features X). Many highly practical models were proposed for that task already [4, 24, 26, 27, 19].

Complementary to the existing work that mainly focuses on training, we consider the problem of cost-saving in test time. Specifically, we would like to design algorithms to reduce the expected number of new annotations needed to label a new item (x, y) using a pre-trained crowdsourcing model as well as the training data set X and $L[\,]$.

2.3 Dawid-Skene Model and Score Functions

The primary model that we work with in this paper is the Dawid-Skene model [4, 26], which assumes the following data generating process.

1. For each $j \in [n]$, $y_j \sim \text{Categorical}(\rho)$.
2. For each $j \in [n], i \in [m]$, $\ell_{ij} \sim \text{Categorical}(\pi_{y_j, i})$.
3. We observe ℓ_{ij} with probability ω_{ij}.

where ρ and $\pi_{y,i}$ denote the probability distributions defined on $[k]$. In particular, $\pi_{y,i}$ is the Column y of the confusion matrix of worker i, which the DS model uses to describe $P_i(\ell|y)$. We denote the confusion matrix associated with Worker i by $\pi_i \in \mathbb{R}^{k \times k}$. Once the DS model is learned, we can make use of the learned parameters π and ρ to infer the true labels using worker annotations via the posterior belief

$$P(y|\ell_1, \ell_2, \ldots \ell_m) \propto P(y) \prod_{i=1}^{m} P(\ell_i|y) = \rho[y] \prod_{i=1}^{m} \pi_i[\ell_i, y]. \tag{1}$$

Take the log for both sides and dropping the additive constant, we obtain the score function that is induced by the DS model

$$S_{DS}(y|\ell_i \forall i \in \mathcal{M}_x) = \log \rho[y] + \sum_{i=1}^{m} \log \pi_i[\ell_i, y]. \tag{2}$$

This is a weighted voting rule based on a pre-trained DS model. Similarly, we can cast the inference procedure of other crowdsourcing models as maximizing such a score function as well. For example, in the naive Majority Voting approach.

$$S_{MV}(y \mid x) = \sum_{i=1}^{m} 1(\ell_i(x) = y). \tag{3}$$

Notably, no training data sets are needed for majority voting. The exposition above suggests that test time involves collecting a handful of worker annotations (choosing x) and calculating a voting score specified by your favorite crowdsourcing model in a form of

$$S(y \mid x) = \sum_{i \in x} S_i(y; \ell_i(x)). \tag{4}$$

where S_i is supplied by the model that connects annotation ℓ_i to label y. Then the label y that maximizes the score is chosen.

2.4 A Statistical Estimation Framework

In the ideal world, when money is not a concern, we will poll all workers and calculate

$$S(y \mid [m]) = \sum_{i}^{m} S_i(y; \ell_i(x)). \tag{5}$$

In practice, however, just as we cannot afford to poll all voters to estimate who is winning the presidential election, we also cannot afford to poll everyone to annotate a single data point. But do we have to?

Notice that we can frame the question as a classical point estimation problem in statistics, where the statistical quantity of interest is

$$v^x(y) := E\left[\frac{1}{m}\sum_{i}^{m} S_i(y; \ell_i(x))\right]; \tag{6}$$

the expectation of the ideal world score function (5), rescaled by $1/m$. In the above, the expectation is taken over the randomness in worker's annotation. For example, if we select each worker independently with probability π, then the approach used in (2) and (3) would be an unbiased estimate of $v^x(y)$ defined using DS and MV respectively, if we rescale them by a factor of π^{-1}.

The advantage of translating the problem into a classical statistical estimation problem is that there is now a century of associated literature that we can tap into, including those on adaptive sampling and variance reduction techniques. We emphasize that while we will be using a crowd-sourcing model, e.g., the Dawid-Skene model, we do not assume that the data is generated according to the model. In fact, we are not imposing any restrictions on how workers annotate items, except that

1. $\ell_i(x) \forall i \in [m]$ are mutually independent given any item x.

2. $\mathrm{Var}[S_i(y, \ell_i(x))] < +1 \quad \forall x, y$.

These are very mild assumptions that are typically true in practice.

It is generally difficult to analytically model human behaviors because it depends on how the item is presented to worker as well as the worker's knowledge and cognitive processes. The agnostic learning point of view helps disentangle the approximation-theoretic questions from the statistical question of estimating the best approximation possible using a given crowdsourcing model.

The remainder of the paper will be about designing estimators of $v^x(y)$ that achieves accurate label-inference at a low cost and their corresponding theory and experiments.

To avoid any confusions, we emphasize again that item x is fixed. All the estimators are defined for each y separately. $\ell_1, \ldots, \ell_m$ are random variables that comes out of the unknown process of workers looking at the item x. Whenever the dependence is clear from context, we drop the conditioning on x for better readability.

3 Benchmark Approaches

In this section, we describe a few baseline approaches for estimating $v^x(y)$ and their corresponding cost in number of annotations. These are the approaches that we will compare to in the experiments.

3.1 Ideal World estimator

In the ideal world, all workers are required to label x. We can use

$$\hat{v}_{IW}(y) = \frac{1}{m} \sum_{i=1}^{m} S_i(y, \ell_i). \tag{7}$$

This estimator incurs a cost of m and it is unbiased, with a variance of $\frac{1}{m^2} \sum_{i=1}^{m} \mathrm{Var}[S_i(y, \ell_i(x))]$. This is arguably the best one can do additional source of information.

3.2 Importance Sampling estimator

A more affordable approach is to directly sample the workers. Specifically, we will include Worker i independently with probability π_i [2].

$$\hat{v}_{IS}(y) = \frac{1}{m} \sum_{i} \frac{\mathbb{1}(i \in \pi)}{\pi_i} S_i(y, \ell_i). \tag{8}$$

The expected cost of the IS estimator is $\sum_{i \in [m]} \pi_i$ and it is clearly an unbiased estimator.

[2] This is called a Poisson sampling [18] in the survey sampling theory

Theorem 1. The $\hat{v}_{IS}(y)$ is unbiased and

$$\mathrm{Var}[\hat{v}_{IS}(y)] = \frac{1}{m^2} \sum_i \left(\frac{1}{\pi_i} \mathrm{Var}[S_i(y;\grave{}_i)] + \left(\frac{1}{\pi_i} - 1 \right) E[S_i(y;\grave{}_i)]^2 \right) .$$

Proof. By the independence of sampling,

$$\begin{aligned} E[\hat{v}_{IS}] &= \frac{1}{m} \sum_{i=1}^{m} E[1(i \in \text{Omega})] \frac{1}{\pi_i} S_i(y;\grave{}_i) \\ &= \frac{1}{m} \sum_{i=1}^{m} \pi_i \frac{1}{\pi_i} E[S_i(y;\grave{}_i)] = v^x(y). \end{aligned}$$

To calculate the variance, we use the independence and then apply the law of total variance on each i:

$$\begin{aligned} \mathrm{Var}[\hat{v}_{IS}] &= \frac{1}{m} \sum_i \frac{1}{\pi_i^2} \mathrm{Var}[1(i \in \quad) \; S_i(y;\grave{}_i)] \\ &= \frac{1}{m} \sum_i \frac{1}{\pi_i^2} \left(\pi_i \mathrm{Var}[S_i(y;\grave{}_i)] + \pi_i E[S_i(y;\grave{}_i)]^2 - \pi_i^2 E[S_i(y;\grave{}_i)]^2 \right) \\ &= \frac{1}{m^2} \sum_i \left(\frac{1}{\pi_i} \mathrm{Var}[S_i(y;\grave{}_i)] + \left(\frac{1}{\pi_i} - 1 \right) E[S_i(y;\grave{}_i)]^2 \right) \end{aligned}$$

as claimed. □

Remark. If π_i , the we will are essentially doing the standard probabilistic inference as in (2) and (3). When $\pi_i = 1$, IS trivially subsumes (7) as a special case. Moreover, since x is fixed, the sampling π_i can be chosen as a function of the item x without affecting the above results.

3.3 Worker Imitation and Direct Method

Finally, there is an option that comes with no cost at all! Recall that we have a data set X and L that were used to train the crowdsourcing model at our disposal. We can reuse the data set and train m supervised learners to imitate each worker's behavior. Let the fictitious annotations provided by these supervised learners be $\hat{}_1, \ldots, \hat{}_m$, we can simply plug them into the ideal world estimator (7) without costing a dime!

$$\hat{v}_{DM}(y) = \frac{1}{m} \sum_{i=1}^{m} E[S_i(y;\hat{}_i)]. \tag{9}$$

Following the convention in the contextual bandits literature, we call this approach the direct method (DM). The additional E is introduced to capture the case when the supervised learner outputs a soft annotation $\hat{\eta}_i$.

The variance of this approach is 0. However, as we mentioned previously, we can never hope to faithfully learn human behaviors, especially when we only have a small number of annotations in the training data for each Worker i. As a result, (9) may suffer from a bias that does not vanish even as $m \to \infty$.

4 Main Results

In this section, we adapt an old statistical technique, called doubly robust estimation, to the crowdsourcing problem.

4.1 Doubly Robust Crowdsourcing

As we established in the last section, the importance sampling estimator is unbiased but suffers from a large variance, especially when we would like to cut cost and use a small sampling probability. The DM estimator incurs no additional annotation cost and has 0 variance, but it can potentially suffer from a large bias due to supervised learners not imitating the workers well enough.

Doubly robust estimation [17, 5] is a powerful technique that allows us to reduce the variance using a DM estimator while retaining the unbiasedness, hence getting the best of both worlds. The doubly robust estimator works as follows:

$$\hat{v}_{DR}(y) = \frac{1}{m}\sum_{i=1}^{m} E[S_i(y;\hat{\eta}_i)] + \frac{\mathbb{1}(i \in \mathcal{E})}{\pi_i}(S_i(y;\ell_i) - E[S_i(y;\hat{\eta}_i)]) \quad : \tag{10}$$

The doubly robust estimator can be thought of using the DM as a baseline and then use IS to estimate and correct the bias. Provided that the supervised learners are able to provide a nontrivial approximation of the workers, the doubly robust estimator is expected to reduce the variance. Just to give two explicit examples of $\hat{v}_{DR}(y)$, under the Dawid-Skene model, the DR estimator is

$$\frac{1}{m}\sum_{i=1}^{m} \log P_{\hat{\eta}_i}(\hat{\eta}_i|y) + \frac{\mathbb{1}(i \in \mathcal{E})}{\pi_i} \log \frac{P_{\hat{\eta}_i}(\ell_i|y)}{P_{\hat{\eta}_i}(\hat{\eta}_i|y)}.$$

Similarly, for the majority voting model, we can write

$$\frac{1}{m}\sum_{i=1}^{m} e_{\hat{\eta}_i} + \frac{1}{\pi_i}\mathbb{1}(i \in \mathcal{E})(e_{\ell_i} - e_{\hat{\eta}_i}).$$

Theorem 2 (DRC). The doubly robust estimator (10) is unbiased and its variance is:

$$\frac{1}{m^2}\left(\sum_i \frac{1}{\pi_i}\mathrm{Var}[S_i(y;\grave{}_i)] + \left(\frac{1}{\pi_i} - 1\right)\mathrm{E}[S_i(y;\grave{}_i) - S_i(y;\hat{}_i)]^2\right) :$$

Proof. Note that the first part $\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}[S_i(y;\hat{}_i)]$ of the estimator is not random. The result follows directly by invoking Theorem 1 on the second part of the estimator, which is an importance sampling estimator of the bias. □

Remark. First, if workers are deterministic, the first part of the variance $\mathrm{Var}[S_i(y;\grave{}_i)] \to 0$. Second, if the supervised learner imitates workers perfectly in expectation, the second part of the variance vanishes. Finally and most importantly, the supervised learner does not have to be perfect. In the simple case of a deterministic workers, the percentage of agreements between supervised learners and their human counterparts directly translate into a reduction of the variance of about the same percentage, for free!

The third point is especially remarkable as it implies that even a trivial surrogate that outputs a label at random could lead to a $1/k$ factor reduction of the variance. On the other hand, a good set of worker imitators with 90% accuracy can lead to an order of magnitude smaller variance and hence allow us to incur a much lower cost on average. We will illustrate the effects of doubly robust estimation more extensively in the experiments. This feature ensures that our proposed method remains applicable even in the case when the training data set contain few annotations from some subset of the features.

4.2 Confidence-based Adaptive Sampling

Doubly robust estimation allows us to reduce the variance. However, doubly robust is still an importance sampling-based method that requires the number of new annotations to be at least linear in the number of data points to label.

In this section, we propose using the supervised learning imitation of the workers to obtain confidence estimates for free and using them to construct confidence-based adaptive sampling schemes.

We propose two rules.

Adaptive item selection For each new data point, run DM first. If DM predicts label y with an overwhelming confidence, then chances are, there is no need to collect more annotations. If not, human workers are needed.

Adaptive worker selection We can adaptively choose which worker to annotate a given item. Instead of sampling at random with probability π, we choose a set of adaptive sampling probability $\pi_1, \ldots, \pi_m$ that makes high confidence workers more likely to be selected. As different workers have different skill sets, confidence may depend strongly on each item x. We propose to calculate such item-dependent confidence using outcome of the imitated workers and the confusion matrices from the DS model.

In both cases, we need a way to measure con dence given a probability distribution. A threshold is introduced to decide whether accept predicted labels or not[2]. Margin in multi-class classi cation is de ned as the di erence between the score of true label and the largest score of other labels[1][4]. Inspired by them, we de ne the con dence margin of a probability as follows.

De nition 1 (Con dence Margin). Given a discrete probability distribution $_1; :::; _m$, its con dence margin is de ned as the di erence between the largest element and the second largest one.

Based on con dence margin, we propose 3 new methods: DRC with Adaptive Item Selection (DRC-AIS), DRC with Adaptive Worker Selection (DRC-AWS), and the combination DRC-AWS-AIS.

In DRC-AIS, DM is performed for all labels. For each item, the surrogate label given by DM follows (1) to get the posterior belief, which describes how con dently DM gives the label of this item. Based on posterior belief, its con dence margin $_{AIS}$ is compared with the given con dence margin parameter . If $_{AIS}$ is larger, DRC-AIS takes the surrogate label provided by DM with no worker cost, otherwise, DRC-AIS follows the regular DRC model which incures cost.

In DRC-AWS, again DM runs rst and gets surrogate labels $\hat{}_i$. Then from each worker i's confusion matrix we can get the labeling probability $P(\hat{}_i|y)$, whose con dence margin is used as the worker score$_i$ and $_1; \ ; _m$ are normalized to be a distribution. For each item, worker i will be sampled with probability $_i$. However, in this case the sampling probability for each worker is usually very small, thus, we can introduce a parameter to multiply with $_i$ to increase the sampling probability. If $_i$ is larger than 1, it needs to be scaled to1. If a worker is sampled, the corresponding label is used as regular DRC model.

Table 1: Summary of benchmark and DRC approaches

Methods	Input	Sampling	Cost
IW	L	No	O(nm)
IS	L		O (nm)
DM	X + f	No	0
DRC	L + + (y) + f		O (nm)
DRC-AIS	L + + (y) + f		O (n m)
DRC-AWS	L + + (y) + f	$_{1:m}$	O($\sum_{i2[m]}$ $_i$n)

n n denotes the number of items AIS selected as low-con dence.

Table 1 shows the summary of benchmark and DRC approaches, including the input elements, sampling rate and the order of worker cost. As we can see, IW and IS take the

fewest input elements while IW has the most worker cost. DM has no cost because it only take advantage of surrogate labels given by worker imitation. DRC has similar cost as IS while it is expected to improve the ground truth inference with less variance. Thanks to the con dence-based adaptive sampling techniques, DRC-AIS and DRC-AWS are able to save even more cost than DRC.

4.3 Weight-clipping in Doubly Robust Crowdsourcing

Adaptive selection of workers involves making $_i$ larger from somei and smaller for other i. According to Theorem 2, the variance is proportional to $\sum_i {}_i{}^{1}$, hence even a single $_i$ being close to0 would result in a huge variance.

In the causal inference and o -policy evaluation problems this issue is addressed by clipping the importance weight at a xed threshold . This results in the clipped doubly robust estimator.

$$\hat{\theta}_{DR}(y) = \frac{1}{m}\sum_{i=1}^{m} E[S_i(y;\hat{\ }_i)] + 1(i\ 2\ )\min\{\ ;\ _i{}^{1}\}(S_i(y;\grave{}_i) \quad E[S_i(y;\hat{\ }_i)]) \ : \qquad (11)$$

The bias and variance of this estimator is given as follows.

Theorem 3. The clipped doubly robust estimator obeys that:

$$\mathrm{Bias}(\hat{\theta}_{DR}(y)) = \frac{1}{m}\sum_{i=1}^{m} \min\{\ _i \quad 1; 0\}E[S_i(y;\grave{}_i) \quad S_i(y;\hat{\ }_i)] \ :$$

$$\mathrm{Var}[\hat{\theta}_{DR}(y)] = \frac{1}{m^2}\sum_{i} \min\{\ ^2\ _i^2; 1\}\ \frac{1}{i}\mathrm{Var}[S_i(y;\grave{}_i)] + (\ \frac{1}{i} \quad 1)E[S_i(y;\grave{}_i) \quad S_i(y;\hat{\ }_i)]^2 \ :$$

Proof. The results follow from straightforward calculations similar to that in Theorem 1 and 2. □

Remark. The bias bound indicates that only those workers we clipped who contribute to the bias. The variance bound implies that the part of variance from Workeri is reduced to $O(\min\{1=\ _i;\ ^2\ _i\})$ from $O(1=\ _i)$. If the total amount of additional Bias^2 introduced by the clipping is smaller than the corresponding savings in the variance, then clipping makes makes the estimator more accurate in MSE. The theory inspires us to design an algorithm to automatically choose the threshold.

Remark (Automatic choice of the threshold). Assume$\grave{}_1; :::; \grave{}_m$ are deterministic, $|E[S_i(y;\grave{}_i) \quad S_i(y;\hat{\ }_i)]|$. Then the bias of$\hat{\theta}_{DR}(y)$ can be bounded by $\sum_i 1(\ _i{}^{1} > \ )=m$, and variance of$\hat{\theta}_{DR}(y)$ can be bounded by $^2\ ^2=m$. Recall that the mean square error can be decomposed in toBias^2 + Var. The optimal choice of that minimizes this upper bound is the one that minimizes$|\ \sum_i 1(\ _i{}^{1} > \ )=m|$. This can be found numerically in time $O(m\log(m))$ by sorting $[\ _1{}^{1}; :::;\ _m{}^{1}]$ and applying binary search.

5 Experiments

5.1 Synthetic Experiments

5.1.1 Experimental Settings

There are plenty of supervised learning datasets, but they usually don't have labels given by workers. In order to do take advantage of them to do experiments, we use worker imitation to generate labels. In this experiment, we use 6 classification datasets, Segment, DNA, Satimage, USPS [8], Pendigits, and MNIST [13], collected by Libsvm [3], which are all publicly available [3].

Figure 1 shows the workflow how we use worker imitation to generate crowdsourcing datasets, which has 3 steps.

Figure 1: The workflow of generating crowdsourcing datasets with item features for experiments.

1. In step 1, the feature matrix and labels of the classification datasets are the raw dataset. If the dataset has been split into training part and test part, we combine them together.

2. In step 2, we uniformly sample the dataset into two equal parts, one for training and the other for test. In order to remove the randomness of this sampling process, the sampling index is fixed and saved for all further experiments. Training means we use this part of data to train m decision trees with maximum depth 5 to simulate the generating process of workers to give labels. Following random forest, for one item only $\log(d)$ features can be observed by each tree where d is the total number of features. Then these m decision trees are used to make predictions on the test set to obtain item label matrix.

3. In step 3, we uniformly sample the test part of step 2 into two equal parts, one working as source part and the other working as target part. For all experiments, this sampling process will be conducted 10 times and mean and standard variance performances are

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

reported. In order to show our proposed approach is able to work with few labels, i.e., our DRC approaches doesn't need very good surrogate labels, we set the sparsity rate of source label matrix to be 0.5, specifically, for any item and any work, the probability of the label exists in source label matrix is 0.5. Also, m classifiers are trained on source dataset and label matrix to simulate worker behaviors to give surrogate labels for all DRC approaches. Evaluations are performed on the target label matrix.

5.1.2 Algorithm Comparison with Increasing Worker Cost

In order to show our approach DRC is able to infer true labels with low worker cost, we do experiments of DRC with Dawid-Skene (DRC-DS), DRC with Majority Voting (DRC-MV), IS, IW, DM, and Majority Voting (MV) with various worker cost. In detail, the number of workers m is fixed on 50 and we uniformly sample workers with sampling rate going from 0.1 to 1.0 with the step of 0.1, which means the number of workers goes from 5 to 50 with the interval of 5. The worker cost is defined as number of workers that per item needs. Logistic regression with default settings in scikit-learn is used as the base classifier for DRC-DS, DRC-MV and DM.

Results are shown in Figure 2. Because DM involves no sampling rate and IW uses all data points, they are two nodes in the figures corresponding to $= 0$ and $= 1$, respectively. For most datasets, given same worker cost, DRC-DS performs better than IS, and DRC-MV is better than MV, which shows the effectiveness of DRC. As a naive crowdsourcing approach, MV usually performs the worst. Also, the accuracy of DRC-DS, IS, DRC-MV, and MV usually increase rapidly at the initial increasing phase of worker cost while trend to be stable in the remaining increasing phase of worker cost. When worker cost is 50, DRC-DS and IS meet with IW, which is easy to understand because in that case, all workers are involved in labeling task.

5.1.3 Algorithm Comparison with Confidence-based Adaptive Sampling

We propose two rules, AIS and AWS, to do confidence-based adaptive sampling. To show their effectiveness, in this experiment, we compare the following four kinds of methods: DRC-DS, DRC-AIS, DRC-AWS, and DRC-AWS-AIS, while DM is used as the baseline. We subtract each method's accuracy and DM's accuracy for results. Because AWS and AIS rules are expected to save a lot of worker cost, logarithmic worker cost is used. For DRC-AIS and DRC-AWS-AIS, the confidence margin parameter is set to be 0.03, 0.06, and 0.09. For DRC-AWS and DRC-AWS-AIS, the multiplier parameter is set to be 1, 2, 3, 4, 5, 7, 10, 15, 25, and 50.

Results are shown in figure 3. As we can observe, DRC-AIS, DRC-AWS, and DRC-AWS-AIS are performing better than regular DRC-DS model especially when the worker cost is very little, which validates adaptive item selection and adaptive worker selection do play a key role in improving inference accuracy and saving worker cost at the same time.

Figure 2: Performance of DRC-DS, DRC-MV, IS, IW, DM, and MV with increasing worker costs.

Figure 3: Performance of DRC-DS, DRC-AIS, DRC-AWS, and DRC-AWS-AIS w.r.t. DM with increasing logarithmic worker costs.

Among them, DRC-AWS-AIS performs with the lowest worker cost, which means AIS and AWS can work together. Also, its performance can be improved by introducing more worker cost. The performances of most methods are above 0, which means they perform better than DM. It makes sense because all methods in this experiment are taking advantage of the surrogate labels provided by DM. As $\to\infty$, DRC-AIS trends to become DRC-DS and DRC-AWS-AIS trends to become DRC-AWS, which is most obvious on MNIST and

Table 2: Performance of DRC-DS, DRC-MV, and DM in realizable and unrealizable cases.

Dataset	Method	DT		LR		GNB	
Segment	DRC-DS	0.7823	0.0097	0.7828	0.0090	0.7811	0.0101
	DRC-MV	0.7939	0.0034	0.7988	0.0036	0.7977	0.0042
	DM	0.7024	0.0077	0.6951	0.0067	0.6863	0.0052
DNA	DRC-DS	0.7074	0.0067	0.7078	0.0052	0.6997	0.0068
	DRC-MV	0.5191	0.0034	0.5191	0.0034	0.5192	0.0034
	DM	0.6112	0.0058	0.6184	0.0095	0.6102	0.0089
Satimage	DRC-DS	0.8493	0.0026	0.8466	0.0020	0.8360	0.0030
	DRC-MV	0.8514	0.0022	0.8436	0.0023	0.8389	0.0023
	DM	0.8350	0.0021	0.7948	0.0028	0.8057	0.0020
USPS	DRC-DS	0.8142	0.0025	0.8145	0.0025	0.8028	0.0026
	DRC-MV	0.7816	0.0013	0.7794	0.0013	0.7718	0.0019
	DM	0.7687	0.0037	0.7818	0.034	0.7255	0.0036
Pendigits	DRC-DS	0.7935	0.0013	0.7910	0.0016	0.7913	0.0017
	DRC-MV	0.7577	0.0025	0.7518	0.0020	0.7586	0.0024
	DM	0.7794	0.0026	0.7706	0.031	0.7802	0.0026
MNIST	DRC-DS	0.6798	0.0057	0.6745	0.0056	0.6347	0.0047
	DRC-MV	0.3905	0.0009	0.3907	0.0010	0.4447	0.0007
	DM	0.5969	0.0038	0.5933	0.0034	0.4405	0.0039

USPS dataset. As increases, which means more workers can be sampled in DRC-AWS and DRC-AWS-AIS, their performances get better.

5.1.4 Algorithm Comparison with Realizable and Unrealizable Cases

Since we use decision trees to generate the crowdsourcing datasets with item features, it is interesting to know if we don't use decision trees as base classi ers for DRC-DS, DRC-MV and DM, what will happen. For methods with decision trees, we call it the realizable case because means all features used to generate labels can be observed by decision trees. We call it unrealizable cases with other classi ers. In this experiment, we compare DRC-DS, DRC-MV and DM with decision trees, logistic regression, and Gaussian naive Bayes, all with default parameter settings. Each item is associated with25 workers while the toal number of workersm is 50.

In Table 2, best results are set in bold font according to the mean. Decision trees perform only slightly better than other two classi ers, which is good because it show our approaches

Table 3: Performance of DM, IS, DRC-DS, MV, and DRC-MV on Music Genre dataset.

DM	IS	DRC-DS	MV	DRC-MV
0.4014±0.0096	0.4914± 0.0102	0.6246±0.0088	0.6509±0.0047	0.6966±0.0045

are able to work without knowing the label generating process. In fact, we have no idea of this process in reality.

5.2 Real-world Experiments

Finally, we do experiments on a real-world dataset called Music Genre. It was created and first used in [16], and it is publicly available [4]. It has 700 items, 10 classes, 44 workers, and 124 feature dimensions. LHS of Figure 4 shows histogram of the number of labels per worker, where we learn most workers give very few labels. In this experiment, we remove labels given by workers if he or she provides 40 or fewer labels. Following the same experimental setting as of synthetic experiments, results of algorithm comparison are shown in Table 3, where DRC-DS is better than IS and DRC-MV beats MV. It again shows the effectiveness of our DRC approach.

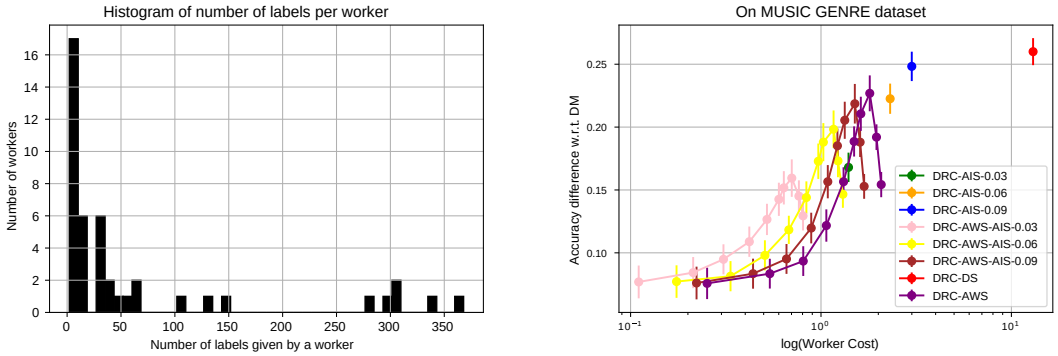

Figure 4: Histogram of number of labels per worker (LHS). Performance of DRC-DS, DRC-AIS, DRC-AWS, and DRC-AWS-AIS w.r.t. DM with increasing logarithmic worker costs (RHS).

Also, to show the effectiveness of AIS and AWS, we do experiments of DRC-AWS-AIS, DRC-AIS, DRC-AWS, and DRC-DS on Music Genre dataset. Experimental settings are similar to synthetic experiments. For DRC-AWS and DRC-AWS-AIS, the multiplier parameter λ is set to be 1, 2, 3, 4, 5, 6, 7, 9, 11, and 13. The results are shown in RHS of Figure 4. As we can see, DRC-AWS-AIS uses the least worker cost and it achieves

[4] `https://eden.dei.uc.pt/~fmpr/malr/`

relatively good performance. For AIS, as ρ increases, the performance becomes better. For all methods, if we put more worker cost, the accuracy goes up. The DRC-DS uses most worker cost and it achieves the best performance.

6 Conclusions

We formulate crowdsourcing a statistical estimation problem and propose a new approach DRC to address it where worker imitation and doubly robust estimation are used. DRC can work with any base models such as Dawid-Skene model and majority voting and improve their performance. With adaptive item/worker selection, our proposed approaches are able to achieve nearly the same accuracy of using all workers but with much less worker cost. In the future, there are many problems worth further research. Since item features are helpful for crowdsourcing problems, worker features can be taken into consideration as well. Also, if there are some new workers joining the project, how to deal with this kind of problem needs more study.

References

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017.

[3] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):27, 2011.

[4] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[5] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[6] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[8] Jonathan J Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[9] Hideaki Imamura, Issei Sato, and Masashi Sugiyama. Analysis of minimax error rate for crowdsourcing and its application to worker clustering model. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2147–2156, 2018.

[10] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

[11] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

[12] Ashish Khetan, Zachary C Lipton, and Animashree Anandkumar. Learning from noisy singly-labeled data. In *Proceedings of the International Conference on Learning Representations*, 2018.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[15] Jungseul Ok, Sewoong Oh, Yunhun Jang, Jinwoo Shin, and Yung Yi. Iterative bayesian learning for crowdsourced regression. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1486–1495, 2019.

[16] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436, 2013.

[17] Andrea Rotnitzky and James M Robins. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

[18] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[19] Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.

[20] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 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*, pages 614–622, 2008.

[21] Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, and Pietro Perona. Lean multiclass crowdsourcing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2723, 2018.

[22] Chong Wang, Xi Chen, Alexander J Smola, and Eric P Xing. Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems 26*, pages 181–189, 2013.

[23] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3589–3597, 2017.

[24] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pages 2424–2432, 2010.

[25] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5934–5938, 2018.

[26] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1):3537–3580, 2016.

[27] Dengyong Zhou, Qiang Liu, John C Platt, Christopher Meek, and Nihar B Shah. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240*, 2015.