The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Cost-Accuracy Aware Adaptive Labeling for Active Learning

Ruijiang Gao
University of Texas at Austin
ruijiang@utexas.edu

Maytal Saar-Tsechansky
University of Texas at Austin
maytal@mail.utexas.udu

Abstract

Conventional active learning algorithms assume a single labeler that produces noiseless label at a given, fixed cost, and aim to achieve the best generalization performance for given classifier under a budget constraint. However, in many real settings, different labelers have different labeling costs and can yield different labeling accuracies. Moreover, a given labeler may exhibit different labeling accuracies for different instances. This setting can be referred to as active learning with diverse labelers with varying costs and accuracies, and it arises in many important real settings. It is therefore beneficial to understand how to effectively trade-off between labeling accuracy for different instances, labeling costs, as well as the informativeness of training instances, so as to achieve the best generalization performance at the lowest labeling cost. In this paper, we propose a new algorithm for selecting instances, labelers (and their corresponding costs and labeling accuracies), that employs generalization bound of learning with label noise to select informative instances and labelers so as to achieve higher generalization accuracy at a lower cost. Our proposed algorithm demonstrates state-of-the-art performance on five UCI and a real crowdsourcing dataset.

Introduction

Supervised learning has achieved great successes over the years and has a significant impact on practice in a growing variety of predictive tasks. In many settings, however, labels for training instances are not readily available, but can be acquired from different labelers at different costs; often, different labelers may exhibit varying labeling accuracies, and a given labeler can have different labeling accuracies across different instances, possibly as a result of experience or prior knowledge. Given this setting and a model induction algorithm, it is important to understand how to best select labelers and instances they will label so as to induce a model with the highest generalization performance for a given labeling cost. In practice, such challenges arise in important applications, where scientists, medical professionals, or crowd of lay workers can be used to label a possibly large number of instances. In recent years, large-scale crowdsourcing and online labor market platforms, such as Amazon Mechanical Turk (AMT), have emerged to offer unprecedented scalability towards such tasks. Nevertheless, in the settings we consider here, selecting labelers' (and corresponding costs and accuracies) and the instances they will label to produce the best model at a given cost remains an open problem.

Traditional active learning (Lewis and Gale 1994; Gal, Islam, and Ghahramani 2017; Tang and Huang 2019) has received significant attention, and considers the problem of selecting instances for labeling when a single labeler produces labels at the same, fixed cost and with perfect labeling accuracy. Because labels of all instances are assumed to have the same accuracy and cost, traditional active learning frameworks aim to identify the most informative training instances from which to induce a model. However, in many real settings, acquiring labels from labelers presents greater complexity. AMT, for example, offers access to workers from around the world, with different expertise and varying costs. Indeed, prior social science research has shown that different payments lead to different qualities of work, and that different relationships between payment and quality can arise at different times or for different tasks (Mason and Watts 2009; Kazai 2011; Kazai, Kamps, and Milic-Frayling 2013).

More recent work considered multiple noisy workers, yet assumed that either all labelers exhibit the same quality (Ipeirotis et al. 2014; Lin, Weld, and others 2014; Donmez, Carbonell, and Schneider 2010), or that all labelers have the same cost per label and that the labeling quality is independent of the instance being labeled (Donmez, Carbonell, and Schneider 2009; Yan et al. 2011; 2014). The closest work to the problem we consider here is by Huang et al. (2017), where instance difficulty, labeler expertise, and varying costs across labelers are considered. Huang et al. (2017) use a different criterion to select labelers and instance than the one we develop here; as we discuss below, in the most common setting in practice, where labelers are not adversarial (Ipeirotis et al. 2014; Yan et al. 2011; Lin, Weld, and others 2014; Donmez, Carbonell, and Schneider 2009), this criterion appears to be prone to consistently choose low payment options, even when the labeling quality is poor and such choices undermine learning significantly.

In this paper, we propose a novel criterion utilizing gen-

eralization bound of learning with label noise to evaluate the cost-effectiveness of labeler-instance pair. The criterion we propose is motivated by the goal of directly minimizing the generation error of the classifier. To the best of our knowledge, this work is the first to use generalization bound for guidance of selecting labeler-instance pairs in this setting. We empirically evaluate the effectiveness and robustness of our method for settings with different cost-accuracy trade-offs reported in prior work to arise in crowdsourcing markets, and for five UCI datasets and a real crowdsourcing dataset. Our results show that our approach offers state-of-the-art performance across settings.

Related Work

Active learning has been studied extensively and is getting more important with the emergence of modern artificial intelligence. Most active learning research has considered settings where there is only one perfect labeler. Active learning algorithms for these settings thus aim to select the most informative training instances to label, so as to reduce the number of instances. However, when crowdsourcing platforms are used to acquire labels, annotation is done by different, noisy annotators, whose labels can be acquired at different costs and who may exhibit different levels of accuracies. Recently, more research has considered acquiring labels and learning from noisy labelers. Donmez, Carbonell, and Schneider (2009) and Zheng, Scott, and Deng (2010) estimate the accuracy rates of labelers and then selects for annotation labelers with high accuracies. Zhao, Sukthankar, and Sukthankar (2011) actively select instances for labeling, but do not select labelers. All these works assume that a labeler exhibits the same accuracy for all instances they label. Yet, in practice, different workers may have different expertise or prior experience and can consequently exhibit different accuracies when labeling different instances. Yan et al. (2011) propose a probabilistic framework to estimate workers' accuracies and select the worker estimated to be the most accurate for a given instance. Fang, Yin, and Tao (2014) and Ambati, Vogel, and Carbonell (2010) consider the different expertise of workers and aim to match instances with different worker with varying accuracies in the task domain; yet, these works do not consider varying labeling costs may be incurred by different labelers. Geva, Saar-Tsechansky, and Lustiger (2019) consider acquiring labels from labelers with varying accuracies and costs based on the estimated effect on generalization error, but do not consider selecting instances for labeling. The most closely related work is by Huang et al. (2017): the proposed method estimates workers' labeling accuracies based on a small set of ground-truth data, and then estimates the value of acquiring the label for a given instance from a given worker as the weighted labeling accuracy divide by worker's cost. As we discuss in more detail below and reflected in our empirical evaluations, in the most common setting in practice, when labelers are not adversarial (Ipeirotis et al. 2014; Yan et al. 2014), this heuristic criterion is prone to select low payments.

Some prior work considered the cost-effectiveness of majority voting by multiple labelers for same instance as compared to singly-labeled data. These works considered workers who exhibit the same accuracy and cost. Ipeirotis et al. (2014) shows how in some cases majority voting can improve the performance of given classifier for a given cost, and Lin, Weld, and others (2014) demonstrates how the optimal choice depends on the dataset, classifier, and labeling accuracy. Yet, these works do not address how to effectively trade-off performance and cost. Importantly for this work, the trade-off between acquiring a single or multiple labels per instance can also be viewed as a special case of having multiple labelers of varying costs and accuracies. Hence, a method that can effectively select amongst different labelers of accuracies can also apply to select whether or not to acquire multiple labels for a given instance. Other research explored the generalization error bounds for learning with label noise. Simon (1996) and Aslam and Decatur (1996) study the error bounds for learning from noisy labels for PAC-learnable concepts. Kearns (1998) develop a bound for concepts that are Statistical-Query-learnable. We rely on these theoretical results and propose a novel criterion to select instance-labeler pairs that minimize the generalization error and achieve state-of-the-art results.

Problem Statement

Suppose we have a dataset $D = \{x_i, y_i\}_{i=1}^{N}$, concept $\mathcal{C}$, an unlabeled set $\mathcal{U} = \{x_i\}_{n_l+1}^{N}$, and a set of labelers $\mathcal{A} = \{a_1, \ldots, a_n\}$, who exhibit costs $\{c_1, \ldots, c_n\}$ and label accuracies $\{\rho_1, \ldots, \rho_n\}$, respectively. In addition, suppose that an initial labeled dataset is available with ground-truth labels $\mathcal{L} = \{x_i, y_i\}_1^{n_l}$ that have also been labeled by each of the labelers in $\mathcal{A}$. We assume the most common settings where labelers are not adversarial, i.e. their labeling accuracy rates are higher than 50% (Ipeirotis et al. 2014; Yan et al. 2011; Lin, Weld, and others 2014; Donmez, Carbonell, and Schneider 2009).

Further, labelers can have different expertise for different instances. Thus, for example, in an image classification task, Amy may be an expert at identifying species of flora while Bob excels in identifying fauna. We illustrate this notion in Figure 1, where labelers, $L1$, $L2$, and $L3$, have diverse expertise across different image categories. If each image category has the same number of samples, the overall labeling accuracies for $L1$, $L2$, $L3$ are 0.83, 0.73, 0.66 respectively. Yet, each labeler exhibits higher (and lower) labeling accuracies on some of the categories. Experiments in online labor markets (Kazai 2011; Kazai, Kamps, and Milic-Frayling 2013; Mason and Watts 2009) reveal that often higher accuracies can be obtained by offering higher payments. For simplicity, in this work we consider the price levels of 3, 2, and 1, for labelers $L1$, $L2$, and $L3$, respectively.

Finally, we consider an iterative setting, where at each iteration, a labeler from $\mathcal{A}$ is selected for labeling a selected instance from $\mathcal{U}$. Given a limited budget B, we aim to acquire labels from labelers for certain instances so as to induce a classifier with the best generalization performance. Thus, an algorithm for selecting labelers and instances ought to decide from which labelers and for what instances to acquire labels so as to yield the best generalization perfor-

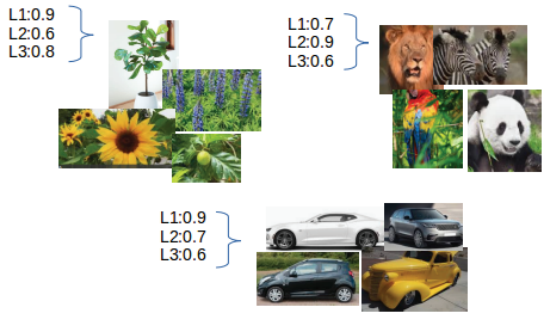

Figure 1: Diverse labelers' performance on different image categories: Flora, Fauna and Cars. Labels L1, L2, L3 demonstrate different labeling accuracies on different category, and their overall accuracies are 0.83, 0.73 and 0.66, respectively (assuming equal numbers of instances in each category).

mance with a given budget; for example, it ought to determine whether it would be more cost-effective to acquire flora labels from the more accurate (and pricier) L1 than acquiring lower accuracy labels for 3 different flora images from L3, thereby compiling a larger training set for learning at the same cost.

Algorithm

Recall that our approach aims to select instance-labeler pairs so as to improve the generalization performance for a given labeling budget. Below, we first discuss briefly how we quantify instance usefulness. Because the closest work to the contributions we present here is by Huang et al. (2017), we then outline the main elements of the CEAL algorithm (Huang et al. 2017) and then describe how our algorithm builds on this contributions.

Instance Usefulness

There are many algorithms that propose various criteria for selecting instances for labeling; the more prominent measures include: uncertainty sampling that uses the posterior probability of predicted class (Gal, Islam, and Ghahramani 2017; Lewis and Gale 1994), model expected change that selects the sample the affects the model most (Freytag, Rodner, and Denzler 2014) and data diversity that chooses the data that helps labeled pool better represent the underlying population (Sener and Savarese 2017; Nguyen and Smeulders 2004). In general, any measure for quantifying instance usefulness can be used in our approach. In this paper, our main focus is on a cost-effective selection of labelers, we simply follow the setting in (Huang et al. 2017) and use uncertainty sampling as shown in Equation 1. $P(y|x_j)$ is the posterior probability predicted by the classifier trained at current iteration.

$$r(x_j) = 1 - \max_{y \in \mathcal{Y}} P(y|x_j) \tag{1}$$

CEAL (Huang et al. 2017)

CEAL estimates annotators' labeling accuracies of given instance $x_j \in \mathcal{U}$ based on the accuracy of labelers' respective responses on the labeled set $\mathcal{L}$ with ground truth labels. Specifically, the accuracy of labeler i for instance x_j, $\rho_i(x_j)$ is a weighted mean of labeling accuracy, weighted by the similarity between x_j and each of its nearest neighbors $x_k \in \mathcal{N}(x_j)$, as shown in Equation 2, where $0 \leq s(x_k, x_j) \leq 1$, $\mathcal{N}(x)$ represents that nearest neighbors of x in $\mathcal{L}$ and $\sum_k s(x_k, x_j) = 1$.

$$\rho_i(x_j) = \sum_{x_k \in \mathcal{N}(x_j)} s(x_k, x_j) I[y_k == \hat{y}_{ik}] \tag{2}$$

The final instance-labeler pair is selected by Equation 3 and 4. At each iteration, the product $q_i(x_j)r(x_j)$ is computed for every instance-labeler pair $(x_j, a_i), x_j \in \mathcal{U}, a_i \in \mathcal{A}$, and the pair with maximum product is selected for labeling. As a result, CEAL tends to select samples that are quite useful, and labelers that are good enough, yet cheap.

$$q_i(x_j) = \frac{\rho_i(x_j)}{c_i} \tag{3}$$

$$(x^\star, a^\star) = \arg\max_{a_i, x_j} q_i(x_j) r(x_j) \tag{4}$$

However, this heuristic tends to select instance-labeler pairs that maximize *labeling* accuracy per cost, and it does not assess the implications of different labeling accuracies and costs on subsequent generalization performance. To illustrate how CEAL prioritizes amongst different labelers, suppose we have five labelers whose labeling costs are 1,2,3,4,5 respectively. We further suppose labelers are not adversarial (Ipeirotis et al. 2014; Yan et al. 2011; Lin, Weld, and others 2014; Donmez, Carbonell, and Schneider 2009), thus $\rho_i(x_j)$ is greater than 0.5 for all i. When a low cost labeler offers near-random accuracy, while all others offer perfect accuracy, it is easy to see that CEAL will always prefer the least costly labeler with near-random labeling accuracy.

$$\frac{\rho_i(x_j)}{1} \geq \frac{0.5}{1} \geq \frac{\max \rho(x)}{2} = \frac{1.0}{2} > \frac{1.0}{3} > \frac{1.0}{4} > \frac{1.0}{5} \tag{5}$$

Similarly, in the example shown in Figure 1, given all labelers' accuracies are above 0.5, CEAL will select the labeler with the lowest cost.

Proposed Algorithm

As we discussed above, in CEAL, the value of a labeler is quantified by the ratio between the labeler's labeling quality and cost, and as such does not necessarily reflect the impact on generalization performance as done in (Geva, Saar-Tsechansky, and Lustiger 2019). We seek to develop an algorithm that aims to address this: identify labeler-instance pairs that would have the greatest benefit to generalization performance per cost.

However, generalization error is intractable in most supervised learning problems. There are some prior works that

explore different ways to estimate it. In the context of traditional active learning, (Roy and McCallum 2001) proposes to use (empirical) Estimated Error Reduction (EER) as an estimation to the generalization error reduction to select useful training instances for labeling. (Settles, Craven, and Ray 2008; Freytag et al. 2013) maximize Expected Model Change (EMC) by evaluating expected changes in model parameters, but this approach lacks a theoretical connection to error reduction. (Freytag, Rodner, and Denzler 2014) proposes to use expected model output change as an upper bound to generalization error. While it gives no guarantee when we are interested in maximization, it shows good performance in active learning problems. However, all these approaches consider settings with a single and perfectly accurate labeler, and it is unclear how they can apply in our setting when multiple noisy workers are present.

Meanwhile, research on generalization error bound for different concepts when learning from noisy labels, provides an upper bound on the decreasing speed on the generalization error of concept classes of interest (Simon 1996; Aslam and Decatur 1996; Kearns 1998), but is rarely used in empirical research. The upper bound of the generalization error can also be thought of as incorporating uncertainty of the classification error. Inspired by the notion of guiding acquisition by expected error reduction (Geva, Saar-Tsechansky, and Lustiger 2019), we propose a novel criterion that utilizes the theoretical results on generalization bound for learning with noisy labels to select cost-effective labelers. Based on the generalization bound proposed in Theorem 1 (Simon 1996; Aslam and Decatur 1996), our algorithm combines theoretical analysis into active learning to select the labeler that minimize error as shown in Equation 8, where $\hat{\rho}_i$, and n_i denote the estimated label accuracies and number of samples labeler i can purchase under a fixed budget. The VC dimension of a classifier measures the size of the largest finite subset of points that it is capable of classifying correctly (shatter). A higher VC dimension thus corresponds to weaker inductive bias. For simplicity, we treat the fail probability term as a constant since the VC dimension is not known.

Theorem 1 *(Simon 1996; Aslam and Decatur 1996) PAC learning a function class $\mathcal{F}$ with Vapnik-Chervonenkis dimension VC($\mathcal{F}$) in the presence of classification noise ρ and fail probability δ requires a sample of size*

$$\Omega\Big(\frac{VC(\mathcal{F})}{\epsilon(2\rho-1)^2}+\frac{\log(1/\delta)}{\epsilon(2\rho-1)^2}\Big) \tag{6}$$

$$i^\star = \arg\min_i \frac{1}{(2\hat{\rho}_i-1)\sqrt{n_i}} \tag{7}$$

$$= \arg\max_i (2\hat{\rho}_i-1)\sqrt{n_i} \tag{8}$$

The cost-normalized benefit to generalization performance from selecting labeler i to label instance x_j can thus be captured by the term in Equation 9.

$$q_i(x_j) = \frac{2\rho_i(x_j)-1}{\sqrt{c_i}} \tag{9}$$

However, recall that we aim to select labelers that lead to lowest generalization error. We should therefore consider the expected *cumulative* accuracies for all options so far. Also, importantly, (Lin, Weld, and others 2014) shows how more accurate labels may be preferred early on the learning curve, while cheaper and noisier labels may be more cost-effective for learning when number of samples are sufficiently large. Methods such as CEAL and Equation 9 neglect such learning dynamics; hence, we propose a second, adaptive criterion shown in Equation 10, where ρ_0, n_0 is the estimated accuracy and number of instances so far, and b denotes the unit budget we consider for estimating the future expected generalization error.

$$q_i(x_j) = \Big(2\frac{\rho_0 n_0 + \rho_i(x_j)\lfloor\frac{b}{c_i}\rfloor}{n_0+\lfloor\frac{b}{c_i}\rfloor}-1\Big)\sqrt{n_0+\lfloor\frac{b}{c_i}\rfloor} \tag{10}$$

$$E(a_i, x_j) = r(x_j)q_i(x_j) \tag{11}$$

$$(x^\star, a^\star) = \arg\max_{a_i, x_j} E(a_i, x_j) \tag{12}$$

Similar to CEAL, we estimate labelers' accuracies using Equation 2. At each iteration, we calculate Equation 11 for each instance-labeler pair and select the pair with the highest value, as in Equation 12; after the label from the chosen labeler for the instance is acquired, the labeled instance is added to current training set, and next iteration begins. The acquisitions continue until the budget is exhausted. The complete Generalization Bound based Active Learning (GBAL) algorithm based on the criterion in Equation 9, and the Adaptive GBAL (AGB)[1] are shown in Algorithms 1, 2.

Algorithm 1 Generalization Bound based Active Learning **(GBAL)**

Input:
L: a small labeled set
U: the pool of unlabeled data for active selection
A: all possible labelers
$\hat{Y}$: the labels given by all labelers in A on L
repeat
  for each $x_j \in U$ and labeler a_i **do**
    calculate the uncertainty for x_j in Equation 1
    calculate expected generalization bound as in Equation 9 or 3
    calculate the effectiveness as in Equation 11
  end for
  Select the pair $(x^\star, a^\star)$ in Equation 12.
  Query the label of $x^\star$ from $a^\star$, denoted by $\hat{y}^\star$.
  $L = L \cup (x^\star, \hat{y}^\star)$; $U = U \setminus x^\star$.
  Train classifier on L and test it on test set.
until the budget is used up

[1]Code is availble in https://github.com/ruijiang81/AGB

Algorithm 2 Adaptive GBAL **(AGB)**

Input:
L: a small labeled set
U: the pool of unlabeled data for active selection
A: all possible labelers
$\hat{Y}$: the labels given by all labelers in A on L
b: unit budget for estimating Equation 10
Initialize:
$\rho = 1$
repeat
for each $x_j \in U$ and labeler a_i **do**
calculate the uncertainty for x_j in Equation 1
calculate the the expected generalization bound as in Equation 10
calculate the effectiveness as in Equation 11
end for
Select the pair $(x^\star, a^\star)$ in Equation 12.
$\rho = (\rho n + \hat{\rho}_\star(x_j))/(n+1)$
Query the label of $x^\star$ from $a^\star$, denoted by $\hat{y}^\star$.
$L = L \cup (x^\star, \hat{y}^\star)$; $U = U \setminus x^\star$.
Train classifier on L and test it on test set.
until the budget is used up

Experiment

We compare our methods with four baselines:

- **ALC:** (Yan et al. 2011) ALC selects the most uncertain sample from the unlabeled set and uses the most accurate labeler for annotation at each iteration.
- **CEAL :** (Huang et al. 2017) CEAL selects instance-labeler pair that maximize Equation 3.
- **All:** Select the most uncertain sample and use majority voting based on all labelers annotations for the sample at each iteration.
- **Random:** Select the most uncertain sample and randomly choose a labeler from the set of all labelers to annotate the sample at each iteration.

The main goal of the evaluation is to compare the effectiveness of labelers (costs and accuracies) chosen by different algorithms. The effectiveness of uncertainty sampling has been established in prior work (Huang et al. 2017; Lewis and Catlett 1994; Gal, Islam, and Ghahramani 2017). Algorithm 1 and 2 are referred to as **GB** and **AGB** in this section.

Label Simulation

We use the publicly available UCI datasets, and therefore we simulate the labels produced by different labelers for these datasets. The label generating process we use is similar to that in (Yan et al. 2011). Specifically, in order to create diverse labelers, we first create 30 clusters using KMeans (Jain, Murty, and Flynn 1999) for each dataset. In addition, as in (Huang et al. 2017), we simulate five labelers with cost levels: 5, 4, 3, 2, 1 which are associated with overall labeling accuracies from high to low, respectively. Each labeler is an 'expert" in some random set of clusters by exhibiting a high probability of correctly labeling instances from the corresponding cluster. In particular, the probabilities that a labeler correctly labels instances in her "expert" clusters are 0.95, 0.925, 0.9, 0.875, and 0.85; these probabilities for "non-expert" clusters are 0.61, 0.585, 0.56, 0.535, and 0.51 respectively. The resulting overall worker accuracies on UCI datasets are shown in Table 1. This process produces a diverse label distribution, where different labelers also incur different costs. As demonstrated in Table 1, given the KMeans produce different clusters for different datasets, a labeler of a given cost can yield different overall labeling accuracy across different datasets – this allows us to explore the robustness of our proposed algorithm under a wide variety of price-accuracy trade-offs.

Labeler	W1	W2	W3	W4	W5
Pen Digits	0.90	0.79	0.70	0.62	0.56
Audit	0.94	0.91	0.71	0.67	0.66
Mushroom	0.92	0.82	0.76	0.68	0.61
Spambase	0.95	0.81	0.79	0.71	0.58
German	0.93	0.87	0.74	0.68	0.57

Table 1: Worker Label Accuracy on UCI Datasets

UCI Dataset

We evaluated our approach using the following datasets: *German, Mushroom, Pen Digits, Spambase, Audit (Hooda, Bawa, and Rana 2018)* from UCI Machine Learning Repository (Bache and Lichman 2013). The statistics of these five datasets can be found in Table 5.

We divide each dataset into initial, train and test set, consisting of 5%, 65% and 30% of the data, respectively. The algorithm's performance will be better if the size of initial set is larger, but more data with ground truth is also harder to acquire. Logistic Regression is used as classifier in our experiments. We report classification accuracy on test set after each acquisition iteration. The results shown are averages over 20 runs.

Our main results are shown in Figure 2. We also report the average cost, query numbers and label accuracies in Tables 2, 3, and 4, respectively. As we discussed in the previous section, our results show that CEAL often selects the cheapest labelers, and the resulting noisy annotations can yield poor generalization error. ALC tends to select many expensive (and accurate) labelers and yields the highest label accuracy amongst all methods. These results also demonstrate that the most accurate labels may not be the most cost-effectiveness to acquire. A similar conclusion is also drawn in (Khetan, Lipton, and Anandkumar 2017; Snow et al. 2008; Geva, Saar-Tsechansky, and Lustiger 2019). The costs incurred by AGB and GB are relatively higher than the cost of random, but lower than the cost of ALC. This allows the AGB and GB methods to compile a larger number of instances with sufficient labeling accuracy to produce good generalization performance. AGB and GB perform quite well in all the tasks, and AGB outperforms all other methods in all datasets, suggesting that an adaptive estimate can offer a better assessment of the expected benefits

Methods	Audit	Pen Digits	Spambase	German	Mushroom
CEAL	1.04	1.01	1.01	1.03	1.01
Random	3.01	3.01	3.04	3.01	3.03
ALC	4.53	4.63	4.59	4.54	4.40
GB	2.97	3.14	1.50	2.68	2.30
AGB	3.26	4.12	2.57	3.59	3.67

Table 2: Average cost on UCI Datasets of different methods

Methods	Audit	Pen Digits	Spambase	German	Mushroom
CEAL	0.67	0.51	0.67	0.59	0.57
Random	0.81	0.70	0.76	0.74	0.71
ALC	0.95	0.94	0.93	0.92	0.79
GB	0.91	0.82	0.81	0.82	0.73
AGB	0.93	0.88	0.88	0.85	0.77

Table 3: Average Label Accuracy on UCI Datasets of different methods

Methods	Audit	Pen Digits	Spambase	German	Mushroom
CEAL	194.8	200	200	195.4	200
Random	68.3	68.2	67.8	68.1	67.4
ALC	46.0	44.7	45.3	45.6	47.3
GB	69.4	66.4	136.7	77.2	91.8
AGB	63.4	50.2	81.5	57.5	57.4

Table 4: Average Number of Queries on UCI Datasets of different methods

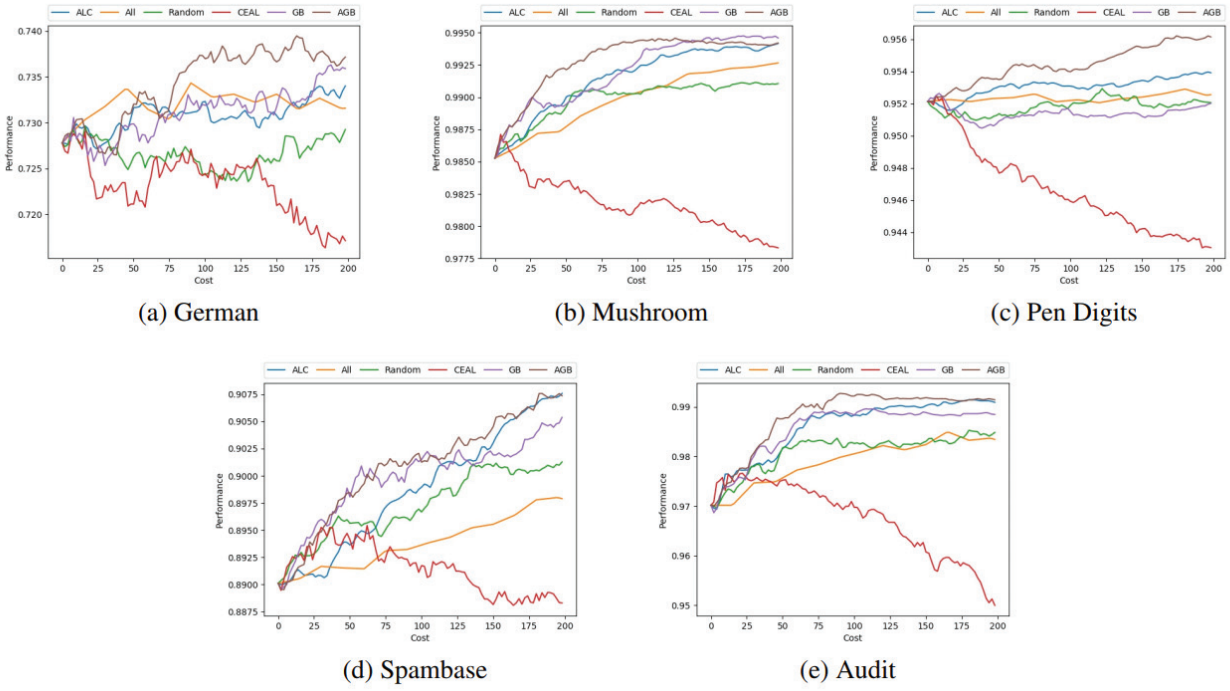

Figure 2: Cost-Accuracy Curves for Active Learning on UCI Datasets: we report accuracy after each iteration and X-axis represents cost so far in active learning. We can see AGB consistently outperforms other baselines. Results are averaged over 20 runs.

	#Instance	#Feature
German	1000	24
Mushroom	8124	117
Pen Digits	10992	16
Spambase	4601	57
Audit	776	24

Table 5: Statistics of five UCI Datasets

of different labelers.

Real Dataset

In addition to the simulated UCI datasets, we also performed experiment using a real crowdsourcing dataset. We use a sentiment analysis dataset, introduced in (Rzhetsky, Shatkay, and Wilbur 2009), and that includes 1000 sentences labeled by five *real* crowdsourcing workers. The annotators labeled each sentence along three dimensions: Focus, Polarity and Evidence. However, the overall accuracies of the five labelers along the Polarity and Evidence labels are all very high and thus not very diverse; we thus used only the Focus dimension, where labelers exhibit diverse accuracies. We henceforth refer to this as the Focus data set. We binarized the response variable and use bag-of-words features for training the model. After removing stopwords, the feature set consists of 292 features. The overall labeling accuracy of each real labeler is shown in Table 6. We simply set a labeler's cost to be the same as the labeler's accuracy in Table 6. As before, Logistic Regression is used as classifier and dataset is randomly split into 5%, 65% and 30% as initial, train and test set, respectively. Result on Focus, averaged over 30 runs, are shown in Figure 3.

Labeler	W1	W2	W3	W4	W5
Label Accuracy	0.82	0.931	0.892	0.904	0.641

Table 6: Label Accuracy on Focus Dataset

As shown, given the different price levels for the Focus dataset are very similar (around 0.9) for all workers, besides W5, the performance of AGB is similar to that of ALC and random. Indeed, when all price options are the same and workers' accuracies are similar, all methods will have the same effect. However, as shown in Figure 3, even in this setting AGB outperforms all other methods. This result is consistent with our results on the UCI datasets.

Conclusion

In this paper, we propose that the generalization bounds from theoretical analysis of settings with noisy labels can be effectively used to address the cost-effective active learning task with labelers of varying expertise and costs. We examine the shortcomings of existing algorithms proposed for this and other similar settings, and empirically demonstrate the effectiveness of our algorithms on various datasets. It is worth noting that our proposed algorithm can also apply to choose between singly labeling and the acquisition of multiple labels per instance for majority voting strategies, which

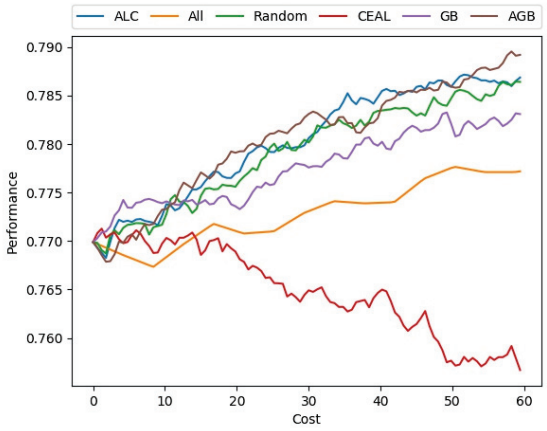

Figure 3: Cost-Effective Active Learning on Real Dataset , Results are averaged over 30 runs

we leave for future work. However, the optimal instance-payment selection ought to account for domain, concept class, and price-accuracy tradeoffs. We use a particular generalization bound as an upper bound which allows us to account for these elements through a model-data free criterion, though clearly not optimally. We leave the design for a more complex model-data dependent algorithm for future work.

References

Ambati, V.; Vogel, S.; and Carbonell, J. G. 2010. Active learning and crowd-sourcing for machine translation.

Aslam, J. A., and Decatur, S. E. 1996. On the sample complexity of noise-tolerant learning. *Inf. Process. Lett.* 57(4):189–195.

Bache, K., and Lichman, M. 2013. Uci machine learning repository.

Donmez, P.; Carbonell, J. G.; and Schneider, J. 2009. 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*, 259–268. ACM.

Donmez, P.; Carbonell, J.; and Schneider, J. 2010. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, 826–837. SIAM.

Fang, M.; Yin, J.; and Tao, D. 2014. Active learning for crowdsourcing using knowledge transfer. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Freytag, A.; Rodner, E.; Bodesheim, P.; and Denzler, J. 2013. Labeling examples that matter: Relevance-based active learning with gaussian processes. In *German Conference on Pattern Recognition*, 282–291. Springer.

Freytag, A.; Rodner, E.; and Denzler, J. 2014. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, 562–577. Springer.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1183–1192. JMLR. org.

Geva, T.; Saar-Tsechansky, M.; and Lustiger, H. 2019. More for less: adaptive labeling payments in online labor markets. *Data Mining and Knowledge Discovery* 1–49.

Hooda, N.; Bawa, S.; and Rana, P. S. 2018. Fraudulent firm classification: A case study of an external audit. *Applied Artificial Intelligence* 32(1):48–64.

Huang, S.-J.; Chen, J.-L.; Mu, X.; and Zhou, Z.-H. 2017. Cost-effective active learning from diverse labelers. In *IJCAI*, 1879–1885.

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.

Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31(3):264–323.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16(2):138–178.

Kazai, G. 2011. In search of quality in crowdsourcing for search engine evaluation. In *European Conference on Information Retrieval*, 165–176. Springer.

Kearns, M. 1998. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* 45(6):983–1006.

Khetan, A.; Lipton, Z. C.; and Anandkumar, A. 2017. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.

Lewis, D. D., and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*. Elsevier. 148–156.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Lin, C. H.; Weld, D. S.; et al. 2014. To re (label), or not to re (label). In *Second AAAI conference on human computation and crowdsourcing*.

Mason, W., and Watts, D. J. 2009. Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation*, 77–85. ACM.

Nguyen, H. T., and Smeulders, A. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 79. ACM.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through monte carlo estimation of error reduction.

Rzhetsky, A.; Shatkay, H.; and Wilbur, W. J. 2009. How to get the most out of your curation effort. *PLoS computational biology* 5(5):e1000391.

Sener, O., and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-instance active learning. In *Advances in neural information processing systems*, 1289–1296.

Simon, H. U. 1996. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences* 52(2):239–254.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. 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*, 254–263. Association for Computational Linguistics.

Tang, Y.-P., and Huang, S.-J. 2019. Self-paced active learning: Query the right thing at the right time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5117–5124.

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

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014. Learning from multiple annotators with varying expertise. *Machine learning* 95(3):291–327.

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

Zheng, Y.; Scott, S.; and Deng, K. 2010. Active learning from multiple noisy labelers with varied costs. In *2010 IEEE International Conference on Data Mining*, 639–648. IEEE.