The ACTIVECROWDTOOLKIT: An Open-Source Tool for Benchmarking Active Learning Algorithms for Crowdsourcing Research

Matteo Venanzi, Oliver Parson, Alex Rogers, Nick Jennings
University of Southampton
Southampton, UK
{mv1g10, osp, acr, nrj}@ecs.soton.ac.uk

Abstract

Crowdsourcing systems are commonly faced with the challenge of making online decisions by assigning tasks to workers in order to maximise accuracy while also minimising cost. To tackle this problem, various methods inspired by active learning research have been proposed in recent years. However such methods are typically evaluated against a limited set of benchmarks in different scenarios. As a result, no general comparative analysis of methods exists. Therefore it is not known which strategies dominate others across a range of domains or which strategies are best suited to specific domains. In this paper, we describe an open-source toolkit that allows the easy comparison of the performance of active learning methods over a series of datasets. Furthermore, the toolkit provides a user interface which allows researchers to gain insight into worker performance and task classification at runtime. This paper provides both an overview of the toolkit as well as a set of results evaluating 16 state-of-the-art active learning strategies over seven public datasets.

Introduction

Researchers who work in the area of crowdsourcing have invested significant effort into the problem of how to gain the best accuracy for a given task at the minimum cost. In general, this problem pertains to the area of active learning research which studies how to maximise the performance of artificial learners by intelligently picking the next best sample to learn from within all possible samples (Settles, 2012). However, in contrast to traditional active learning problems, adaptive crowdsourcing algorithms also need to reason about which worker should be allocated to the task. This selection must be driven by knowledge of the worker's accuracy and the estimated belief over the task's true label estimated from multiple workers' judgments[1] collected so far. As a result, an active learning strategy for crowdsourcing usually consists of a combination of three components: *(i)* a judgement aggregation model, *(ii)* a task selection method and *(iii)* a worker selection method.

[1]In our terminology, we use the term *label* to indicate a possible classification of a task, and the term *judgement* to indicate a label provided by a worker for a task.

In recent years, various active learning strategies have been proposed. These methods aim to intelligently select the most promising task to be crowdsourced next (i.e. the task for which an additional judgement will maximise the increase in overall accuracy), and the best worker to be allocated to the task (Costa et al., 2011; Zhao, Sukthankar, and Sukthankar, 2011; Kamar, Hacker, and Horvitz, 2012). A common approach is to use statistical models to infer the quantities of interest, such as the worker's accuracy and the task's aggregated answer from the set of judgments and subsequently use this inferred knowledge to inform the decision of which task and worker to select next (Bachrach et al., 2012; Bragg and Weld, 2013) or potentially remove workers considered as spammers from the worker selection (Welinder and Perona, 2010). As a result, existing methods vary in the types of models used to aggregate the workers' judgments and their strategies to perform task and worker selections. Crucially, all these methods agree that the proposed active learning strategies provide significant cost saving by achieving the same accuracy with significantly fewer judgments. However, their performance is typically evaluated against a limited set of benchmarks in different scenarios and no comparative analysis of a broad range of methods currently exists. As a result, it is not known which strategies dominate others across a range of domains, or which strategies are best suited to specific domains.

In addition, a number of public datasets have recently been proposed and employed for the evaluation of these methods, for example the Weather Sentiment dataset (Venanzi et al., 2014), Music Genre dataset (Rodrigues, Pereira, and Ribeiro, 2013), Sentiment Polarity dataset (Pang and Lee, 2004) and ZenCrowd dataset (Demartini, Difallah, and Cudré-Mauroux, 2012). New active learning strategies are often evaluated over one or two of these datasets, but little attention is paid to the domain of the dataset nor the individual characteristics of a dataset. For example, datasets collected from a domain in which the workers are required to make trivial decisions might not demonstrate the effectiveness of intelligent judgement aggregation models, while datasets with few judgements per worker might not demonstrate the effectiveness of worker selection methods. As such, the performance of an active learning strategy may be easily over or under-generalised without considering a number of datasets which vary over different domains and in size.

In this paper, we propose the open-source .NET ActiveCrowdToolkit – available on Github at `orchidproject.github.io/active-crowd-toolkit`. This toolkit allows the easy benchmarking of active learning strategies over a series of datasets and offers an extensive set of features for monitoring the performance of algorithms as they are executed in large-scale experiments. The toolkit aims to provide a number of features to aid researchers to reproduce, benchmark and extend the most prominent active learning and crowdsourcing algorithms, in a similar way to toolkits in other domains, such as NLTK for natural language processing (Bird, 2006), SQUARE or CrowdBenchmark for computing crowd consensus (Sheshadri and Lease, 2013; Nguyen et al., 2013) and NILMTK for energy disaggregation (Batra et al., 2014). Using our toolkit, we present the results of an extensive empirical comparison of the combination of five judgement aggregation models, two task selection methods and two worker selection methods, resulting in a total of 16 active learning strategies, evaluated over seven public datasets. Finally, we also release two new datasets for the Weather Sentiment and Sentiment Polarity domains, each of which were collected using the Amazon Mechanical Turk (AMT) platform, that provide new testbeds for crowdsourcing research. Thus, our contributions are summarised as follows:

- We propose the .NET ActiveCrowdToolkit for analysing the performance of active learning strategies in crowdsourcing environments. The toolkit provides interfaces for judgement aggregation models, task selection methods and worker selection methods, allowing new strategies to be constructed by combining novel and existing modules. Furthermore, our common dataset format allows existing active learning strategies to be easily evaluated over new datasets. In addition, our toolkit provides a user interface, allowing researchers to gain intuition into the inner workings of active learning strategies, such as the real-time belief updates over the label of individual tasks and the accuracies (confusion matrices) of individual workers.
- We release two datasets along with this paper: the Weather Sentiment dataset (WS-AMT) and the Sentiment Polarity dataset (SP-2015).[2] Both datasets were collected using the AMT platform for the same domain as existing datasets. Our dataset analysis shows that despite this similarity in domains, the datasets show significant differences in certain dimensions, which we also show contribute to variations in the performance of active learning strategies.
- We provide an empirical evaluation of 16 active learning strategies over seven datasets. Our results show that Bayesian aggregation models which capture the uncertainty in latent parameters provide the most accurate judgement aggregation, entropy-based task selection ensures the most efficient order for selecting the next task, and strategies which specifically select the best worker yield the greatest overall accuracy. Furthermore, our results highlight the variance between different datasets even when collected from the same domain. Finally, our results show that average recall is a preferable metric over the accuracy metric, especially when the number of gold labels in a dataset is highly skewed.

The remainder of this paper is structured as follows. We first describe a number of existing and novel datasets, paying specific attention to the domain and scale of the datasets. We then define an active learning strategy as a combination of a judgement aggregation model, task selection method and worker selection method, before describing how each element is implemented as part of the ActiveCrowdToolkit. Last, we present our empirical evaluation of 16 active learning strategies over seven datasets and conclude by discussing future work in the last section.

Datasets

A number of datasets have been used for the evaluation of the active learning strategies. In particular, we consider seven public datasets commonly used for crowdsourcing research, five of which were released by related work, and two of which are novel contributions of this work. We chose to collect additional datasets in order to determine whether differences between datasets were due to the scale or domain of such datasets. Importantly, *all* these datasets have gold-standard labels, (i.e., external labels provided by more reliable judges) for *all* the tasks. We provide details of the payment per task to each worker for our datasets and for existing datasets where this information is available in publications. Table 1 provides the scale of each dataset in terms of the number of tasks, workers, labels and judgements, while Figures 1 and 2 show the number of judgements per task and the number of tasks per gold label respectively for each dataset. The following sections describe the domain of each dataset and their relevant features.

Weather Sentiment - CF (WS-CF)

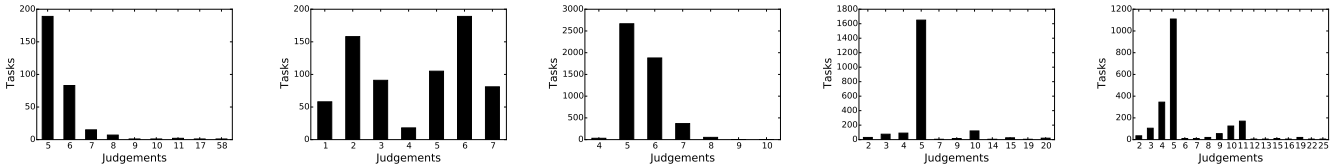
The Weather Sentiment dataset was provided by CrowdFlower for the 2013 Crowdsourcing at Scale shared task challenge.[3] The workers were asked to classify the sentiment of tweets with respect to the weather into the following categories: negative (0), neutral (1), positive (2), tweet not related to weather (3) and can't tell (4). The competition organisers did not release information regarding any restrictions of the worker pool or the worker payment per task. This dataset contains the least number of judgements but has the largest pool of workers, with each worker providing less than 4 judgements on average, as shown by Table 1. Furthermore, Figure 2 shows that the most common gold label is *unrelated*, while only five tasks were assigned the gold label *can't tell*. This suggests that, in this dataset, strategies which correctly label tweets in the *can't tell* class will have a much higher accuracy improvement than for tweets in any other class.

[2] Both the released datasets are available at `bit.ly/1EJdtBt`

[3] `https://www.kaggle.com/c/crowdflower-weather-twitter`

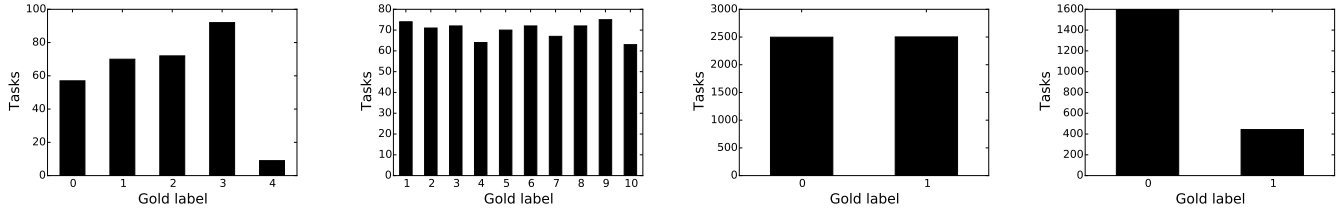
Dataset	Judgements	Workers	Tasks	Labels	Judgement accuracy	Judgements per task	Judgements per worker
WS-CF	1720	461	300	5	0.766	5.733	3.731
WS-AMT	6000	110	300	5	0.704	20.000	54.545
MG	2945	44	700	10	0.560	4.207	66.932
SP-2013	27746	203	5000	2	0.789	5.550	136.680
SP-2015	10000	143	500	2	0.893	20.000	69.930
ZC-IN	11205	25	2040	2	0.678	5.493	448.200
ZC-US	12190	74	2040	2	0.770	5.975	164.730

Table 1: Crowdsourcing datasets



(a) WS-CF (b) MG (c) SP-2013 (d) ZC-IN (e) ZC-US

Figure 1: Judgements per task. WS-AMT and SP-2015 are not shown as all tasks received exactly 20 judgements.



(a) WS (b) MG (c) SP (d) ZC

Figure 2: Tasks per gold label. Variants of the same task set (e.g. WS-CF and WS-AMT) are not shown individually as they contain identical distributions.

Weather Sentiment - AMT (WS-AMT)

We recollected the WS-CF dataset using the AMT platform. The workers were asked to complete the same task as in the CrowdFlower shared task challenge, although in this case many more judgements were collected from a smaller pool of workers for the same set of tasks. This was achieved by acquiring exactly 20 judgements per task, and as such the dataset contains many more judgements per task and judgements per worker than the CrowdFlower dataset, as shown in Table 1. We did not place any restrictions on the worker pool and each worker was paid $\$0.03$ per judgement. This dataset collected using AMT yielded a 0.06 lower judgement accuracy than the CrowdFlower dataset, which can be explained by the use of a different platform and system design. We also collected task acceptance and submission timestamps, which are not used in this work but might be of use to future research to inform the reliability of individual judgements.

Music Genre (MG)

The Music Genre dataset was collected by Rodrigues, Pereira, and Ribeiro (2013) using the AMT platform and contains samples of songs of 30 seconds in length taken from the audio dataset collected by Tzanetakis and Cook (2002). The workers were asked to listen to the music sample and classify the task into one of 10 music genres: country (1), jazz (2), disco (3), pop (4), reggae (5), rock (6), blues (7), classical (8), metal (9), hip-hop (10). Gold labels were assigned to each task by experts. The crowdsourcing platform required that workers were restricted to only those with an average task acceptance rate of higher than 95% across the whole AMT platform, although the worker payment per task was not released. The judgement accuracy for this dataset is the lowest of all our datasets, indicating the difficulty of each task. Figure 1 shows a bi-modal distribution of judgements per task, in which many tasks receive either 2 or 6 judgements, while few tasks receive 4 judgements. Figure 2 shows that the gold labels are equally distributed across the 10 genres. Thus, this dataset presents a

scenario of uniform classes of tasks where each correct classification provides the same accuracy gain.

Sentiment Polarity 2013 (SP-2013)

The SP-2013 dataset contains worker annotations using the AMT platform for sentences taken from movie reviews, extracted from the website Rotten Tomatoes (Pang and Lee, 2004). The workers were asked to classify the polarity of each sentence as either positive (1) or negative (0), with no option to express their uncertainty. Gold labels were assigned to each task by experts. As with the MG dataset, the crowdsourcing platform required that workers were restricted to only those with an average task acceptance rate of higher than 95% across the whole AMT platform, although the worker payment per task was not released. This is the largest dataset considered in this paper, containing 27,747 sentiment judgements for 5,000 tasks. Figure 1 shows that the vast majority of tasks receive 5 or 6 judgements, while Figure 2 shows that the tasks with positive and negative gold labels are equally weighted.

Sentiment Polarity 2015 (SP-2015)

We recollected the SP-2015 dataset also using the AMT platform. Similarly to WS-AMT, we did not set any requirements on each worker's qualifications and we paid \$0.03 per judgment. As with WS-AMT, we again collected 20 judgements per task including time information about the worker's submissions, although such judgements were collected over a subset of 10,000 tasks of the SP-2013 dataset. As a result, SP-2015 contains approximately 4 times the number of judgements per task compared with SP-2013, although it also contains only 10% of tasks of SP-2013. Table 1 shows that both SP datasets contain judgements of the highest average accuracy, indicating that the tasks are easier and require less subjective judgement than other domains. Furthermore, the average accuracy of SP-2015 was 0.11 higher than for the SP-2013 dataset. This could be due to a different task design that may have made the task easier, and also a reward scheme that might have attracted different types of workers.

ZenCrowd - India (ZC-IN)

The ZenCrowd India dataset contains links between the names of entities extracted from news articles and URIs describing the entity (Demartini, Difallah, and Cudré-Mauroux, 2012). The dataset was collected using the AMT platform, with each worker being asked to classify whether a single URI was either irrelevant (0) or relevant (1) to a single entity. Gold standard labels were collected by asking expert editors to select the correct URI for each entity. No information was released regarding the restrictions on the worker pool, although all workers are known to be living in India, and each worker was paid \$0.01 per judgment. A total of 11,205 judgements were collected from a small pool of 25 workers, giving this dataset the highest number of judgements per worker. Figure 1 shows that the vast majority of tasks receive 5 judgements, while Figure 2 shows a skewed distribution of gold labels, in which 78% of links between entities and URIs were classified by workers as *unrelated*. As such, it should be noted that any aggregation methods with a bias towards unrelated classification will correctly classify the majority of tasks and thus receive a high accuracy. Therefore it is necessary to select more than one performance metric (e.g., accuracy and recall) to evaluate different aspects of the methods.

	RT		ET	
	RW	**BW**	**RW**	**BW**
Majority vote	✓	✗	✓	✗
Vote distribution	✓	✗	✓	✗
Dawid & Skene	✓	✓	✓	✓
IBCC	✓	✓	✓	✓
CBCC	✓	✓	✓	✓

Table 2: Compatibility of active learning strategies

ZenCrowd - USA (ZC-US)

The ZenCrowd USA dataset was also provided by Demartini, Difallah, and Cudré-Mauroux (2012) and contains judgements for the same set of tasks as ZC-IN, although the judgements were collected from AMT workers in the US. The same payment of \$0.01 per judgement was used. However, a larger pool of 74 workers was involved, and as such a lower number of judgements were collected from each worker, as shown by Table 1. Figure 1 shows a similar distribution of judgements per task as the India dataset, although slightly fewer tasks received 5 judgements, with most of the remaining tasks receiving 3-4 judgements or 9-11 judgements. The judgement accuracy of the US dataset is 0.09 higher than the India dataset despite an identical crowdsourcing system and reward mechanism being used.

Active Learning Strategies

Having described our datasets, we now focus on the active learning problem of optimising the classification accuracy using the minimum number of judgments. In general, an active learning strategy consists of a single loop in which a judgement aggregation model is first required to update its estimates of the task labels and, in some cases, of the workers accuracies given a set of judgements. The task selection method then uses the output of the aggregation model to select the next task to receive a new judgement. Finally, the worker can be selected if the model also maintains its belief over the accuracy of each worker, otherwise the worker can be selected randomly to simulate the situation of no control over the worker assigned to the task. Table 2 shows how strategies can be formed as combinations of the three elements described in the following sections.

Judgement Aggregation Models

The judgement aggregation model is the algorithm that combines the set of collected judgments into a set of aggregated estimates of the true label of each task. Furthermore, to handle the uncertainty around the judgments, these estimates are

usually provided in terms of probabilities of a task to be assigned to each category.

A wide spectrum of aggregation models currently exists, which vary in terms of complexity and their ability to take into account different aspects of labelling noise in the aggregation process, such as worker accuracy and task difficulty. In particular, the most sophisticated aggregation algorithms use statistical models of the worker's reliability and filter unreliable judgments. Other simpler methods combine judgments assuming that all the workers are equally reliable. We select the following set of models in order to provide a balance between simple, intuitive aggregation models and more complex aggregation models. Furthermore, we focus on aggregation models which have a low computational overhead and are therefore able to make decisions within a few seconds, since these methods are most promising for deployment in real systems. Thus, we consider the following five judgement aggregation models:

- *Majority vote* is a simple yet popular algorithm that assigns a point mass to the label with the highest consensus among a set of judgments. Thus, the algorithm does not represent its uncertainty around a classification, and considers all workers to be equally reliable. (Littlestone and Warmuth, 1989; Tran-Thanh et al., 2013).
- *Vote distribution* assigns the probability of a label as the fraction of judgments corresponding to that label. Thus, the algorithm treats the empirical distributions of judgments as the estimate of the true label.
- *Dawid & Skene* is a well-known method that uses confusion matrices, i.e., matrices expressing the labelling probabilities of the worker conditioned on the task label values, to model the reliability of individual workers (Dawid and Skene, 1979). It treats both the workers' confusion matrices and the task labels as unobserved variables and applies an iterative expectation-maximisation inference algorithm to simultaneously estimate both these quantities from the judgment set.
- *Independent Bayesian Classifier Combination (IBCC)* learns the confusion matrices using a Bayesian inference framework that, in contrast to Dawid & Skene, considers uncertainty over the confusion matrices and the task labels (Ghahramani and Kim, 2003). It then applies an iterative variational method to compute approximate estimates over all latent variables.
- *Community-Based Bayesian Classifier Combination (CBCC)* is an extension of IBCC that models communities of workers with similar confusion matrices (Venanzi et al., 2014). This model is able to learn both the confusion matrices of each community and each worker as well as the true label of each task from the judgment set.

Importantly, the two Bayesian aggregation models (IBCC and CBCC) require several hyperparameters to be set. These hyperparameters regulate the prior belief of the algorithm over the confusion matrices and task labels. Following the same setting used in the original papers (Ghahramani and Kim, 2003; Venanzi et al., 2014), we set the hyperparameter of the diagonal counts of the confusion matrices to be higher (1.5) than the off-diagonal counts (1). This means that a priori the workers are assumed to be better than random. The other hyperparameters are set uninformatively. Furthermore, we run CBCC with two communities for all the datasets to allow the algorithm to learn two micro-classes of accurate and inaccurate workers.[4] Finally, Dawid & Skene was implemented using the open-source code available at `bit.ly/1zkExXf` in which, by default, the model inference of the confusion matrices is initialised with the accuracy of the majority votes.

Task Selection Methods

Different methods for selecting tasks given estimates of their labels can be devised depending on different approaches for measuring uncertainty in the label estimates. For instance, some approach consider the entropy, the label uncertainty, the model uncertainty or the combination of these elements as utility. Other approaches might use a more expensive forward planning framework to predict the expected utility of judgments for a particular task (Kamar, Hacker, and Horvitz, 2012). Here, we focus on approaches with low computational cost that can select tasks within a few seconds. We consider the following two methods:

- *Random task* (RT) selection is a simple method that selects tasks uniformly at random in a way that, if enough sequential rounds are allowed, all the tasks will have a uniform number of labels.
- *Entropy task* (ET) selection is the method that selects the most uncertain task with respect to the uncertainty measured by the entropy of the estimated label distribution. It seeks to reduce the uncertainty of the aggregation model by acquiring extra judgments for tasks for which the model has lower confidence of the true label.

Worker Selection Methods

A crucial step for an active learning strategy is to select the worker which will provide a new judgment for the selected task. Depending on the level of control of the crowdsourcing system, one can assign a task to a specific worker to reproduce the scenario of expert crowdsourcing (Tran-Thanh et al., 2012) or select workers at random to reproduce the scenario of AMT and similar platforms. Thus, many existing strategies are built around the two following methods:

- *Random worker* (RW) selects the workers uniformly at random to simulate the scenario where task requestors do not have direct control over task assignments such as in AMT.
- *Best worker* (BW) selects the workers with the highest estimated reliability of correctly labelling a task, calculated by taking the maximum of the diagonal of each worker's confusion matrix. However, it should be noted that other methods of estimating the best worker based

[4]The toolkit allows users to set any number of communities for CBCC, although we do not focus on exploring how active learning performance varies w.r.t the number of communities here.

on each worker's confusion matrix can be easily implemented with our toolkit. Notice we can only use this method with aggregation models that learn each worker's accuracy such as Dawid & Skene, IBCC and CBCC, as shown by Table 2.

It should be noted that while the best worker selection strategy is intuitively the most efficient, it also more problematic due the risk of potentially overloading a single worker with many tasks and the delays introduced by constraining the task to that worker. This makes it less practical to be applied to crowdsourcing processes.

ActiveCrowdToolkit

We have released the ActiveCrowdToolkit as open-source software, with the aim that the toolkit will be used and extended by the active learning community. The following sections describe the structure of toolkit and the toolkit's graphical interface.

Toolkit Structure

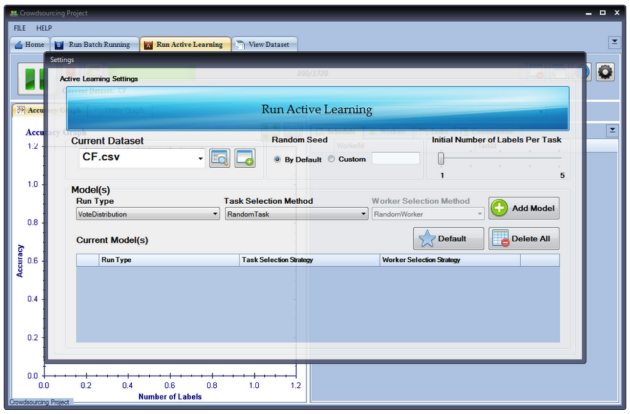
The toolkit has been structured to allow it to be easily extended with new datasets and active learning strategies. Datasets are described on disk in CSV format, in which the columns represent: task identifier, worker identifier, judgement label and gold label. The toolkit automatically discovers datasets in a specific directory, although external datasets can also be loaded into memory using the graphical interface. No data type constraints are placed on the values contained in the CSV files, allowing identifiers from the original crowdsourcing platform to be used if desired.

As described in the previous section, an active learning strategy consists of a combination of a judgement aggregation model, a task selection method and a worker selection method. By defining a single interface between these three components, strategies can be easily constructed by combining existing or novel components. Through the use of such interfaces, arbitrarily complex aggregation models can be implemented, for example the IBCC and CBCC models use the Infer.NET engine (Minka et al., 2013) to perform probabilistic inference over the task labels and worker confusion matrices.

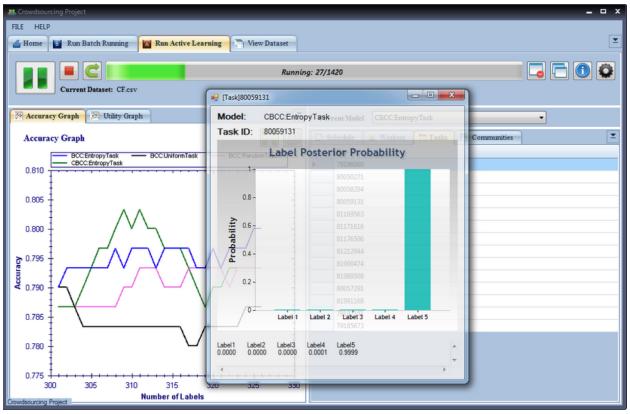
The toolkit has two interfaces: a command line interface and a graphical interface. The command line interface is designed to allow experiments to repeated many times across multiple machines, which we use to produce the results presented in the empirical evaluation section of this paper. The graphical interface is designed to provide researchers with the intuition behind the behaviour of active learning strategies, which we describe in the following section.

Graphical Interface

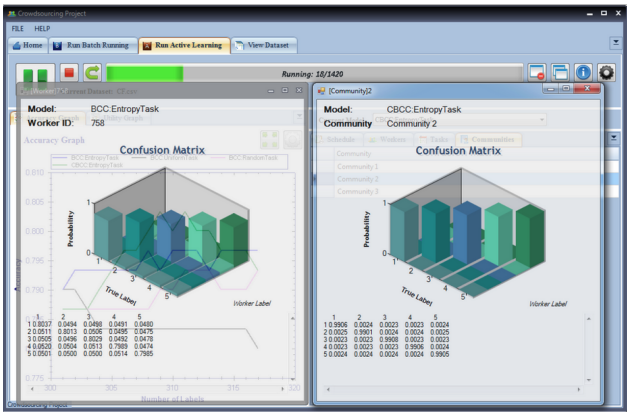
Figure 3 (a) shows the interface which allows researchers to set up experiments which run multiple active learning strategies over a single dataset. Using this dialog, the user can construct an active learning strategy by combining an aggregation model, a task selection method and a worker selection method. The user can also select the number of judgements each task should receive during an initial exploration phase, i.e., before the active learning process begins.



(a) Experiment set up.



(b) Real-time strategy accuracy and belief over task label.



(c) Confusion matrix of individual workers.

Figure 3: Screenshots of the ActiveCrowdToolkit's graphical interface.

Once an experiment has been started, the interface displays the accuracy graph of each strategy in real-time as judgements are selected from the dataset, as shown by Figure 3 (b). The figure also shows the ability to visualise the estimate over the true label of a task as provided by the ag-

gregation model. Figure 3 (c) demonstrates the ability to visualise the confusion matrix of individual workers over each label for models which include a worker model. This allows researchers to understand which workers consistently provide correct or incorrect judgments, and even which gold labels are repeatedly misclassified by an individual worker.

Empirical Evaluation

We now compare all possible active learning strategies as combinations of the five judgement aggregation models, two task selection methods and two worker selection methods. However, as shown in Table 2, the best worker method is not compatible with the majority vote and vote distribution models, since these models do not represent the performance of individual workers. This results in a total of 16 active learning strategies, each of which is evaluated over seven datasets. All strategies are initialised with one judgement per task for each dataset, in order to provide each strategy with some information upon which an intelligent strategy can be executed. We present results of the active learning phase which follows this initialisation, such that the first judgement to be selected by the strategy is actually the second judgement for that task. Each strategy is run up to 30 times on each dataset, for which we present the mean value as well as error bars representing the standard error in the mean. For each dataset, we present graphs of the following two standard metrics:

- *Accuracy* is the proportion of tasks that were classified correctly by the aggregation model.
- *Average recall* is the mean across all classes of the recall rate, defined as the fraction of positive instances of a given class that were correctly labelled.

Figures 4 and 5 show the accuracy and average recall for the five judgement aggregation models and two worker selection methods over seven datasets for entropy task selection and random task selection respectively. The following sections discuss the key findings as a result of differences in the datasets, judgement aggregation models, task selection methods and worker selection methods respectively.

Datasets

It can be seen that each of the seven datasets exhibit distinct behaviour. Most evidently, all strategies converge slowly on datasets with a low judgement accuracy (e.g. ZC-IN). Such datasets show a large difference between the performance of each strategy, indicating the benefit of intelligent judgement aggregation models. In contrast, all strategies converge quickly on datasets with a high judgement accuracy (e.g. SP-2015). Such datasets show little difference between the performance of various strategies. However, the fast convergence rate indicates the benefit of stopping early, allowing the cost of acquiring additional unnecessary judgements to be avoided. Furthermore, it can be seen that even datasets from the same domain show vastly different performance of active learning strategies. In the WS domain the difference in performance is due to differences in the total number of judgements and also the number judgements per task and judgements per worker, while in the SP domain the difference is due to the average accuracy of judgements.

Figures 4 and 5 show the accuracy and average recall after each judgement. Although accuracy represents a highly intuitive metric, it assigns more weight to gold labels which occur more frequently a dataset. In contrast, the recall metric weights each gold label equally, independent of the number of tasks per gold label. The difference between these metrics is most evident for highly skewed datasets, such as ZC-IN as shown in Figure 4 (f), in which the IBCC and CBCC strategies show decreasing accuracy while the average recall increases. Furthermore, the Dawid & Skene strategies show a high accuracy and low average recall, due to all tasks being classified as the most common gold label. As a result, we recommend the use of the recall metric where a single measure is required to compare the performance of active learning strategies.

Judgement Aggregation

Each graph in Figures 4 and 5 compares the performance of five judgement aggregation models. Consistent with previous work (Bragg and Weld, 2013; Venanzi et al., 2014), the IBCC and CBCC models consistently outperform the Dawid & Skene, Majority vote and Vote distribution models. This is due to their modelling of the skill of individual workers, as well as their ability to represent uncertainty over latent parameters.

It is also worth noting that although the Dawid & Skene model reaches its highest average recall at the endpoint of most datasets, it often requires a large number of judgements before the average recall rapidly increases. Such rapid changes in accuracy and average recall are a result the reclassification of all tasks as a new label. This long initial phase means that the model is likely to be competitive in scenarios where a large number of judgements are available before the active learning phase begins, but uncompetitive in cold-start scenarios. In addition, compared to IBCC and CBCC that accept prior distributions over their parameters, Dawid & Skene requires a good guess of initial parameters to provide good accuracy in its inference results. For example, for the MG dataset, the accuracy of the majority vote is close to 0.5 (see Table 1) and, given the default initialisation of the confusion matrices of Dawid & Skene with majority vote accuracy on the diagonal values, the model starts its inference assuming that the workers are almost random labellers that does not produce good quality results.

Task Selection

It can be seen by comparing the performance of similar strategies between the two figures that entropy task selection outperforms random task selection across all datasets. This is due to the faster convergence rate, which is most evident for the ZC-US dataset, in which the CBCC best worker strategy reaches its maximum average recall with less than 2,000 judgements through the entropy task selection method, while the same strategy requires nearly 10,000 judgements for the random task selection method.

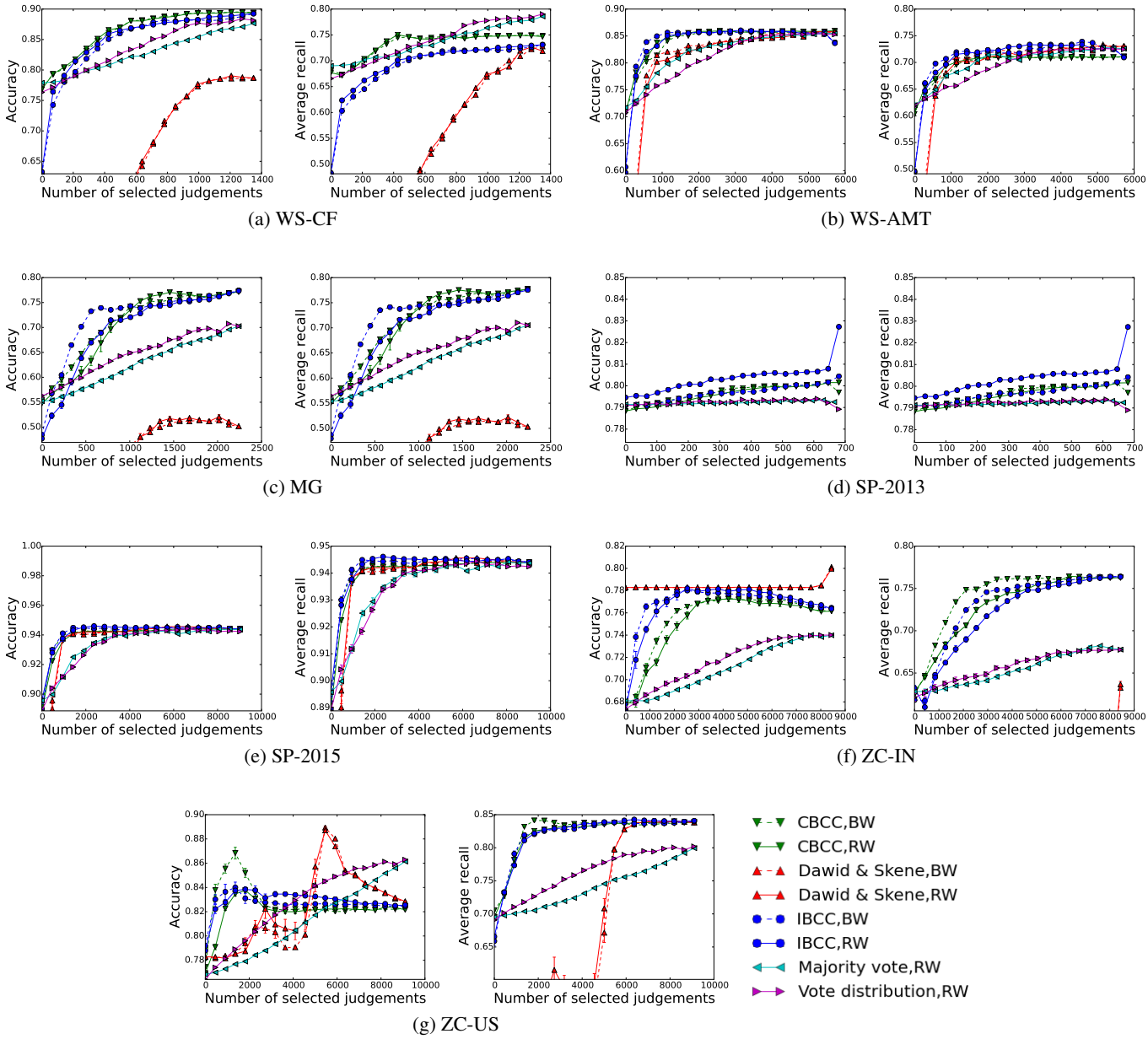


Figure 4: Accuracy graphs versus number of selected labels for the five models with *entropy task selection*.

Worker Selection

Figures 4 and 5 show that in almost all cases the best worker strategies outperform or perform comparably to the corresponding random worker strategies. As such, the ability to select the next worker allows convergence to be achieved earlier than through random worker selection, and as such can reduce crowdsourcing costs significantly in practice. However, it should be noted that there are multiple ways of implementing a best worker strategy as discussed in the active learning strategies section of this paper, and other implementations might yield a consistently superior performance. Furthermore, the ability to select specific workers might not be possible in many crowdsourcing systems.

Conclusions and Future Work

In this paper, we have proposed the ActiveCrowdToolkit as a tool for benchmarking active learning strategies for crowdsourcing research. As part of this toolkit, we have released two new crowdsourcing datasets characterised by a high redundancy in their judgment set. The toolkit allows active learning strategies to be easily constructed by combining judgement aggregation models, task selection methods and

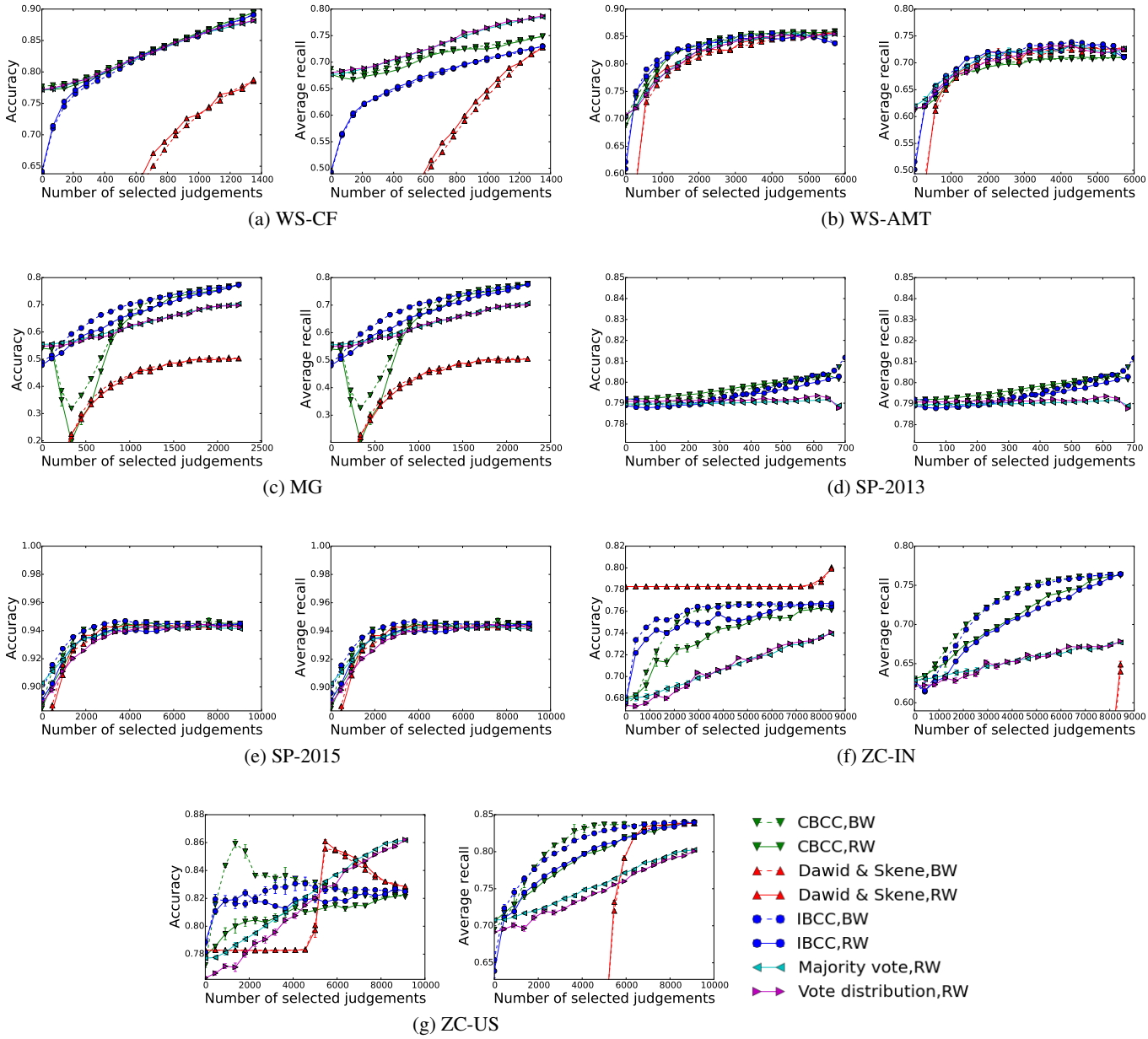


Figure 5: Accuracy graphs versus number of selected labels for the five models with *random task selection*.

worker selection methods, and evaluated as the strategy is executed using the toolkit's graphical interface. We have also used the toolkit to evaluate 16 active learning strategies across seven datasets. We have shown that existing datasets vary widely due to differences in domain, scale, worker restrictions and rewards, and we have also shown that such differences have a significant impact on the evaluation of active learning strategies. We evaluated each strategy using both accuracy and average recall metrics, and showed that the recall metric is most representative of performance in cases of unevenly distributed task gold labels. Our empirical evaluation showed that the IBCC and CBCC judgement aggregation models show the best performance across the range of datasets, while majority vote and vote distribution models converge more slowly to a constant accuracy level, and the Dawid & Skene model performs most competitively in scenarios with longer initialisations. In addition, we have shown that entropy-based task selection results in faster converge than random task selection, and also that the selection of the best worker to perform each task also results in faster convergence compared to random worker selection.

There are a number of potential directions for future work. First, integration of the ActiveCrowdToolkit with live crowdsourcing platforms, such as AMT or CrowdFlower,

would allow state-of-the-art active learning strategies to be deployed in real-time. Second, the trade–off between an initial uniform task selection phase followed by a more intelligent task selection phase has not yet been investigated. Third, the trade-off between the performance and run-time of active learning computation has not been investigated in this work. Fourth, although different strategies of our toolkit can be easily run in parallel via the command line, the graphical interface does not currently support multithreading or parallel computation.

References

Bachrach, Y.; Graepel, T.; Minka, T.; and Guiver, J. 2012. How to grade a test without knowing the answers — a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of the 29th Int. Conf. on Machine Learning (ICML-12)*, 1183–1190.

Batra, N.; Kelly, J.; Parson, O.; Dutta, H.; Knottenbelt, W.; Rogers, A.; Singh, A.; and Srivastava, M. 2014. NILMTK: An open source toolkit for non-intrusive load monitoring. In *Proceedings of the 5th international conference on Future energy systems*, 265–276. ACM.

Bird, S. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, 69–72. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bragg, J., and Weld, D. S. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing*, 25–33.

Costa, J.; Silva, C.; Antunes, M.; and Ribeiro, B. 2011. On using crowdsourcing and active learning to improve classification performance. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, 469–474.

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

Demartini, G.; Difallah, D. E.; and Cudré-Mauroux, P. 2012. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, 469–478. New York, NY, USA: ACM.

Ghahramani, Z., and Kim, H. 2003. Bayesian classifier combination. *Gatsby Computational Neuroscience Unit Technical Report GCNU-T., London, UK*.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 467–474.

Littlestone, N., and Warmuth, M. K. 1989. The weighted majority algorithm. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, 256–261. IEEE.

Minka, T.; Winn, J.; Guiver, J.; and Knowles, D. 2013. Infer.NET 2.6. Microsoft Research Cambridge. *See http://research. microsoft. com/infernet*.

Nguyen, Q. V. H.; Nguyen, T. T.; Lam, N. T.; and Aberer, K. 2013. BATC: a benchmark for aggregation techniques in crowdsourcing. In *Proceedings of the 36th international Conference on Research and Development in Information Retrieval (SIGIR 2013)*, 1079–1080.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013. Learning from multiple annotators: Distinguishing good from random labelers. *Pattern Recognition Letters* 34(12):1428 – 1436.

Settles, B. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6(1):1–114.

Sheshadri, A., and Lease, M. 2013. SQUARE: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2013)*, 156–164.

Tran-Thanh, L.; Stein, S.; Rogers, A.; and Jennings, N. R. 2012. Efficient crowdsourcing of unknown experts using multi-armed bandits. In *European Conference on Artificial Intelligence*, 768–773.

Tran-Thanh, L.; Venanzi, M.; Rogers, A.; and Jennings, N. R. 2013. Efficient Budget Allocation with Accuracy Guarantees for Crowdsourcing Classification Tasks. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13*, 901–908. International Foundation for Autonomous Agents and Multi-Agent Systems.

Tzanetakis, G., and Cook, P. 2002. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on* 10(5):293–302.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proc. of the 23rd Int. Conf. on World Wide Web*, 155–164.

Welinder, P., and Perona, P. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 25–32.

Zhao, L.; Sukthankar, G.; and Sukthankar, R. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *Privacy, security, risk and trust*, 728–733. IEEE.