Human Computation and Crowdsourcing: Works in Progress Abstracts:
An Adjunct to the Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing

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

We present an open-source toolkit that allows the easy comparison of the performance of active learning methods over a series of datasets. The toolkit allows such strategies to be constructed by combining a judgement aggregation model, task selection method and worker selection method. The toolkit also provides a user interface which allows researchers to gain insight into worker performance and task classification at runtime.

Introduction

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 aid researchers to reproduce, benchmark and extend state-of-the-art active learning methods for crowdsourcing systems, we developed the open-source .NET ActiveCrowdToolkit.[1] The toolkit offers a set of features for monitoring the performance of crowd consensus and active learning algorithms as they are executed in large-scale experiments,[2] thus extending the current features of other toolkits such as SQUARE and CrowdBenchmark for computing crowd consensus (Sheshadri and Lease 2013; Nguyen et al. 2013).

The toolkit design is based on a generalised representation of active learning strategies for crowdsourcing systems. This consists as a single loop in which a judgement aggregation model is first required to update its estimates of the task labels and 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, a specific worker can be selected if the model also maintains its belief over the accuracy of each worker, otherwise workers can be selected randomly to simulate the situation of no control over the worker assigned to each task. Table 1 shows how strategies can be formed as combinations of these components.

Datasets are described on disk in CSV format, in which the columns represent: task identifier, worker identifier, worker judgement label and, if available, gold label. The toolkit automatically discovers datasets in a specific directory, although external datasets can also be loaded 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.

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 or high performance clusters. The graphical interface is designed to provide researchers with the intuition behind the behaviour of active learning strategies. Figure 1 (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 active learning begins.

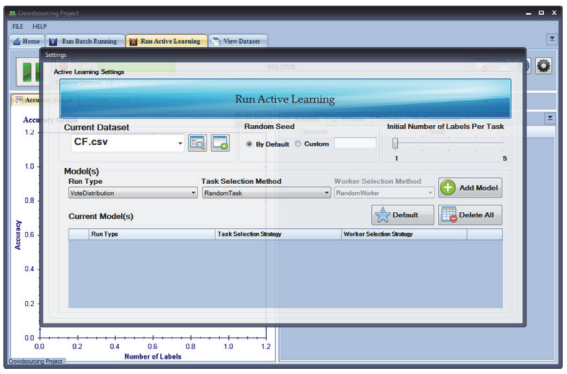
Once an experiment has been started, the interface displays the accuracy graph of each strategy in real-time as judgements are selected from the dataset, as shown by Figure 1 (b). The figure also shows the ability to visualise the estimate over the true label of a task as provided by the aggregation model. Figure 1 (c) demonstrates the ability to visualise the confusion matrix of individual workers over each label for strategies which include a worker model. This allows researchers to understand which workers consistently provide correct or incorrect judgements, and even which gold labels are repeatedly misclassified by an individual worker.

Toolkit Overview

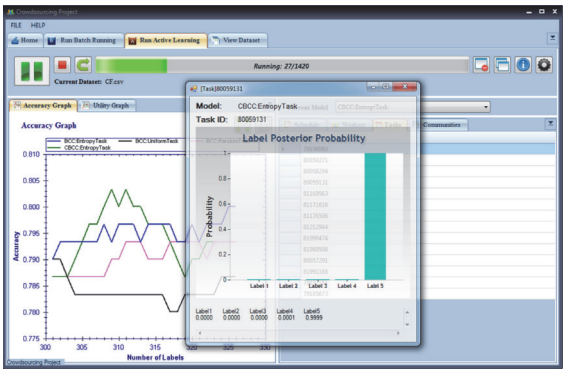
In the toolkit, an active learning strategy consists of a combination of a judgement aggregation model, a task selection method and a worker selection method. Currently, the toolkit supports five judgement aggregation models: *Majority vote* is a popular baseline that assigns a point mass to the label with the highest consensus among a set of judgements. *Vote distribution* assigns the probability of a label as the fraction of judgements corresponding to that label. *Dawid & Skene* is a well-known method that uses confusion matrices to model the reliability of individual workers (Dawid and

[1] orchidproject.github.io/active-crowd-toolkit

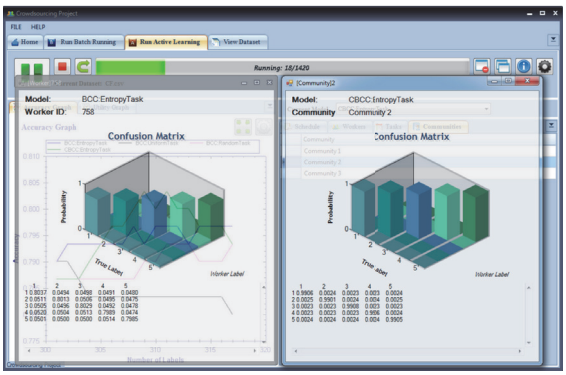
[2] See the full version of this paper for a description of experiments run using the toolkit: eprints.soton.ac.uk/380666



(a) Experiment set up.



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



(c) Confusion matrix of individual workers.

Figure 1: ActiveCrowdToolkit's graphical interface.

Skene 1979). *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 (Kim and Ghahramani 2012). *Community-Based Bayesian Classifier Combination (CBCC)* is an extension of IBCC that also learns communities of workers with similar confusion matrices (Venanzi et al. 2014).

In terms of task allocation strategies, the toolkit implements two standard 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.

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

Table 1: Compatibility of active learning strategies currently supported by the ActiveCrowdToolkit

Finally, the toolkit implements the following worker selection methods: *Random worker* (RW) selects the workers uniformly at random to simulate the scenario where the task requester does not have direct control over task assignments, such as in Amazon Mechanical Turk. *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. 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 1.

Conclusions

Our ActiveCrowdToolkit is a tool to support the benchmarking of active learning strategies for crowdsourcing research. This toolkit is under active development and we aim to keep extending this toolkit with new models.

References

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.

Kim, H., and Ghahramani, Z. 2012. Bayesian classifier combination. In *International Conference on Artificial Intelligence and Statistics*, 619–627.

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.

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.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web (WWW 2014)*, 155–164.