Activist: A New Framework for Dataset Labelling

Jack O'Neill[1], Sarah Jane Delany[1], and Brian MacNamee[2]

[1] Dublin Institute of Technology
jack.oneill1@mydit.ie,sarahjane.delany@dit.ie
[2] University College Dublin
brian.macnamee@ucd.ie

Abstract. Acquiring labels for large datasets can be a costly and time-consuming process. This has motivated the development of the semi-supervised learning problem domain, which makes use of unlabelled data — in conjunction with a small amount of labelled data — to infer the correct labels of a partially labelled dataset. Active Learning is one of the most successful approaches to semi-supervised learning, and has been shown to reduce the cost and time taken to produce a fully labelled dataset. In this paper we present *Activist*; a free, online, state-of-the-art platform which leverages active learning techniques to improve the efficiency of dataset labelling. Using a simulated crowd-sourced label gathering scenario on a number of datasets, we show that the *Activist* software can speed up, and ultimately reduce the cost of label acquisition.

1 Introduction

The availability of a large corpus of labelled training data is a key component in developing effective machine learning models. In many cases, such as speech recognition systems and sentiment analysis, labels are time-consuming or expensive to obtain, and must be provided by human annotators, constituting a bottleneck in the predictive model development life-cycle. Recent trends have seen an increased interest in using crowd-sourcing platforms such as CrowdFlower[3], and Amazon Mechanical Turk[4] to distribute the task of dataset labelling over a large number of anonymous oracles [21]. While crowd-sourced labels may reduce both the cost and time required to obtain a fully labelled dataset, further reductions may be realized by employing active learning to reduce the number of labels required.

The key insight behind active learning is that *"a machine learning algorithm can perform better with less training if it is allowed to choose the data from which it learns"* [16]. By allowing the active learning system to select the most informative data, and pose *queries* for labels for this data to the label provider, or *oracle*, the cost and time required to train an effective machine learning model can be greatly reduced.

[3] https://www.crowdflower.com
[4] https://www.mturk.com/mturk/welcome

Although the actual utility of a label may not be known in advance, an active learning system may employ one or more heuristics to predict the utility of querying for a particular label. This decision-making process, or *selection strategy*, is a key component of the active learning process. An active learning system begins with a small amount of pre-labelled, — or *seed* — data, and proceeds in iterations. Through its selection strategy, the system generates a query for a batch of labels from the unlabelled data. These labels are provided by the oracle, and the data is added to the labelled set. The process continues until a pre-determined *stopping criterion* is reached. A stopping criterion may be a straightforward label budget, or a more complex prediction of the marginal utility of each new label. Once this stopping criterion is met, a predictive model is trained using the set of labelled data. While active learning is primarily used in the context of predictive model generation, these same principles may be applied to a dataset labelling task. The process is carried out as above, but the resulting model is used to predict the labels of the remaining unlabelled data. The output of an active labelling task is, then, a fully labelled, approximately correct dataset.

A dataset labelling task may be seen as an instance of active learning in a pool-based setting, i.e. a setting in which the learner has access to a large, static pool of unlabelled instances from which to generate label requests. By submitting some, but not all data to oracles for labelling, the goal of the active learning system in this context is to reduce the cost accrued and time spent per correct label acquired, while maintaining accuracy.

This paper presents *Activist*, an extensible framework which assists users in all aspects of the data labelling process. As well as giving users the ability to configure an active labelling task, *Activist* provides a front-end UI for providing labels to the active learning system. The system covers all aspects of the dataset labelling process from loading and pre-processing the data, to creating a fully labelled output dataset once the process is complete. In addition to assisting users in producing fully labelled datasets, *Activist* allows multiple active learning strategies to be compared on simulated dataset labelling tasks, creating a detailed performance analysis for each approach under examination.

In this paper we describe the *Activist* system, and show how it can be used in an evaluation investigating the cost-benefit of applying active learning to a number of dataset labelling tasks. We show that while the impact of active labelling varies depending on the task, an active labelling approach consistently outperforms full dataset labelling.

The rest of the paper is structured as follows: Section 2 discusses related research in the areas of active learning and cost-sensitive labelling; Section 3 describes the *Activist* framework, and how it can be used to support the active learning process; Section 4 evaluates the use of *Activist* on a number of datasets, exploring the cost-benefits of applying active learning to a dataset labelling task; finally, Section 5 discusses the findings, suggesting avenues for future research.

2 Related Work

This paper examines the use of active learning in a pool-based setting, i.e. a setting in which the learner has access to a large, static pool of unlabelled instances from which to generate label requests. The problem of pool-based active learning was introduced by Lewis and Gale [10] in response to the need to develop text classification models for document retrieval. One of the key components which differentiates approaches to active learning is the selection strategy — the heuristic used to predict the informativeness of a particular label. Initial approaches to selection strategies favoured some measure of uncertainty sampling [3, 4], selecting those instances for labelling which are closest to the decision boundary of the model, *i.e.* those which the model was most likely to classify incorrectly.

An alternative selection strategy to uncertainty sampling is the Query-By-Committee (QBC) approach, introduced by Seung *et al.* [17]. QBC describes a general approach in which a number of diverse classifiers are trained on the currently labelled data, such that the classifiers can be expected to produce slightly different results for each unseen instance. The learner then measures the level of disagreement between the classifiers for each unlabelled instance and selects those instances which induce the highest level of disagreement between the classifiers in the committee. Variations on the QBC algorithm continue to be popular in the literature [5, 11].

Although measures of diversity have often been incorporated into other active learning selection strategies [2, 8], diversity measurements were first proposed as a sole metric in a selection strategy by Baram *et al.* [1]. Their Kernel-Farthest-First diversity algorithm seeks to label those instances which are least similar to the currently labelled data. Diversity, as a selection strategy, has been shown to work well in text classification [7], and in regression problems [13].

Research has shown that the labelling process of text-based classification datasets may be made more efficient by using visualisations to assist in the labelling process [19] or by using machine learning techniques to reduce the number of labels required of the annotator [14, 9]. Active learning has also been shown to improve the efficiency of dataset labelling for image classification [12], while the availability of commercial platforms such as CrowdFlower attest to the viability of active learning as a dataset labelling tool.

For a more in-depth discussion of the components comprising an active learning system (*e.g.* selection strategies, stopping criteria, *etc.*) see [16, 8].

3 Activist

The *Activist* Framework provides an end-to-end solution for dataset labelling tasks. Using *Activist*, the dataset labelling process consists of 4 stages:loading, pre-processing, labelling and output. The life-cycle of an Activist task is illustrated in Figure 3.

The simplest data format understood by *Activist* is the comma-separated values (csv) file. However, for many real-world problems (image or document

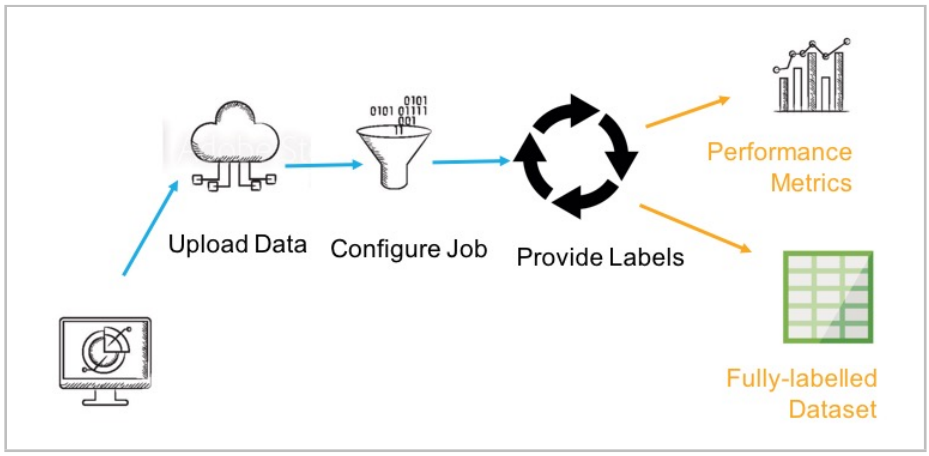

Fig. 1. Flow diagram illustrating the life-cycle of an *Activist* task

classification), data is not always available as a csv in its raw form. *Activist* provides a number of parsers capable of converting raw data into a structured dataset for processing by a machine learning algorithm. *Activist* allows users to transform images in the Portable Network Graphics (.png) format to multi-channel pixel maps, and collections of text files to bag-of-words representations. Depending on the size of the data and resources available, datasets may be stored in-memory or using a Redis No-SQL database.

Pre-processing is an important step in dataset generation. Pre-processing is used either to generate new aggregate features, or to remove uninformative features from a dataset. The activist framework gives users the capability to perform commonly required standard pre-processing on image or document-based datasets, such as word-stemming, stop-word removal, feature-filtering and row and column-wise pixel-map aggregation.

The heart of the *Activist* software is the dataset labelling loop. Having chosen a dataset, users can construct an active learning task to assist in the dataset labelling process. *Activist* offers a range of seeding strategies, selection strategies, stopping criteria, and predictive models to create an active learning system. Once the active learning task is initialised, the dataset labelling loop begins and visual representations of the data, such as the original image or document requested, are then presented to the user for labelling. Once the stopping criterion has been fulfilled, the *Activist* framework trains a prediction algorithm using the previously supplied labelled data, and predicts the most likely labels for the remaining unlabelled data, producing a fully labelled dataset in csv format.

By allowing users to simulate the labelling process (in cases where the true labels are available from the outset), the *Activist* software gives researchers the ability to perform evaluative experiments to compare the effectiveness of active learning options — such as selection strategies or stopping criteria — on a given

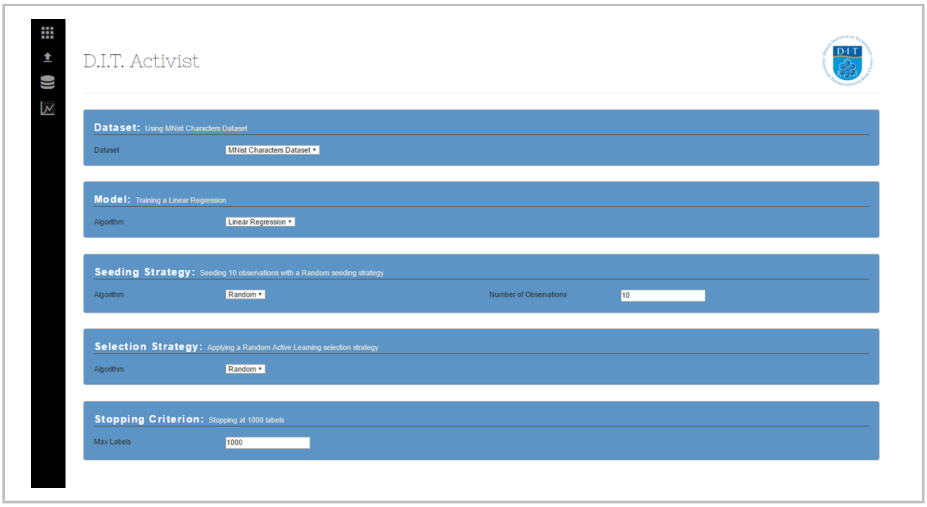

Fig. 2. A screenshot of the Activist System, showing the task configuration options

dataset. Labels are hidden from the system until requested. After each batch of label requests is issued, the chosen predictive model is trained and used to predict the labels of the remaining data. Accuracy and execution times are recorded and returned to the researcher as a csv file when the process is complete, allowing for direct comparison of multiple approaches.

4 Evaluation

The aim of the evaluation is to explore the potential of the *Activist* framework to reduce the number of manually required labels needed to produce a fully labelled dataset. This section describes the data and methodology used in the experiment, and reports the findings.

4.1 Datasets Used

Three datasets were used in this experiment, the MNist handwriting recognition dataset, the CIFAR-10 image classification dataset and the 20 Newsgroups document classification dataset. The MNist dataset[5] consists of 50,000 28x28 pixel gray-scale images of hand-written digits between 0 and 9. Each image is represented as a pixel map containing the value of each pixel as an unsigned byte. Another image classification dataset, CIFAR-10[6] consists of 60,000 32x32 colour images in 10 equally distributed classes, indicating the content of the

[5] `http://yann.lecun.com/exdb/mnist/`
[6] `https://www.cs.toronto.edu/~kriz/cifar.html`

image — all subcategories of vehicles and animals, for example: airplane, automobile, bird, cat, dog *etc.*. Images are represented as a pixel graph containing RGB values for each pixel as unsigned bytes. Rather than using the raw pixel values directly, individual pixels were aggregated into row and column totals for each colour channel, resulting in a vector of 192 features. The 20 Newsgroups[7] dataset is a freely available document classification dataset, consisting of approximately 20,000 documents partitioned approximately evenly across 20 different newsgroups. Each document was represented as a bag of words. The data was stemmed, with stop words removed, and words occurring in fewer than 3 separate documents removed as part of the data pre-processing stage. In order to reduce dataset size and the problem complexity, a subsection of the data containing 5 of the 20 newsgroups, — *alt.atheism*, *comp.windows.x*, *rec.autos*, *sci.space*, *talk.politics.guns* — was chosen.

4.2 Experimental Methodology

The active learning approach used in these experiments was set up using the *Activist* framework. As part of the task configuration, choices need to be made for the active learning components used in this the task: seed data, a batch size, a selection strategy, a stopping criterion and a predictive model algorithm. The following system was used for each of the datasets under consideration.

Seed Data: 50 initial labels were randomly selected and provided to the active learning system as seed data

Batch Size: To keep the batch sizes roughy proportional to the size of the datasets, the MNist dataset used a batch size of 10, while the CIFAR-10 and 20 Newsgroups datasets were evaluated with a batch size of 50

Stopping Criterion: The active learning loop was run until no unlabelled data remained, with performance recorded after each batch was complete.

Selection Strategies: A Query-by-Committee algorithm, using a committee of 5 k-nearest neighbour models was created, using k=5, with each committee member trained on a subset consisting of 80% of the data, selected randomly with replacement. An alternative, diversity-based selection strategy was also employed, using cosine distance as its distance metric. Finally, a random selection strategy, which makes no effort to select the *best* labels for querying, was evaluated as a baseline for selection strategies.

Predictive Model: A k-nearest neighbour predictive model with k=5 was used to classify the remaining unlabelled data, after each iteration.

After each new batch of labels was added to the labelled dataset, a predictive model was trained using the currently labelled data, and used to predict the

[7] `http://qwone.com/~jason/20Newsgroups/`

labels of the remaining unlabelled data. The number of correct labels (labels provided by oracle + correctly predicted labels) was recorded at each step.

4.3 Findings

Figure 3 shows the results of the experiment. After each batch of labels was requested, a predictive model was trained using the currently labelled data, and used to predict labels for the as-yet unlabelled data. The overall accuracy is recorded on the y-axis while the number of labels provided by the oracle is recorded on the x-axis. The black dashed line represents the accuracy obtained in the absence of an active labelling system *i.e. the number of labels provided by the oracle*. The difference on the y-axis between the dashed and solid lines represents the accuracy-gain provided by the active labelling framework.

The MNist dataset demonstrates that *Activist* can significantly improve the labelling rate of some datasets. Although less pronounced, the CIFAR10 and Newsgroups datasets benefit from employing active labelling techniques. These results also show that the benefit gained from active labelling is dependent on the characteristics of the dataset being used on the related prediction problem.

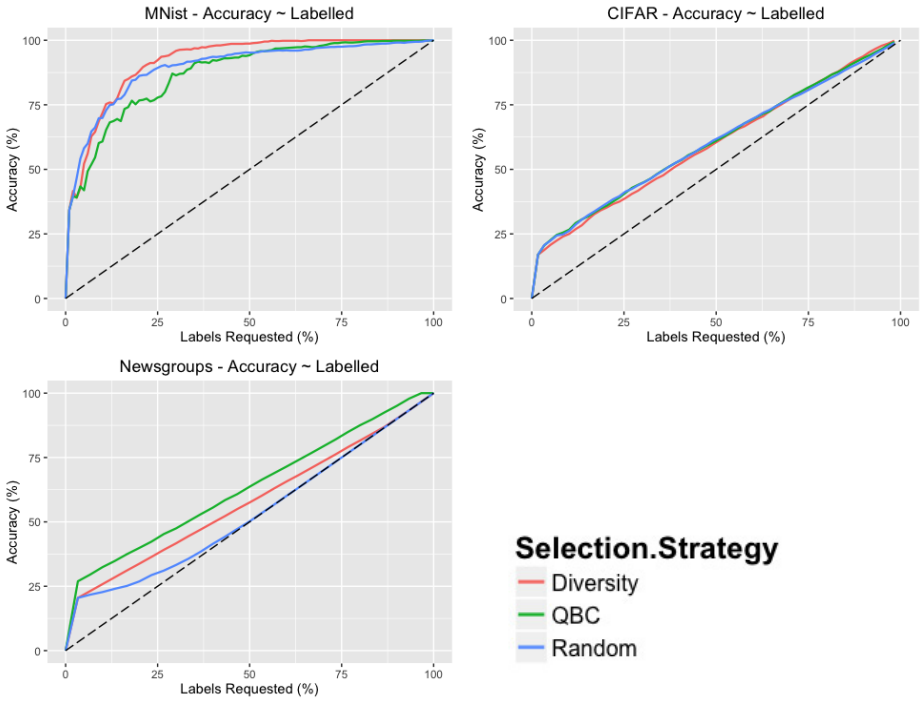

Fig. 3. Graphs showing the accuracy achieved per labels requested on each of the datasets examined. The dashed black line represents the number of correct labels in the absence of an active labelling system.

The results show that in all cases, a random selection strategy can yield demonstrable performance benefits over manual labelling, represented by the x=y baseline. This indicates that, although the performance of the *Activist* system differs depending on the selection strategy chosen, applying active learning techniques to dataset labelling yields a visible performance improvement irrespective of the particular selection strategy used.

5 Conclusions and Future Work

This paper presented *Activist*, a platform for applying active learning techniques to the problem of dataset labelling. Activist reduces the amount manual dataset labelling required to produce a fully labelled, approximately correct dataset. The *Activist* platform is under active development and is available for download online[8].

This evaluation has demonstrated the potential benefits of applying active learning to dataset labelling. Future work will expand the capabilities of the framework to further facilitate labelling large datasets. In order to take advantage of the benefits of crowd-sourced labelling, future work will incorporate an API to allow users to obtain labels from on-line crowdsourcing platforms.

The *Activist* framework will be expanded to include a wider variety of active learning components, particularly predictive models. Convolutional neural networks have been shown to be effective at classifying the CIFAR10 dataset [18, 6], while SVMs have been shown to work well classifying the 20 newsgroups dataset [15]. The inclusion of a wider range of predictive models is anticipated to yield a greater benefit for a larger number of datasets.

In its current format, the *Activist* system relies on a single label per instance. This approach is known to be problematic due to errors or subjectivity in the labelling process. Strategies for coping with this problem have been discussed in further detail by Tarasov [20]. Future work will aim to allow the *Activist* system to handle multiple responses per instance in an effort to mitigate the impact of subjectivity and rater unreliability on the labelling process.

The experiment has shown that the performance of active labelling depends to some extent on the selection strategies used. This suggests that a deeper investigation of the relative impact of all active learning components may prove promising. In addition to adding a wider range of components to the *Activist* platform, we hope to develop heuristics which will guide users in tailoring an active learning task to the problem at hand.

References

1. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. The Journal of Machine Learning Research 5, 255–291 (2004)
2. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: ICML. vol. 3, pp. 59–66 (2003)

[8] `https://github.com/joneill87/Activist`

3. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of artificial intelligence research (1996)
4. Engelson, S.P., Dagan, I.: Minimizing manual annotation cost in supervised training from corpora. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics. pp. 319–326. Association for Computational Linguistics (1996)
5. Gilad-Bachrach, R., Navot, A., Tishby, N.: Query by committee made real. In: Advances in neural information processing systems. pp. 443–450 (2005)
6. Graham, B.: Spatially-sparse convolutional neural networks. CoRR abs/1409.6070 (2014), http://arxiv.org/abs/1409.6070
7. Hu, R.: Active learning for text classification. Dublin Institute of Technology (2011)
8. Hu, R., Delany, S.J., Mac Namee, B.: Egal: Exploration guided active learning for tcbr. In: International Conference on Case-Based Reasoning. pp. 156–170. Springer (2010)
9. Hu, R., Mac Namee, B., Delany, S.J.: Sweetening the dataset: Using active learning to label unlabelled datasets (2008)
10. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 3–12. Springer-Verlag New York, Inc. (1994)
11. Li, S., Xue, Y., Wang, Z., Zhou, G.: Active learning for cross-domain sentiment classification. In: IJCAI (2013)
12. Li, X., Wang, L., Sung, E.: Multilabel svm active learning for image classification. In: Image Processing, 2004. ICIP'04. 2004 International Conference on. vol. 4, pp. 2207–2210. IEEE (2004)
13. O'Neill, J.: An evaluation of selection strategies for active learning with regression (2015)
14. Palmer, A.M.: Semi-automated annotation and active learning for language documentation (2009)
15. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: ICML. pp. 839–846. Citeseer (2000)
16. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(1), 1–114 (2012)
17. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 287–294. ACM (1992)
18. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
19. Sun, Y., Leigh, J., Johnson, A., Lee, S.: Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In: International Symposium on Smart Graphics. pp. 184–195. Springer (2010)
20. Tarasov, A.: Dynamic estimation of rater reliability using multi-armed bandits (2014)
21. Yuen, M.C., King, I., Leung, K.S.: A survey of crowdsourcing systems. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. pp. 766–773. IEEE (2011)