Efficient Active Learning for New Domains

Kiri L. Wagstaff kiri.l.wagstaff@jpl.nasa.gov
Steven Lu you.lu@jpl.nasa.gov
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109, USA

Abstract

The promise of active learning is to reduce the number of labeled examples required by supervised machine learning algorithms. The largest potential benefits lie in entirely new domains, for which no labeled examples yet exist. Yet to date, most active learning studies are retroactive and demonstrate the benefits that *could* have been gained if active learning had been used. What are the barriers to true adoption and utilization of active learning? We focus on two: (1) the cold start or class discovery problem, in which active learning methods may struggle to make progress with zero labeled examples, and (2) the cost of having the classifier in the loop to select the next example to be labeled. We assess different active learning approaches in the context of these two barriers and conclude with recommendations for how to employ active learning in new domains. As an example, we report on the use of active learning on a large, novel data set of Mars surface images.

Keywords: Active learning, Efficiency, Class discovery, Mars surface image data set

1. Introduction and Related Work

Active learning strategies aim to reduce the labeling burden required for a machine learning classifier to perform well on a new task. Typical heuristics for example selection prioritize examples about which the classifier is most uncertain or which may be most informative by causing the largest changes in the decision boundary (Settles, 2010). Active learning is a tool best used when we encounter an entirely new domain, with no pre-existing labeled data. Yet most active learning studies report instead on retroactive studies where all of the labels already exist, as also noted by Lowell et al. (2019). In so doing, many convenient yet unrealistic assumptions are made, such as assuming that the number and identity of all classes are known, that a number of examples for each class are already labeled, and that the optimal hyperparameters for the base classifier are given. These assumptions may prevent the lessons learned from applying in new, real-life settings.

We focus on two important barriers to the use of active learning in new domains: (1) lack of knowledge or examples of classes when learning begins and (2) high computational cost for example selection. We recommend that these factors influence how active learning methods are designed and evaluated to yield lessons of greater generalization potential.

First, most active learning studies assume that the number and identity of the classes of interest are already known and that at least one (or 2, or 10) examples of each class are given. Working in a new domain raises the *cold start* problem, in which some or all of the classes in the data set have zero labeled examples. Clustering the unlabeled data to generate (ideally) a representative sampling with a small number of items to label can identify some,

but is not guaranteed to find all, of the classes present (Zhu et al., 2008). If some classes are rare, the goal of discovering all classes may be at odds with maximizing overall accuracy, suggesting the need for a joint optimization (Hospedales et al., 2013) or a choice by domain experts about which objective is most valuable. Either way, for realism we recommend that active learning evaluations provide only a single initial labeled item, with the expectation that the active learning method should be responsible for class discovery.

Second, most active learning studies focus solely on sample efficiency or minimizing the cost of human labeling effort. Yet in many cases, especially when employing deep learning systems, the cost required to compute the selection heuristic itself is non-negligible. For uncertainty-based sampling, the classifier must be re-trained before it can generate new predictions so the most uncertain ones can be identified. Most active learning studies sidestep this issue by selecting a batch of items at a time rather than re-training the classifier (Guo and Schuurmans, 2007; Settles, 2010). Batch selection has other benefits such as enabling distributed labeling by a team of labelers, but because they were all chosen based on the same classifier state, learning may be slower. Some heuristics incorporate diversity into the process of batch selection (Brinker, 2003) yet the tradeoff between batch size and performance gain is rarely discussed. In this aspect, model-agnostic selection heuristics that do not require re-training the classifier at each step are advantageous, such as the use of curriculum learning (Bengio et al., 2009; Hacohen and Weinshall, 2019) or purely diversity-based prioritization (Wagstaff et al., 2013). Further, the selections can be re-used by other algorithms without risking the generalization issues identified by Tomanek and Morik (2011) and Lowell et al. (2019). We propose that the cost of the selection heuristic itself be incorporated into active learning evaluations as a guide to future experimenters, who will necessarily pay this price.

2. Assessing Active Learning Methods in New Domains

We conducted experiments to investigate both issues (cold start and selection heuristic cost). In all experiments, we start with a single labeled example from one class, and we assume that the total number of classes is not known and must be discovered during the course of active learning and labeling.

Data sets. We conducted experiments with seven benchmark data sets obtained from the UCI Machine Learning Repository (Dua and Graff, 2017) and scikit-learn (Pedregosa et al., 2011) (see Table 1). We selected data sets with different characteristics to explore the effectiveness and cost of active learning. We expected dimensionality to affect computational cost and class balance to affect class discoverability. The `optdigits` and `pendigits` data set have separate training and test sets; our experiments employed cross-validation using their training sets only.

We also conducted experiments in a new domain using Mars rover images (Lu and Wagstaff, 2020). The Mars Science Laboratory (MSL) rover collected these images in its first 2224 sols (days) on Mars, using the Mastcam and MAHLI cameras. We randomly sampled 2900 of the 54,850 images available for this study. Our goal was to expand on classes previously identified in MSL images (mostly rover parts) (Wagstaff et al., 2018) to include classes of scientific interest, such as layered rocks, veins, and sand. While applying

Table 1: Benchmark data sets: `digits` is from scikit-learn and the rest are from the UCI Machine Learning Repository.

Data set	Samples	Features	Classes	Class distribution
`digits`	1797	64	10	balanced
`wine`	178	13	3	balanced
optical digits (`optdigits`)	3823	64	10	balanced
pen digits (`pendigits`)	7494	16	10	balanced
`sonar`	208	60	2	imbalanced
Wisconsin breast cancer, Diagnostic	569	32	2	imbalanced
`page blocks`	5473	10	5	severely imbalanced

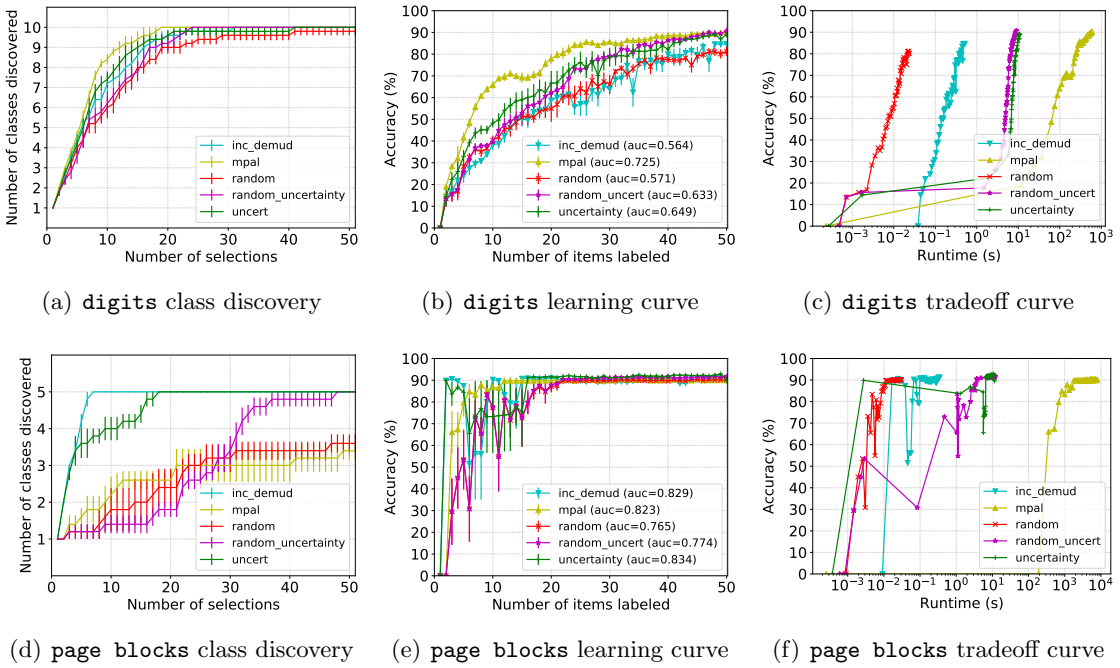
active learning, we discovered and labeled other classes such as night sky, wheel tracks, and artifacts.

Selection heuristics. We compared five item selection heuristics in terms of class discovery and computational cost. The model-agnostic heuristics include random (passive) selection, diversity-based class discovery (DEMUD), and marginal-probability based active learning (MP-AL). DEMUD uses an incremental singular value decomposition of previous selections to iteratively select the most novel next item based on reconstruction error (Wagstaff et al., 2013). DEMUD selects (ranks) all items up front; we adapted it to select one batch at a time for comparison with other heuristics. We set K (number of singular vectors) to preserve 80% of data variance. MP-AL selects a batch of items so as to match the distribution of remaining unlabeled examples (representative) while also minimizing within-batch similarity (redundancy) and similarity to already-labeled items (diversity) (Chattopadhyay et al., 2012). We used an RBF kernel with $\gamma = 1.0$ and solved the quadratic programming problem using CVXOPT (Andersen et al., 2020). The model-dependent heuristics include uncertainty-based selection (Tong and Koller, 2002) and a random-uncertainty hybrid motivated by Mussmann and Liang (2018) that uses random sampling until internal validation accuracy exceeds a threshold value[1] and then switches to more costly uncertainty-based sampling.

Methodology. We assessed active learning with 5-fold cross-validation on the benchmark data sets with three base learners: logistic regression, support vector machine (SVM) (Cortes and Vapnik, 1995), and random forest (Breiman, 2001); implementations are those of scikit-learn. We did not assume that the optimal hyperparameters for the base classifier were known and instead estimated them after each selection using internal cross-validation on the currently labeled data set. The number of folds for this optimization was set to the size of the smallest known class, up to a maximum of 10. If the smallest known class had fewer than 2 items, we instead used the scikit-learn default parameter values.

We assessed active learning on the MSL image data set using a convolution neural network (CNN). We utilized transfer learning to adapt AlexNet (Krizhevsky et al., 2012), whose weights were pretrained using images from ImageNet (Deng et al., 2009), to apply to the MSL image data set. The weights of the network were re-fine-tuned after the addition of each batch of 10 labeled images. To prevent the model from overfitting to the training

1. For the threshold value, we used 70% for the benchmark data sets and 60% for the MSL image data set.



(a) `digits` class discovery (b) `digits` learning curve (c) `digits` tradeoff curve

(d) `page blocks` class discovery (e) `page blocks` learning curve (f) `page blocks` tradeoff curve

Figure 1: SVM results for balanced (`digits`, top row) and imbalanced (`page blocks`, bottom row) benchmark data sets, over 5 folds (standard error indicated with error bars).

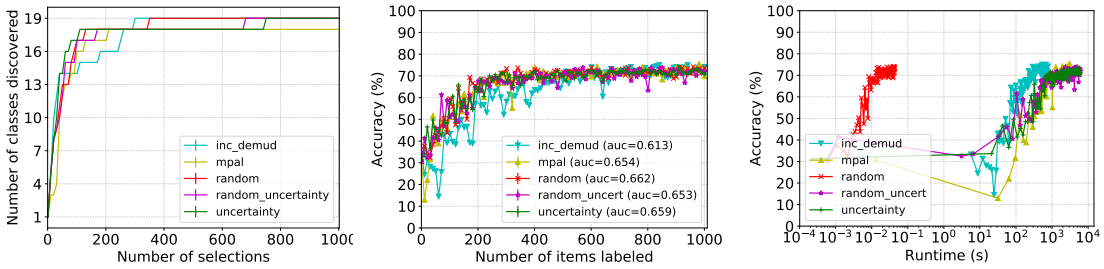
data, we employed an early stopping technique to terminate the training processes if the validation loss did not improve over three consecutive epochs.

Metrics. Active learning performance is often characterized with a learning curve that reports test accuracy as a function of the number of labeled items (Settles, 2010). To assess performance on the issues highlighted in this paper, we also employ class discovery curves (number of classes discovered for a given number of selections) and accuracy as a function of (selection heuristic) runtime. All benchmark experiments were run sequentially on a machine with Intel Xeon CPU that has 32 GB RAM available; the CNN experiment was run on the same machine with NVIDIA Tesla M20 24GB GPU.

3. Results

We first show results for two benchmark data sets, one with balanced classes (`digits`) and one with imbalanced classes (`page blocks`). Due to space constraints, we show results for the SVM base learner only. We used an RBF kernel with γ set to "auto" ($\frac{1}{d\sigma^2}$ where σ is the data standard deviation and d is the dimensionality) and searched data set specific values for C, determined by pairwise distances in feature space, following Chapelle and Zien (2005).

The class discovery curves in Figure 1(a,d) show that MP-AL achieved the fastest class discovery on `digits` while DEMUD was the best performer on the `page blocks` data set. We found that MP-AL also performed the best on the `optdigits`, `pendigits`, and `wine` data sets, all of which are approximately balanced. The combined emphasis on representativity



(a) Class discovery (b) Learning curve (c) Tradeoff curve

Figure 2: CNN active learning results for the imbalanced MSL image data set.

and diversity enabled MP-AL to identify items that span the data set yet are different from previously labeled items. However, MP-AL performed about the same as random selection on the severely imbalanced `page blocks` data set. DEMUD's emphasis on diversity with respect to the previously selected items allowed it to find rare classes quickly.

The learning curves for `digits` and `page blocks` data sets are shown in Figure 1(b,e). For the balanced `digits` data set, MP-AL significantly outperformed other heuristics in early selections; the uncertainty heuristic outperformed the random-uncertainty hybrid heuristic during its random selection phase; and DEMUD performed about the same as the random heuristic. For the severely imbalanced `page blocks` data set, the DEMUD and uncertainty heuristics outperformed the other heuristics for the first 5 selections, with mixed performance after that until selection 23, at which point all heuristics performed about the same. The performance tradeoff curves in Figure 1(c,f) show accuracy scores as a function of runtime. Random selection was, of course, the fastest heuristic, followed by DEMUD. We found that the MP-AL heuristic was by far the most expensive heuristic (note log x axis). Although it achieved higher accuracy for a given number of selections, other methods were able to achieve the same overall accuracy at a tiny fraction of the MP-AL computational cost. In real-world applications, one must assess the relative costs of item labeling versus the cost of selecting the next item to be labeled. For `digits`, MP-AL required ∼10 seconds to select each item, while for `page blocks` it required ∼2 minutes per item (cost scales with the data set size). We also observed the same pattern in the `breast cancer`, `optdigits`, and `pendigits` data sets. Can you wait that long to receive the next item to be labeled?

We also assessed active learning in a real setting using the MSL data set containing images from the surface of Mars. The total number and nature of the image classes was not known in advance. We used DEMUD to rank the images for labeling and identified 19 severely imbalanced classes. Given the now-labeled data set, we retroactively assessed the other methods. Figure 2(a) shows that all methods identify several of the classes quickly, but DEMUD found all 19 with the fewest selections. MP-AL did not find the 19th class within 1000 selections. Figure 2(b,c) show the learning curve and performance tradeoff results for DEMUD, MP-AL, random, and uncertainty heuristics. Strikingly, we did not see a benefit in overall accuracy from using active learning, a phenomenon that has been noted elsewhere (Settles, 2010; Lowell et al., 2019). Mussmann and Liang (2018) found that active learning benefits correlated with inverse error (i.e., how easy or separable or noise-

free a data set was), and the inability of active learning to improve over random sampling, along with the relatively low asymptotic performance, suggests that the concepts in data set are quite challenging to learn. The runtimes of MP-AL and DEMUD include the time to extract image feature vectors using AlexNet's "fc6" layer (Wagstaff and Lee, 2018).

4. Conclusions and Next Steps

Our goal in this work is to identify and explore issues associated with the use of active learning in new domains. Large, unlabeled data sets present exactly the setting for which active learning was designed, yet to be of true utility, the selection heuristics must operate starting with very few labeled examples (perhaps only one) and without knowledge of the total number and nature of the classes or knowledge about the optimal hyperparameters for the base classifier. In this study we have placed more responsibility on the active learning heuristic to bootstrap itself entirely by starting with a single labeled example.

We found that selection heuristics that emphasize diversity are, unsurprisingly, the ones that can most quickly discover all classes that are present. MP-AL achieved the fastest class discovery for balanced data sets, while DEMUD was the best performer on imbalanced data sets. Fast class discovery did not always lead to the most sample-efficient learning in terms of overall accuracy, because rare classes have less impact on total accuracy. However, complete knowledge of the classes in a domain increases understanding of the domain and can guide the next analysis steps. Determining when all classes have been discovered remains an open question.

In terms of accuracy, we found that model-agnostic selection heuristics such as random, DEMUD, and MP-AL were often competitive with model-sensitive heuristics based on uncertainty. Because they are model-agnostic, they can be computed once and employed for multiple different base classifiers rather than requiring that the selection heuristic re-train the classifier each time new examples are chosen. MP-AL was the most costly heuristic to compute by far and may be infeasible for large data sets or those with high dimensionality. Of the heuristics assessed so far, we found DEMUD to provide the best balance between class discovery and efficient operation. If overall accuracy is more important than finding all classes, then simple random selection achieved this goal with the least computational time (but perhaps more selections) on the benchmark data sets, and it was the strongest performer on the more challenging MSL data set. Others have noted that random selection may be the best choice for challenging or novel domains (Settles, 2010; Lowell et al., 2019).

Our next steps include the investigation of additional selection heuristics and a broader assessment of performance with more data sets to more fully answer this question.

Acknowledgments

We would like to acknowledge support for this work from the NASA Planetary Data System, which also provided the MSL image data set. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Copyright 2020 California Institute of Technology. U.S. Government sponsorship acknowledged.

References

Martin Andersen, Joachim Dahl, and Lieven Vandenberghe. CVXOPT, 2020. URL `https://cvxopt.org/index.html`.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning*, pages 59–66, 2003.

Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, 2005.

Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. In *Proceedings of the Knowledge Discovery and Data Mining Conference*, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector network. *Machine Learning*, 20: 273–297, 1995.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Vision and Pattern Recognition*, 2009.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 593–600, 2007.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Timothy M. Hospedales, Shaogang Gong, and Tao Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):374–386, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, 2012.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 21–30, 2019.

Steven Lu and Kiri L. Wagstaff. MSL Curiosity rover images with science and engineering classes, 2020. URL `https://zenodo.org/record/3892024`.

Stephen Mussmann and Percy Liang. On the relationship between data efficiency and error for uncertainty sampling. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2010.

Katrin Tomanek and Katharina Morik. Inspecting sample reusability for active learning. In *Proceedings of the Active Learning and Experimental Design Workshop in conjunction with AISTATS 2010*, pages 169–181, 2011.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.

Kiri L. Wagstaff and Jake Lee. Interpretable discovery in large image data sets. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, pages 107–113, 2018.

Kiri L. Wagstaff, Nina L. Lanza, David R. Thompson, Thomas G. Dietterich, and Martha S. Gilmore. Guiding scientific discovery with explanations using DEMUD. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence*, pages 905–911, 2013.

Kiri L. Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, Thamme Gowda, and Jordan Padams. Deep Mars: CNN classification of Mars imagery for the PDS Imaging Atlas. In *Proceedings of the Thirtieth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 7867–7872, 2018.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1137–1144, 2008.