Inactive Learning? Difficulties Employing Active Learning in Practice

Josh Attenberg
NYU Polytechnic Institute
Brooklyn, NY, 11201
josh@cis.poly.edu

Foster Provost
NYU Stern School of Business
New York, NY, 10012
fprovost@stern.nyu.edu

ABSTRACT

Despite the tremendous level of adoption of machine learning techniques in real-world settings, and the large volume of research on active learning, active learning techniques have been slow to gain substantial traction in practical applications. This reluctance of adoption is contrary to active learning's promise of reduced model-development costs and increased performance on a model-development budget. This essay presents several important and under-discussed challenges to using active learning well in practice. We hope this paper can serve as a call to arms for researchers in active learning—an encouragement to focus even more attention on how practitioners might actually use active learning.

1. INTRODUCTION

The rich history of predictive modeling has culminated in a diverse set of techniques capable of making accurate predictions on many real-world problems. Many of these techniques demand *training*, whereby a set of instances with ground-truth *labels* (values of a dependent variable) are observed by a model-building process that attempts to capture, at least in part, the relationship between the features of the instances and their labels. The resultant model can be applied to instances for which the label is not known, estimating or predicting the labels. These predictions depend not only on the functional structure of the model itself, but on the particular data with which the model was trained.

In many applications, acquiring a label for a particular instance comes at some cost. For example, one may employ human labor to "manually" examine the instance and record its label. In other applications, costly incentives, interventions or experiments may reveal labels. In such cases, simply labeling all available instances may not be practicable, due to budgetary constraints or simply a strong desire to be cost efficient. The dependence of a model's predictive performance on the selection of data suggests that care should be taken. The importance of selective acquisition is evidenced by the vast number of research papers on *active learning*—using the models learned "so far" in the selection of subsequent data for labeling.

However, while active learning is theoretically appealing, it seems that the techniques have had difficulty gaining traction with practitioners. For example, few papers in the literature report on the use of active learning for "real" applications.[1]

This essay discusses how the settings typically used in active learning research papers don't necessarily represent the settings faced by real-world applications. This can result in the literature not providing sufficient guidance for the practitioner actually to apply active learning techniques. The purpose of these observations is to serve as a call to arms to active learning research—a motivation to focus on developing active learning techniques that can be applied effectively.

2. WHAT ACTIVE LEARNING TECHNIQUE SHOULD I USE?

Consider a typical use case for which active learning may seem appealing: given a particular classification problem with some reasonable loss structure (e.g. is a document relevant to a topic or not?), a pool of unlabeled instances, a labeling workforce or procedure that incurs some cost, and a budget that restricts the number of instances to be labeled to a (small) subset of the available pool, construct a predictive model with the best possible performance for the budget—or at least one that is accurate enough to be useful for the practical problem. While this situation may seem extremely simple and an obvious application for active learning, there are several complexities that may stymie the practitioner's application of active learning.

The first and most obvious difficulty is the selection of the active learning technique itself. This is a non-trivial task: there are hundreds of published papers espousing different techniques for active learning, with neither a clear "winner" among them, nor an agreed-upon set of rules of thumb for when to use which technique. The quality of the resultant model can rest on the choice: poor selection may yield a model that performs far worse than would be achieved if one were to select instances for labeling at random. We can see examples of this in Figure 1 (discussed below) and in [9; 26]. The typical post-hoc analysis seen in the active learning research literature, comparing learning curves generated by different techniques on a given problem, simply does not apply in this situation. The practitioner does not have the data required.

[1]The notion of "real" applications of machine learning and data mining technology can be a touchy subject with researchers. Here we simply mean applications with true commercial or scientific import, where the labeling actually is done via active learning. This is in contrast to studies with "real" data, where researchers apply and compare different active learning strategies.

Similar choices faced in machine learning are solved by performing cross-validation. For instance, one may use cross-validation to decide the optimal value of a hyper-parameter, or to choose the best performing model class for a given data set. This is typically done by building and evaluating candidate systems on a data set and picking the setting offering the best performance. However, cross-validation is not applicable to selecting an active learning technique. Before choosing how to sample the data to label, there is no labeled data to perform cross-validation.

One possible solution is to use some "generally safe" active learning strategy initially, then using the acquired data for performing a subsequent cross-validation experiment comparing the induced learning curves to determine the best strategy to use from then on. Such an experimental setup is inadequate. The data sample is biased by the preferences of the initial active learner, and results would be unreliable. For example, derived estimates of generalization performance could be arbitrarily inaccurate. Furthermore, it is unclear how much of the budget would need to be expended in order to choose the active learning technique (the very purpose of which is to optimize budget allocation). While some hybrid active learning heuristics have been proposed, potentially combining the benefits of their constituent sub-methods [11; 5; 17], these techniques suffer from the same failings as their component methods: if the component methods do not work well, or violate the assumptions of the hybrid technique on a particular problem, then the greater techniques will fail to perform as promised. Furthermore, these techniques rely on performance estimations, which if based on the sample drawn for training via the active learner itself, evoke the same problems discussed above.

Alternately, one may simply perform random sampling to gather an initial, experimental data set free from the biases of a particular active acquisition strategy. While this would help achieve more reliable learning curves, such a strategy would defeat the intent of performing active learning in the first place by wasting valuable budget on this random acquisition. The selected instances are likely to differ substantially from those that would be selected by a more intelligent selection process, particularly in a large pool. Furthermore, again it's unclear how many sample instances would be necessary to produce reliable comparable *active* learning curves, and even reliable initial learning curves may not be an indicator of the eventual performance of the strategies considered. Some strategies have been observed to gather a degree of knowledge quickly in certain settings, only to taper off without offering exceptional performance for many subsequent selections, other techniques have been seen to excel at refining already "smart" classifiers, therefore being preferable in the latter stages of active learning [11].

While recent work has examined data acquisition strategies purely for the assessment of model performance [25], it remains unclear how to distribute a limited budget between a system intended for self-evaluation, and a system intended for model induction; improved data gathering for model evaluation would only serve to reduce the expenses incurred by using random sampling for evaluation as described above. Additionally, techniques have been proposed for performing unsupervised assessment [24; 17]. While these surrogate metrics are convenient and may be useful in certain contexts, the reliance on approximations derived from the currently model may be unreliable. Subsequent decisions based on, for instance, minimizing the pool entropy, may simply strengthen the biases already held by the trained model [2]. Furthermore, these metrics may differ substantially from the actual loss describing the problem.

To our knowledge, the only safe way for a practitioner to proceed is to use the literature to select an active learning strategy that is reasonably stable—i.e., one that performs reasonably well on a wide variety of tasks. For example, uncertainty sampling (selecting for labeling those instances from the available pool with the least certain predictions [15]) is by far the most widely studied active learning technique. Uncertainty sampling is the typical baseline for studies of more elaborate active learning strategies, and with good implementations is equivalent to selecting the instances closest to the separating hyperplane of a linear classifier like a support-vector machine [31] and to practical implementations of query-by-committee [28]. However, using uncertainty sampling leaves the practitioner feeling inadequate; as with other techniques used widely as baselines for research papers, uncertainty sampling also is the technique most widely shown to be *worse* than other strategies![2]

3. WHAT BASE LEARNER SHOULD I USE?

A second obvious question also has a subtle dimension. For most not-yet well-understood predictive modeling problems, practitioners face the question of what base learner should be used. Similarly to the case for selecting an active learning strategy (just discussed), we do not have a set of training data on which to inform the choice of a base learner (e.g., via cross-validation). Settles [27] suggests that in settings where the ideal base learner is unknown, a practitioner may be advised to play it safe and prefer random sampling to an active learning strategy that may result in a poor model:

> This ... brings up a very important issue for active learning in practice. If the best model class and feature set happen to be known in advance– or if these are not likely to change much in the future–then active learning can probably be safely used. Otherwise, random sampling (at least for pilot studies, until the task can be better understood) may be more advisable than taking ones chances on active learning with an inappropriate learning model. [27]

However, when considering active learning in practice, under what conditions *would* one know the best model class and feature set in advance? For most practical applications, that would assume that you already have a large, representative set of labeled training data! Expending some of the budget to randomly (or actively) sample a small data set to choose the base learner is not a satisfactory answer. Perlich et al. [22] show quite clearly that given two popular classifier inducers, choosing the learner that performs well for a small data subset often will lead to the wrong choice for a large data set: learning curves often cross. However, to our knowledge there is no good guidance besides experimentation *with labeled data* to know what exactly "large" means in this context for a particular application. There exists a body of work exploring halting heuristics for the active

[2]Sometimes including random sampling.

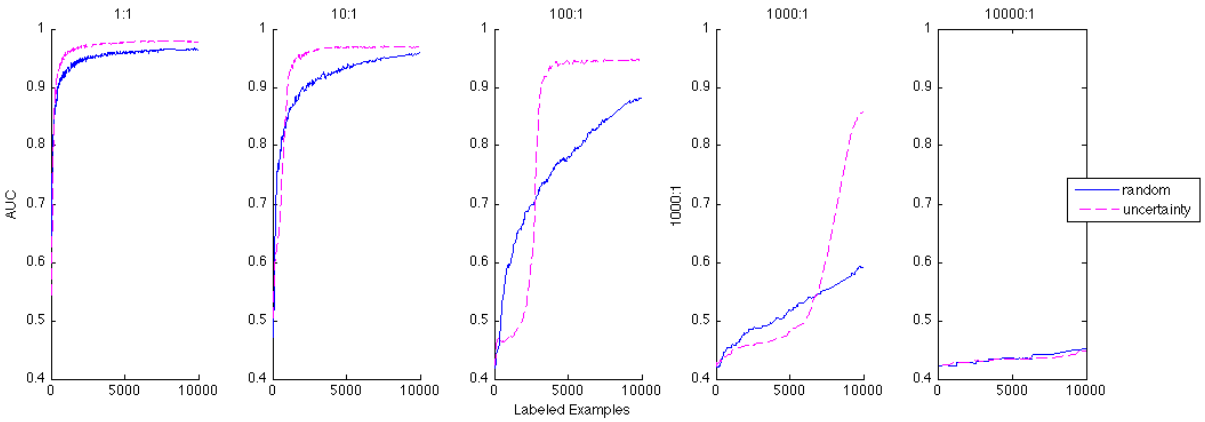

Figure 1: Comparison of random sampling and uncertainty sampling on the same data set with induced skews ranging from $1:1$ to $10,000:1$.

learning process, for instance when further data acquisition is no longer beneficial [38; 32; 14; 12; 37; 6; 17]. However, this work all assumes that one decides on a particular model (functional) form and learning algorithm prior to performing active learning. Convergence of this particular technique may have resulted from a poor initial choice and may not approach what a better model choice could achieve.

The lack of data for deciding on a base model seems to have particular import for model-specific active learning techniques, for example those designed specifically for support-vector machines [31; 35; 12]. Since any given modeling procedure is better on some domains and worse on others,[3] under what conditions would one of these model-specific active learning techniques be justified?

Moreover, model-specific techniques aside, there is the issue of the interaction between the choice of active learning strategy and the base learner used. If the data to label have been chosen based on one active learning strategy using one base learner, are those good data for use with a different base learner (and possibly a different active learning strategy)? Given that some learners are indeed much better for small amounts of data, and others for larger amounts of data, should active learning strategies be designed appropriately? The literature here has expressed mixed results, with several papers expressing positive results in this "reuse" setting: transferring a dataset selected my one class of base learners towards the induction of another model class [16; 30]. Other work has observed difficulties transferring actively selected data sets amongst heterogeneous base learners [18; 4; 20]. One proposed solution is to perform active learning using ensembles consisting of diverse classes of base learners, potentially alleviating the bias towards a particular type of model [18; 4].

Returning to model-specific strategies, it may be the case that the best choice for active learning is a base learner that would be suboptimal if one were to have all the data, possibly because the combination of this learner and its model-specific active learning technique are better than a generic active learning strategy combined with a would-be better base learner.

Thus, Settles' advice amounts to: don't use active learning on real problems where you do not already have a large amount of labeled training data! Exceptions may include fine-tuning a model that's already known to perform well (possibly due to some previous learning or knowledge engineering), and systems where the model form is fixed for other, domain-specific reasons (e.g., a credit-scoring application may demand a logistic regression model).

The practice of active learning could benefit from a different sort of research than "my active learning algorithm is better than yours, *assuming that I'm using learner L*." Instead (or in addition), it is important to study robust active learning techniques: techniques that (i) have good worst-case performance across learners and domains—as compared to using random sampling with a good learner for that domain—and (ii) that also often perform significantly better than random sampling. An alternative (and possibly more ambitious) goal would be (partially) unsupervised methods for estimating the learning technique and active learning technique that in concert would perform well on a given domain, for instance, using a grand expected-utility formulation over the space of learner/active-learner combinations.

4. WHAT SHOULD I DO WHEN MY DATA DISTRIBUTION IS "SKEWED"?

Practical applications rarely provide us with data that have equal numbers of training instances of all the classes. In many applications, the imbalance in the distribution of naturally occurring instances is extreme. For example, when labeling web pages to identify specific content of interest, uninteresting pages may outnumber interesting ones by a million to one or worse (consider identifying web pages containing hate speech, in order to keep advertisers off them, cf. [3]).

Unfortunately, when the data distribution is skewed, active learning strategies can fail completely—and the failure is not simply due to the challenges of learning models with skewed class distributions, which has received a good bit of study [33]. The lack of labeled data compounds the problem, because techniques cannot concentrate on the minority instances, as the techniques are unaware which instances to focus on.

Figure 1 compares the area under the ROC curve (AUC) of logistic regression text classifiers induced by labeled instances selected with uncertainty sampling and with random sampling. The learning task is to differentiate sports web pages from non-sports pages. Depending on the source of

[3]For example, in one well-cited comprehensive experimental comparison [8], support-vector machines were not the best model to choose for *any* data set.

the data (e.g., different impression streams from different on-line advertisers), one could see very different degrees of class skew in the population of relevant web pages. The panels in Figure 1, left-to-right, depict increasing amounts of induced class skew. On the far left, we see that for a balanced class distribution, uncertainty sampling is indeed better than random sampling. For a 10:1 distribution, uncertainty sampling has some problems very early on, but soon does better than random sampling—even more so than in the balanced case. However, as the skew begins to get large, not only does random sampling start to fail (it finds fewer and fewer minority instances, and its learning suffers), uncertainty sampling does substantially worse than random for a considerable amount labeling expenditure. In the most extreme case shown,[4] both random sampling and uncertainty sampling simply fail completely. Random sampling effectively does not select any positive examples, and neither does uncertainty sampling.[5]

A practitioner well-versed in the active learning literature may decide she should use a method other than uncertainty sampling in such a highly skewed domain. A variety of techniques have been proposed for performing active learning specifically under class imbalance [29; 7; 36; 12; 13], as well as for performing density-sensitive active learning, where the geometry of the problem space is specifically included when making selections [38; 10; 21; 35; 19]. While initially appealing, these techniques may not provide results better than more traditional active learning techniques—indeed class skews may be sufficiently high as to thwart these techniques completely [3].

Attenberg and Provost [3] proposed an alternative way of using human resources to produce labeled training set, specifically tasking people with finding class-specific instances ("guided learning") as opposed to labeling specific instances. In some domains, finding such instances may even be cheaper than labeling (per instance). Guided learning can be much more effective per instance acquired; in one of Attenberg and Provost's experiments it outperformed active learning as long as searching for class-specific instances was less than eight times more expensive (per instance) than labeling selected instances. The generalization performance of guided learning is shown in Figure 3, discussed below, for the same setting as Figure 1.

5. DEALING WITH DISJUNCTIVE CLASSES

Even more subtly still, certain problem spaces may not have such an extreme class skew, but may still be particularly difficult because they possess important but very small disjunctive sub-concepts, rather than simple continuously dense regions of minority and majority instances. Prior research has shown that such "small disjuncts" can comprise a large portion of a target class in some domains [34]. For active learning, these small subconcepts act like rare classes: if a learner has seen no instances of the subconcept, how can it "know" which instances to label? Note that this is not simply a problem of using the wrong loss function: in an active learning setting, the learner does not even know that the instances of the subconcept are misclassified if no instances of a subconcept have yet been labeled. Nonetheless, in a research setting (where we know all the labels) using an undiscriminative loss function, such as classification accuracy or even the area under the ROC curve (AUC), may result in the researcher not even realizing that an important subconcept has been missed.

To demonstrate how small disjuncts influence (active) model learning, consider the following text classification problem: separating the *science* articles from the *non-science* articles within a subset of the 20 Newsgroups benchmark set (with an induced class skew of 80 to 1). Figure 2 examines graphically the relative positions of the minority instances through the active learning. The black curve shows the AUC (right vertical axis) of the models learned by a logistic regression classifier using uncertainty sampling, rescaled as follows. At each epoch we sort all instances by their predicted probability of membership in the majority class, $\hat{P}(y = 0|x)$. The blue dots in Figure 2 represent the minority class instances, with the value on the left vertical axis showing their relative position in this sorted list. The x-axis shows the active learning epoch (here each epoch requests 30 new instances from the pool). The blue trajectories mostly show instances' relative positions changing. Minority instances drop down to the very bottom (certain minority) either because they get chosen for labeling, or because labeling some other instance caused the model to "realize" that they are minority instances.

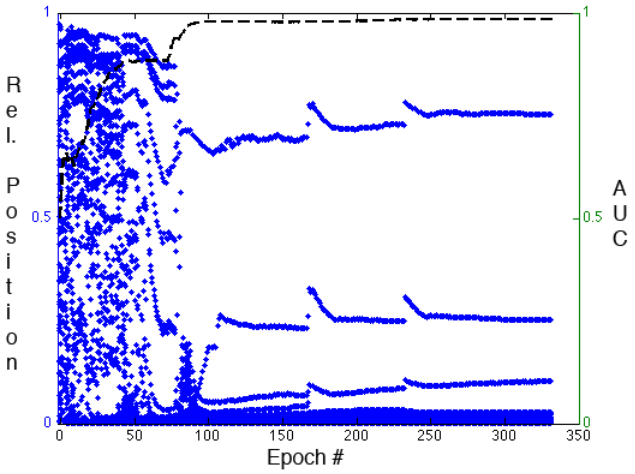

Figure 2: A comparison of the learned model's ordering of the instance pool, along with the quality of the cross-validated AUC.

We see that, early on, the minority instances are mixed all throughout the range of estimated probabilities, even as the generalization accuracy increases. Then the model becomes good enough that, abruptly, few minority class instances are misclassified (above $\hat{P} = 0.5$). This is the point where the learning curve levels off for the first time. However, notice that there still are some residual misclassified minority instances, and in particular that there is a cluster of them for which the model is relatively certain they are *majority* instances. It takes many epochs for the active learning to select one of these, at which point the generalization performance increases markedly—apparently, this was a subconcept that was strongly misclassified by the model, and so it was not a high priority for exploration by the active

[4] 10,000:1 — still orders of magnitude less skewed than some categories

[5] The curious behavior of AUC< 0.5 here is due to overfitting. Regularizing the logistic regression "fixes" the problem, and the curve hovers about 0.5. See another article in this issue for more insight on models exhibiting AUC< 0.5 [23].

learning.

On the 20 Newsgroups data set we can examine the minority instances for which $\hat{P}$ decreases the most in that late rise in the AUC curve (roughly, they *switch* from being misclassified on the lower plateau to being correctly classified afterward). Recall that the minority (positive) class here is "Science" newsgroups. It turns out that these late-switching instances are members of the cryptography (sci.crpyt) subcategory. These pages were classified as non-Science presumably because before having seen any positive instances of the subcategory, they looked much more like the many computer-oriented subcategories in the (much more prevalent) non-Science category. As soon as a few were labeled as Science, the model generalized its notion of Science to include this subcategory (apparently pretty well).

Density-sensitive active learning techniques did not improve upon uncertainty sampling for this particular domain. This was surprising, given the support we have just provided for our intuition that the concepts are disjunctive. One would expect a density-oriented technique to be appropriate for this domain. Unfortunately in this domain—and we conjecture that this is typical of many domains with extreme class imbalance—the *majority* class is *even more disjunctive* than the minority class. For example, in 20 Newsgroups, Science indeed has four very different subclasses. However, non-Science has 16 (with much more variety). Techniques that (for example) try to find as-of-yet unexplored clusters in the instance space are likely to select from the vast and varied majority class. We need more research on dealing with highly disjunctive classes, especially when the less interesting[6] class is more varied than the main class of interest.

6. STARTING COLD

The *cold start problem* has long been known to be a key difficulty in building effective classifiers quickly and cheaply via active learning [38; 11]. Since the quality of data selection directly depends on the understanding of the space provided by the "current" model, early stages of acquisitions can result in a vicious cycle of uninformative selections, leading to poor quality models and therefore additional poor selections. The difficulties posed by the cold start problem can be particularly acute in highly skewed or disjunctive problem spaces; informative instances may be difficult for active learning to find due to their variety or rarity, potentially leading to substantial waste in data selection. Difficulties early in the active learning process can, at least in part, be attributed to the base classifier's poor understanding of the problem space. This cold start problem is particularly acute in otherwise difficult domains. Since the value of subsequent label selections depends on base learner's understanding of the problem space, poor selections in the early phases of active learning propagate their harm across the learning curve.

In many research papers active learning experiments are "primed" with a preselected, often class-balanced training set. As pointed out by [3] if the possibility and procedure exists to procure a class-balanced training set to start the process, maybe the most cost-effective model-development alternative is not to do active learning at all, but to just continue using this procedure. This is exemplified in Figure 3 [3], where the red lines show the effect of investing resources to continue to procure a class-balanced, but otherwise random, training set (as compared with the active acquisition shown in Figure 1).

[6]How interesting a class is could be measured by its relative misclassification cost, for example.

7. CONCLUSIONS

Active learning as a field has shown tremendous theoretical potential to help us to build predictive models quickly and cheaply. However, slow adoption in practice suggests that practitioners face difficulties realizing this potential. This paper illustrates a surprising array of practical difficulties, including:

1. how to choose (cost-effectively) the active learning technique when one starts without the labeled data needed for methods like cross-validation;
2. how to choose (cost-effectively) the base learning technique when one starts without the labeled data needed for methods like cross-validation, given that we know that learning curves cross, and given possible interactions between active learning technique and base learner;
3. how to deal with highly skewed class distributions, where active learning strategies find few (or no) instances of rare classes;
4. how to deal with concepts including very small subconcepts ("disjuncts")—which are hard enough to find with random sampling (because of their rarity), but active learning strategies can actually *avoid* finding them if they are misclassified strongly to begin with;
5. how best to address the cold-start problem, and especially
6. whether and what alternatives exist for using human resources to improve learning, that may be more cost efficient than using humans simply for labeling selected cases, such as guided learning [3], active dual supervision [2], guided feature labeling [1], etc.

We do not intend this essay to be an indictment of active learning research, a field responsible for substantial strides in understanding the problem of cost-effectively acquiring labeled training data. Rather, we hope that it can serve as a call to arms to the research community. We cannot take the current volume of published papers on active learning as a sign that the problem is "solved." As practitioners, we need more research focused on these fundamental questions. It would benefit both the research and practitioner communities if active learning researchers were to view the practical application of active learning techniques as a motivating framework within which to select the important research questions on which to work.

8. REFERENCES

[1] J. Attenberg, P. Melville, and F. Provost. Guided feature labeling for budget-sensitive learning under extreme class imbalance. In *BL-ICML '10: Workshop on Budgeted Learning*, 2010.

[2] J. Attenberg, P. Melville, and F. J. Provost. A unified approach to active dual supervision for labeling features and examples. In *ECML/PKDD*, 2010.

[3] J. Attenberg and F. Provost. Why label when you can search? strategies for applying human resources to build classification models under extreme class imbalance. In *KDD*, 2010.

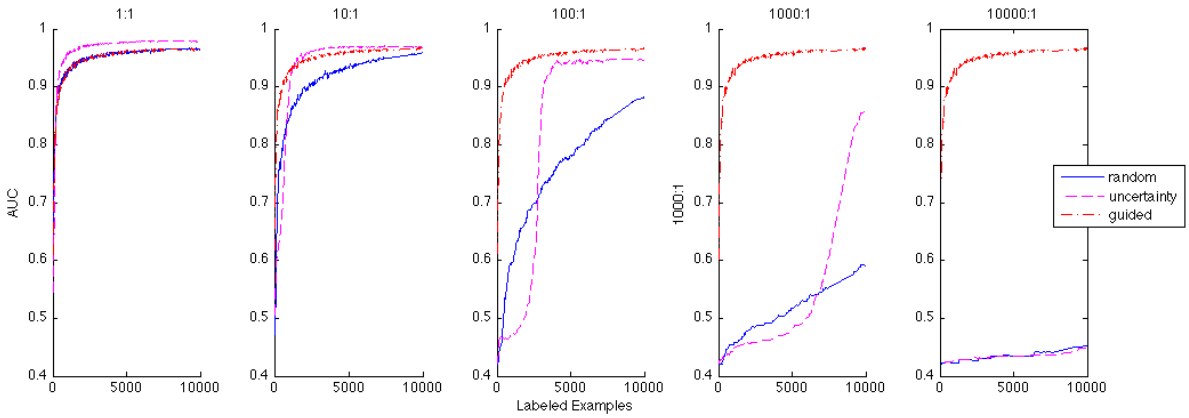

Figure 3: Comparison of random sampling and uncertainty sampling and guided learning on the problem seen in Figure 1.

[4] Baldridge, Jason and Osborne, Miles. Active learning and the total cost of annotation. In *EMNLP*, 2004.

[5] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5:255–291, 2004.

[6] M. Bloodgood and K. V. Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *CoNLL*, 2009.

[7] M. Bloodgood and K. V. Shanker. Taking into account the differences between actively and passively acquired data: the case of active learning with support vector machines for imbalanced datasets. In *NAACL*, 2009.

[8] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *ICML*, pages 161–168, 2006.

[9] Dasgupta, Sanjoy. Analysis of a greedy active learning strategy. In *NIPS*, pages 337–344, 2004.

[10] P. Donmez and J. Carbonell. Paired Sampling in Density-Sensitive Active Learning. In *Proc. 10 th International Symposium on Artificial Intelligence and Mathematics*, 2008.

[11] P. Donmez, J. G. Carbonell, and P. N. Bennett. Dual strategy active learning. In *ECML '07*, 2007.

[12] S. Ertekin, J. Huang, L. Bottou, and L. Giles. Learning on the border: active learning in imbalanced data classification. In *CIKM*, 2007.

[13] J. He and J. G. Carbonell. Nearest-neighbor-based active learning for rare category detection. In *NIPS*, 2007.

[14] F. Laws and H. Schütze. Stopping criteria for active learning of named entity recognition. In *COLING*, Morristown, NJ, USA, 2008.

[15] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR '94*, 1994.

[16] Lewis, David D. and Catlett, Jason. Heterogeneous Uncertainty Sampling for Supervised Learning. In *ICML*, pages 148–156, 1994.

[17] R. Lomasky. *Active Acquisition of Informative Training Data.* PhD thesis, Tufts University, 2010.

[18] Lu, Zhenyu and Bongard, Josh. Exploiting Multiple Classifier Types with Active Learning. In *Proceedings of the Conference on Genetic and Evolutionary Computation*, 2009.

[19] A. K. Mccallum and K. Nigam. Employing em in pool-based active learning for text classification. In *ICML*, 1998.

[20] Melville, Prem and Mooney, Raymond J. Diverse Ensembles for Active Learning. In *ICML*, 2004.

[21] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML*, 2004.

[22] C. Perlich, F. J. Provost, and J. S. Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255, 2003.

[23] C. Perlich and G. Swirszcz. On cross-validation and stacking: Building seemingly predictive models on random data. *SIGKDD Explorations*, 12(2):This issue, 2010.

[24] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[25] Sawade, Christoph, Bickel, Steffen, and Scheffer, Tobias. Active Risk Estimation. In *ICML*, 2010.

[26] A. Schein and L. Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265–265, October 2007.

[27] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[28] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT '92*, 1992.

[29] K. Tomanek and U. Hahn. Reducing class imbalance during active learning for named entity annotation. In *K-CAP '09: Intl. Conf. on Knowledge capture*, 2009.

[30] Tomanek, Katrin, Wermter, Joachim, and Hahn, Udo. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *In Proc. of EMNLP/CoNLL07*, volume 3, pages 486–495, 2007.

[31] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2002.

[32] A. Vlachos. A stopping criterion for active learning. In *Computer Speech and Language*, volume 22(3), pages 295–312, 2007.

[33] G. M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[34] Weiss, Gary M. The Impact of Small Disjuncts on Classifier Learning. *Annals of Information Systems*, 8:193–226, 2010.

[35] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. Representative sampling for text classification using support vector machines. In *ECIR*, 2003.

[36] J. Zhu and E. Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *EMNLP-CoNLL '07*.

[37] J. Zhu, H. Wang, and E. Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *COLING*, 2008.

[38] J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING '08*, 2008.