Why Label when you can Search? Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance

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

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

ABSTRACT

This paper analyses alternative techniques for deploying low-cost human resources for data acquisition for classifier induction in domains exhibiting extreme class imbalance—where traditional labeling strategies, such as active learning, can be ineffective. Consider the problem of building classifiers to help brands control the content adjacent to their on-line advertisements. Although frequent enough to worry advertisers, objectionable categories are rare in the distribution of impressions encountered by most on-line advertisers—so rare that traditional sampling techniques do not find enough positive examples to train effective models. An alternative way to deploy human resources for training-data acquisition is to have them "guide" the learning by searching explicitly for training examples of each class. We show that under extreme skew, even basic techniques for guided learning completely dominate smart (active) strategies for applying human resources to select cases for labeling. Therefore, it is critical to consider the relative cost of search versus labeling, and we demonstrate the tradeoffs for different relative costs. We show that in cost/skew settings where the choice between search and active labeling is equivocal, a hybrid strategy can combine the benefits.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—data mining; I.2.6 [**Artificial Intelligence**]: Learning—induction; I.5.1 [**Pattern Recognition**]: Models—statistics

General Terms: Design, Performance, Human Factors

Keywords: active learning, machine learning, class imbalance, human resources, on-line advertising, micro-outsourcing

This work was conducted while the authors were at AdSafe Media. Foster Provost thanks NEC for a Faculty Fellowship.

KDD'10, July 25–28, 2010, Washington, DC, USA.
Copyright 2010 ACM 978-1-4503-0055-1/10/07 ...$10.00.

1. INTRODUCTION

This paper concerns the interaction of humans in the data acquisition phase of the process of building classification models from data. Consider the following example data mining application: classifying web pages for the purpose of *safe advertising*. Advertisers and advertising networks (hereafter, advertisers) would like a rating system that estimates whether a web page or web site displays certain objectionable content. With such a system, advertisers can control the destination of their ads, advertising only on those pages deemed unlikely to display such unacceptable content (depending on the advertiser, objectionable categories include: adult content, kids content, hate speech, malware, etc.).[1] Evaluating each potential advertising opportunity involves classifying the web page with respect to these objectionable categories. The classification system can take into account various evidence, including the URL, the page text, anchor text, DMOZ categories, third-party classifications, position in the network of pages, and so on [18, 1, 3]. For this paper, we will consider only the textual html source for each page, but the ideas generalize to any type of available feature data.

Manually examining every page encountered by such a system would be prohibitively expensive. This is particularly true in safe advertising, where models for new classification categories must be built rapidly to meet the changing demands of each customer and campaign. Furthermore, assuming that these classifications are based on statistical models, predictions will be more or less effective depending on the particular cases used for training and on the amount and distribution of training data used in their construction. For a given budget, some subset of the cases can be examined by humans—potentially at very low cost using a micro-outsourcing system [24] such as Amazon's Mechanical Turk [2]—to produce training data.

A critical question then is: *which cases should be selected for training?* Simply sampling cases uniformly at random is unlikely to be the best strategy, as is evidenced by the rich field of research comprising active learning [22]. Settings with extreme class imbalance, as is frequently the case on the web, further reduce the effectiveness of random sampling since for reasonable labeling budgets, only rarely would such sampling produce a positive training example at all. For instance, we would hope that the distribution of pages that

[1] This site rating system may be best developed and maintained by a third party to avoid conflicts of interest [4], but that complexity does not affect the development here.

are offered an advertiser for ad placement contain only a tiny fraction of pages containing hate speech. For safe advertising, depending on the category, the base rate of the minority class can be $\frac{1}{10^4}$ to $\frac{1}{10^7}$ or lower. Occasionally filters can be provided on the data (for example, based on selected phrases), to reduce the skew by orders of magnitude. However, generally we see base rates of less than $\frac{1}{10^2}$. These filters introduce a new problem: they bias the data—training and test—toward particular "disjuncts" [29] of the objectionable category.

In extreme cases, active learning simply finds no minority-class examples—examples of the positive class (e.g., adult content, hate speech) appear too infrequently in the pool of cases considered for labeling. Even in moderately high skew settings, any strategy for selecting examples automatically that does not solve the classification problem itself, is likely to select mostly negative examples. For generality this paper mainly considers moderate skews; the trends to the far extreme are clear. As demonstrated below in Section 5, as the minority class becomes more and more scarce, standard active learning strategies have increased difficulty finding instances that improve performance on held-out data.

The techniques we use in this paper for the most part either are existing techniques or are fairly straightforward. The main novelty is the idea of considering on equal footing a variety of strategies for applying human resources for data mining, and an analysis of different strategies applying human resources for inducing classifiers in domains where the base rate of an important category is very low. The main contributions of the paper are threefold (as follows).[2]

First, the paper provides an empirical analysis of the performance of traditional active learning techniques in highly skewed settings. This investigation reveals the deficiency of many active learning techniques in the extreme imbalance setting: the dependency of the active learner on an uninformative model leads to repeatedly making the same mistakes, often leading to selecting mainly majority-class instances for labeling, while important portions of the minority class remain completely unrepresented in the training data.

Second, the paper contrasts active learning with "guided learning." Active learning is based on the availability of human resources that can be applied to the *labeling* of specially selected examples. Guided learning applies those resources specifically to *search* for examples satisfying some criteria; for example, humans could be directed to search specifically for positive examples. For safe advertising, adult content or hate speech may have a fairly low prevalence among the pages supported by a particular advertiser (or on the web more generally); however, a human with a search engine may be able to find examples fairly quickly. We examine the relative benefit of using humans for labeling cases (in combination with an active learning strategy) versus using humans for searching for cases.

The results are striking. Looking ahead to Figure 2, it is clear that that straightforward guided learning strongly dominates active learning—improving accuracy much faster as examples are added to the training set. The details of all techniques will be covered below. This is important not only as practical guidance. This result shows that the dominant problem in these domains is simply finding minority-class examples, not finding otherwise "informative" examples or examples near the classification boundary. The paper proceeds with a deeper empirical analysis of this phenomenon. For example, does this result still hold if search for examples is several times more expensive than labeling examples?

Third, the paper presents and evaluates a *hybrid* strategy for cost-effective guided learning, that utilizes both search and active labeling. The results show that a hybrid strategy can perform better than either pure guided learning or pure active learning, when the setting does not provide clear dominance of one over the other. An ultimate goal for this sort of learning would be to judge the relative benefit-per-unit-cost of each sub-strategy, and allocate resources to labeling or to search accordingly.

The techniques described here are operating in production as part of the technology underlying the rating system of AdSafe Media.[3] Human analysts are tasked with labeling and with search for guided learning; they are supported by systems for web-page labeling and for web search. Models are built across various categories of objectionable content, including adult content, hate speech, violence, and others. Guided/active learning procedures feed analysts with search tasks and with examples to be labeled. The resultant models are used to reduce objectionable on-line advertising adjacencies. In practice, we find that mixing guided learning and active learning is preferable to either in isolation.

The remainder of this paper proceeds as follows: Section 2 covers the baseline techniques used for comparison and explains the details and motivation behind guided learning. Section 3 covers prior work on classification and active learning in unbalanced settings. Section 4 presents the experimental framework and datasets used for evaluation. Section 5 covers the results of these experiments. Section 7 covers the behavior of guided learning under different cost settings, and presents and evaluates hybrid guided/active data acquisition strategies. Section 8 compares with the results of an actual production guided learning system. Section 9 provides further discussion of the issues raised by this work, offers concluding remarks, and notes. Earlier versions of this paper appeared previously [6, 7].

2. LEARNING & HUMAN INTERVENTION

As discussed, this paper analyzes two different methods for incorporating human resources in the data mining process. Specifically, via labeling carefully chosen examples, or via searching for examples. We assume that the reader is familiar with the notion of active learning for choosing examples for labeling; Settles provides a comprehensive survey [22]. We call the search for examples based on particular criteria "Guided Learning." Here we describe the particular techniques that we study in this paper.

2.1 Active Learning

For active learning, this study employs two strategies. The first, uncertainty sampling, is by far the most popular active learning strategy, and is closely related to model-specific strategies such as actively selecting instances closest

[2] As a minor contribution we introduce to the research community the application of classification for safe advertising [6, 7], which exhibits extremely skewed class distributions and illustrates the issues addressed here. The interested reader should also see `http://www.adsafemedia.com` and the contemporaneous work of Rajan et al. [21].

[3] `http://www.adsafemedia.com`

to a separating hyperplane [26].[4] The second is a variation of the popular Query-by-committee [23] technique, specifically introduced to deal with skewed class distributions [25].

- **Uncertainty Sampling**: instances with the smallest margin are chosen for inclusion at each fold. Here we calculate margin as $|p(0) - p(1)|$ [15].
- **Boosted Disagreement with QBC**: instances are ordered by a class-weighted disagreement measure, $-\sum_{j\in\{0,1\}} b_j \frac{V(k_j)}{|C|} \log \frac{V(k_j)}{|C|}$, where $V(k_j)$ is the number of votes from a committee of size $|C|$ that an instance belongs to a class k_j. b_j is a weight corresponding to the importance of including a certain class; a larger value of b_j corresponds to a increased tendency to include examples that are thought to belong to this class. From a window W of examples with highest disagreement, instances are selected greedily based on the model's estimated class membership probabilities so that the batch selected from the window has the highest probability of having a balanced class membership [25].

We also assessed several other candidate active learning techniques. None of them offered substantial improvement over the techniques presented above. Of particular note, we ran several experiments with the density-sensitive pre-clustering technique of Nguyen and Smeulders [19], because of the expected ”disjunctiveness” of the classes (see extended discussion below). Surprisingly, this technique performed no better than uncertainty sampling on the datasets under consideration, while the added computational complexity extended the experimental run time to prohibitive levels. We offer some conjectured insight into the deficiencies of density-sensitive techniques in Section 6.

2.2 Guided Learning

Guided Learning is an alternative technique for utilizing human resources for model development, beyond traditional (active) instance labeling. Here, humans are tasked with *seeking* examples satisfying some criteria. For this paper, the basic guided learning task is straightforward: find examples representing the different classes in some proportion, ρ. These instances are provided as input to classifier induction.

Guided learning is motivated by the results of Weiss & Provost [28, 30], who address the question “if only n training examples are to be selected, in what proportion should the classes be represented?” Their results show that the best proportion varies across domains; however, if one wants to maximize the ranking of cases (i.e., the AUC) a proportion of $\rho = 0.5$ is a very good choice. In principle the problem of this paper is different: how to use human resources to *search* for valid examples using all tools available to them—including both active learning and guided learning. Nevertheless, this paper's analysis could be seen as a follow-on to this prior work; in our experimental setting we simulate guided learning by class-conditional random sampling. We describe the simulation below.

More specifically, a thorough evaluation of a guided learning system in the wild would require a sizable labeled pool of instances, in effect defeating the cost savings of the techniques proposed here. In order to compare and contrast different techniques, all guided learning experiments presented here are performed in the following way: given an initial pool of labeled instances P with some subset of minority and majority instances, P_+ and P_- respectively, along with a selection ratio, ρ, at each batch, the guided learning simulator selects $\rho|b|$ instances from the P_+ uniformly at random and $(1 - \rho)|b|$ instances uniformly at random from P_-, where $|b|$ is the size of the batch selected at each selection epoch. This process proceeds until either pool is exhausted, at which point the process switches over to purely random sampling from the other class. This simulation is similar to the procedure of Weiss & Provost who assume that examples can be produced randomly by class.

3. RELATED WORK

As mentioned above, guided learning was motivated by the results of Weiss and Provost [30, 28]; the authors investigate the influence of class distribution on classifier performance, empirically showing that given a training set of n examples, barring domain-specific information, a balanced class distribution tends to offer the best AUC on held-out data.[5] Lomasky et al. [17], as well as Weiss and Provost, also investigate the setting where instances can be drawn randomly by class, and address the issue of actively choosing classes for sampling. This is a complementary task to what we address in this paper. Our analysis assumes simply that examples will be provided in a particular proportion (balanced by class for our experiments). Incorporating techniques for better choosing the class distribution in the training data could improve the guided learning results presented below and is a direction for future work

There is an extensive body of work investigating strategies for learning in highly skewed settings. This work includes over-sampling the minority class or under-sampling the majority class [9, 16]. A different branch of work investigates the application of non-uniform misclassification costs during training in order to give additional consideration to the class of interest [10].

There has been some work on active learning on skewed data. Tomanek [25] investigates Query By Committee-based approaches to sampling labeled sentences for the task of named entity recognition. The goal of their selection strategy is to encourage class-balanced selections by incorporating class-specific costs. This work assumes that classifiers can often accurately infer which instances belong to the minority class, giving higher weight to instances thought to belong to the minority class and with a high degree of uncertainty. Our work differs from this by extending to extreme cases where initial performance is poor. Additionally, our techniques are more general, able to extend beyond the tasks faced in NLP.

Bloodgood and Shanker [8] use a similar approach to [25], incorporating class specific cost factors to encourage choosing from the minority class in the skewed setting. Here the base rate is estimated on a small random sample. We note that in many realistic settings, random samples may not reveal any minority instances, thereby foiling this technique.

Zhu and Hovy [31] investigate active learning in conjunc-

[4] For example, for an SVM that produces probability estimates via the common technique of applying a simple logistic regression to the orthogonal distance of an example from the separating hyperplane, uncertainty sampling will select the unlabeled examples closest to the separator.

[5] Many practitioners used this as a rule of thumb prior to Weiss & Provost's research.

tion with over and under-sampling to alleviate the class imbalance problem. Here active learning is used to choose a set of instances for labeling, with sampling strategies used to improve the class distribution. Our work differs by seeking strategies for acquiring a good class distribution in the data, removing the necessity for performing sub-sampling.

Ertekin et al. [13] address active learning on highly unbalanced data sets. Given a large, unbalanced pool of labeled instances, the authors randomly sub-sample instances, choosing to keep only those that are positioned close to the margin of a SVM classifier. The authors do not address the problem of seeking unlabeled instances in the wild. Furthermore, the margin-based active learning heuristic is very similar to uncertainty sampling, a strategy that we demonstrate to exhibit difficulty in the extremely skewed cases.

We note that many active learning strategies depend to some degree on the quality of the current model—until the model "warms up," the instance selection is essentially random. This cold-start problem has been examined by Zhu et al. [32], work extended by Donmez and Carbonell [11]. This work seeks to find "clusters" of distinct content among the unlabeled instances. While this offers greater potential overcoming the cold-start than many common active learning techniques, it is still unlikely to succeed in the extremely skewed case; there is often so much diversity within the majority that the method will miss any minority instances. Additionally, these complex methods don't scale well to the data sizes necessary to experience an extreme class skew.

Donmez et al. [12] propose a hybrid active learning technique whereby a density-sensitive learning technique is used to overcome the initial deficiencies of uncertainty sampling until the derivative of the learning rate decreases below some threshold. After this point, traditional uncertainty sampling is incorporated to the instance selection. The intuition here is that the density-sensitive technique is better for exploring the space, while uncertainty sampling is better at "fine tuning" the decision boundary.

4. EXPERIMENTAL SETUP

The experiments for this paper are performed on six data sets with similar characteristics; all represent a task of separating examples of one minority class from examples of a diffuse collection of other topics. While all use text as the raw feature data, the techniques illustrated here apply to any other type of input. The first two are taken directly from the domain of safe advertising; the others are publicly available surrogate data sets with similar problem structure. Specifically, the data sets are:

1. **Safe-Adult** A set of 35,000 pages labeled based on the presence of adult content. Positive instances here are deemed unsafe for advertising.
2. **Safe-Guns** A set containing 55,000 pages labeled based on the presence of guns, ammunition, bombs, or other destructive equipment. Often advertisers choose not to be associated with this type of content.

The two previously mentioned datasets represent random sub-samples from much larger datasets for experimental convenience. Safe-Adult has a class skew of approximately 20 : 1 while Safe-Guns has a class skew of roughly 150 : 1.

The next three data sets were taken from urls contained in the topical hierarchical taxonomy of the Open Directory Project [1]. This data set is a result of a crawl of approximately 4,000,000 urls, and instances are assigned class labels based on their membership in top-level DMOZ categories. To eliminate confusion, pages belonging more than one category were eliminated from this experiment. Data sets were further down-sampled in order to induce a greater degree of skew.

3. **DMOZ-Science**: Positive instances belong to the top-level category of Science, while the minority instances belong to all other categories. This set has approximately 130,000 instances. While this data set was used for experiments of varying skew, the nominal class distribution for this set is 200 : 1.
4. **DMOZ-News**: Here positive instances are pages found in the News top-level DMOZ category. This data set has 100,000 instances with a class skew of 100 : 1.
5. **DMOZ-Games**: Urls sampled from the Games category make up the positive category in this data set, sub-sampled to give 100,000 instances with a 100 : 1 class skew.
6. 20-**News-Groups**: This data set is derived from the popular 20 News Groups set frequently used in text mining evaluation [5]. The data set is modified from the original data by assigning a positive label to all science-related articles, and a negative label otherwise. Positive instances are randomly removed from the data set to give a highly skewed label distribution of roughly 80 to 1.

Classification and probability estimation are performed with logistic regression trained using stochastic gradient descent using feature hashing [27]. The choice of logistic regression was based on this algorithm's efficiency during training and induction, critical given the massive numbers of experimental runs performed in this work. Smaller-scale experiments indicate that the main effects described in this work are independent of the type of model used; other machine learning algorithms could be substituted with similar results.

All experiments compare the area under the receiver operating characteristic curve (AUC) at various stages in learning [14]. This metric allows a comparison of model performance that is largely insensitive to the class prior in the evaluation set, and also to the difference in class priors between the training and test sets. This is critical in a highly skewed setting where simply choosing the majority label for each instance would yield very high (and misleading) classification accuracy, and where one often wants to dope the training set with additional minority-class examples. The results presented are averages computed over ten-fold cross-validation for every experiment. We use uniform random sampling as the baseline against which to compare human-intervention strategies.

5. ACQUISITION RESULTS

Figure 1 compares the four acquisition strategies covered in sections 2.2 and 4 on the six high-skew data sets under consideration. These plots show how the the area under the ROC curve (AUC; vertical axis) improves with additional labeled training data, as acquired by the different strategies.

The results show that searching for examples of each class in balanced proportion, *even without any "active" selection at all*, by and large provides substantially more informative data to the modeling process. These results are similar for every experiment: guided learning very quickly achieves very good class separation (AUC in the high .90s) with considerably fewer examples than are required by active strategies or random sampling. In comparison, the "smart" labeling strategy, uncertainty sampling, often offers no benefits over simply selecting instances at random, in every case requiring thousands of examples to achieve the performance levels of a few hundred instances selected through guided learning.

Boosted disagreement—though specifically designed for active learning in skewed settings by favoring uncertain instances belonging to the minority class—seems to suffer from the same problems as uncertainty sampling, with few exceptions. The models forming the committee are simply unaware of much of the minority class; the method has limited ability to find instances that would improve the generalization performance of the system.

On 20-News-Groups we see an interesting behavior: uncertainty sampling and boosted disagreement perform quite well initially, followed by plateaus, as the active learning apparently cannot find the instances that will improve the model beyond a certain point; only after they exhaust a large number of seemingly uninformative examples do they choose examples for labeling that again provide improvement over random sampling. This behavior suggests a disjunctive minority class, with portions of the class lying within the high-certainty (of majority) regions of the example space. Examples from these disjuncts are only selected when active learning exhausts the less-certain instances. As a result, little improvement is offered after repeated example selection. In effect, the certainty of the underlying model that the as-of-yet unexplored disjuncts are of the wrong class hinders the success of the active learning. We return to this in Section 6.

From these results we can conclude that many active learning heuristics are ill-suited for learning highly skewed, possibly disjunctive concepts. In these cases, gathering rare-class examples and exploring sub-concepts is critical. In these settings, the base-learner induced by the early stages of active learning often has a poor understanding of the problem space, making poor selection of subsequent instances as a result, in turn offering little improvement in model performance. Guided learning, as presented here, does not depend on the quality of the base-learner, relying on an oracle to explore the details of the space.

There is an interesting phenomena in some of the learning curves shown in Figure 1. Notice that many learning curves traced by guided learning achieve their optimal AUC very quickly, then slowly decay. This may be due to several factors: the base learner may be susceptible to certain high skew settings, or to noise present on certain features. Alternately, introducing a large number of diverse pages may cause a large number of hash collisions, an artifact of the feature hashing used, thereby reducing the clarity of each dimension. It is also possible that there is sufficient label noise in this case that the good data set provided in the initial epochs is simply washed out as more data is added.

The primary motivation for introducing guided learning was to facilitate cost-effective learning in settings with high-to-extreme class skew. We now assess the relative benefit of guided learning for different skews. Figure 2 presents the learning curves for three different instance sampling strategies at different base rates, as induced on the DMOZ-Science data set. Specifically, this plot displays the AUC for three different labeling strategies, uniform random sampling, uncertainty sampling, and guided learning with $\rho = 0.5$. We see that as the skew increases, uncertainty sampling and random sampling have increased difficulty selecting minority instances, resulting in poor generalization performance. In the most skewed setting, random sampling and uncertainty sampling are unable to select reliably any minority instances given $10,000$ draws from the available pool. This likely results in a complete lack of understanding of the space.

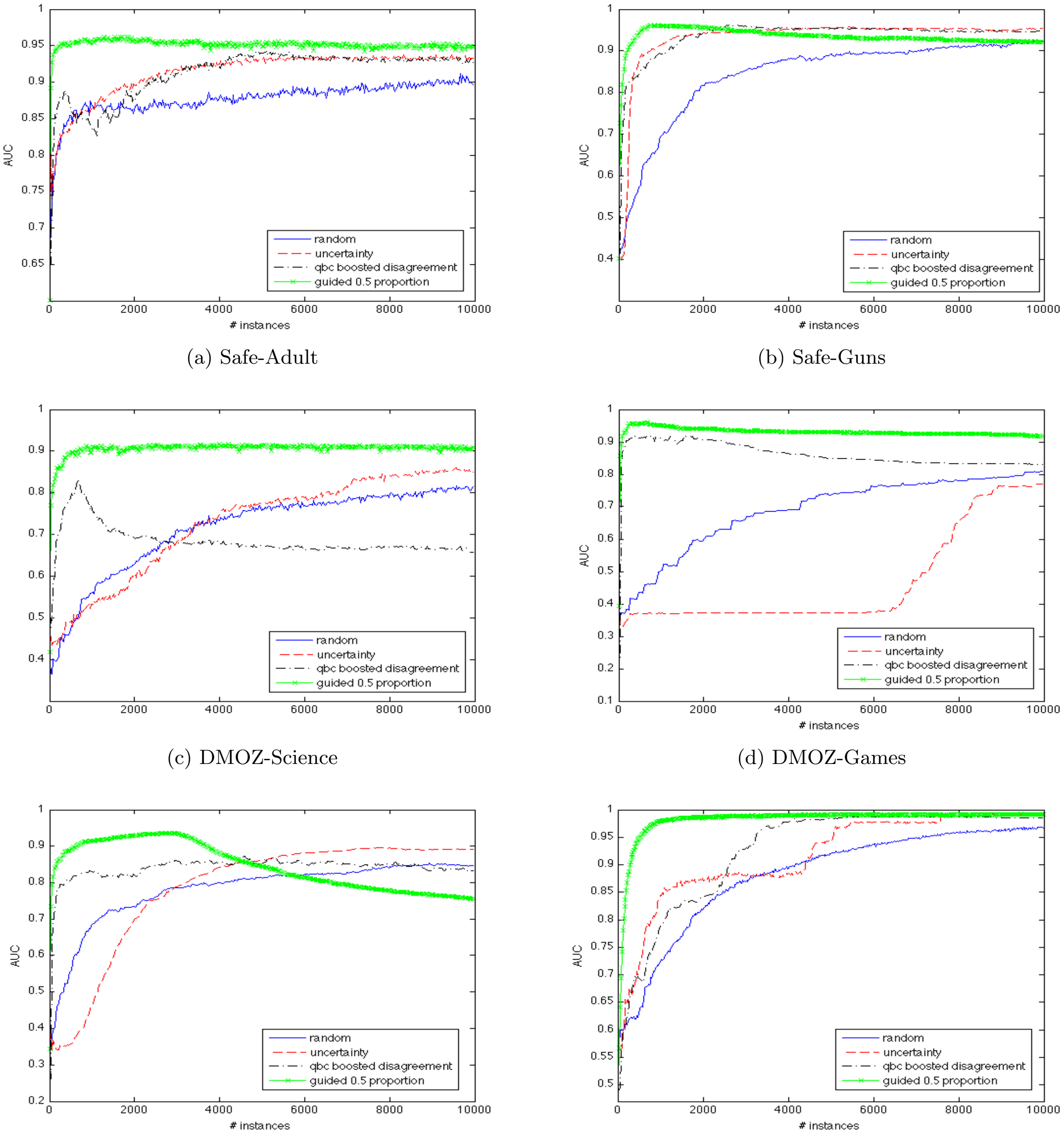
For this experiment, skew was induced by taking a large data set and randomly removing instances belonging to the minority class. This design choice exposes a limitation of our simulation framework; the absolute number of available minority instances critically influences the maximum performance that a model can achieve, becoming particularly evident as skew increases. We note that this limitation has the effect of handicapping guided learning; on the web, even cases that occur extremely infrequently in relation to the web as a whole still occur in great numbers in the absolute sense. While a guided learning scheme may require greater cost and effort to get new minority instances as the number of epochs increases, there is unlikely to be a hard cap on performance as seen in this experiment. It is clear that given enough information, a model can reach generalization performance in excess of AUC= 0.9; the performance of guided learning over all the learning curves present in this work, and in particular the high-skew settings in Figure 2, may therefore be underestimated.

6. INFLUENCE OF DISJUNCTIVE CLASSES ON ACTIVE LEARNING

As mentioned above, "disjunctive" concepts—classes to be modeled that are made up of distinct subclasses—have been shown repeatedly to cause problems for supervised induction. We hypothesized above that the unexpectedly poor performance of active learning for some of our data sets may be (in part) due to poorly modeled subclasses.

Figure 3 examines graphically the relative positions of the minority examples through the active learning. The black curve shows the AUC (right vertical axis) of the models learned by uncertainty sampling on the 20-News Groups data set as in Figure 5 (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 3 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 examples' relative positions changing. Minority examples drop down to the very bottom (certain minority) either because they get chosen for labeling, or because labeling some other example caused the model to "realize" that they are minority examples.

We see that early on the minority examples 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 examples are misclassified (above $\hat{P} = 0.5$). This is the point where the learning curve levels off. However, notice that there still are some residual misclassified minority examples, and in particular that there is a cluster of them for which the model is relatively certain they are majority examples. It takes many epochs for the active learning to select one of these, at which



(a) Safe-Adult

(b) Safe-Guns

(c) DMOZ-Science

(d) DMOZ-Games

(e) DMOZ-News

(f) 20-News-Groups

Figure 1: Comparison of active learning strategies and guided learning. The vertical axis shows the generalization performance of the learned models, measured by the area under the ROC curve (AUC). The horizontal axis shows the number of examples labeled/acquired. Uncertainty sampling and boosted disagreement outperform random sampling. Guided learning dominates by a large margin.

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 learning.

On the 20-News Groups data set we can examine the minority examples 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 switching examples are members of the cryptography (sci.crpyt) subcategory. These pages were classified as non-Science presumably because before having seen any positive examples 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).

Finally, let's return briefly to the surprising (to us) result that density-sensitive active learning techniques did not

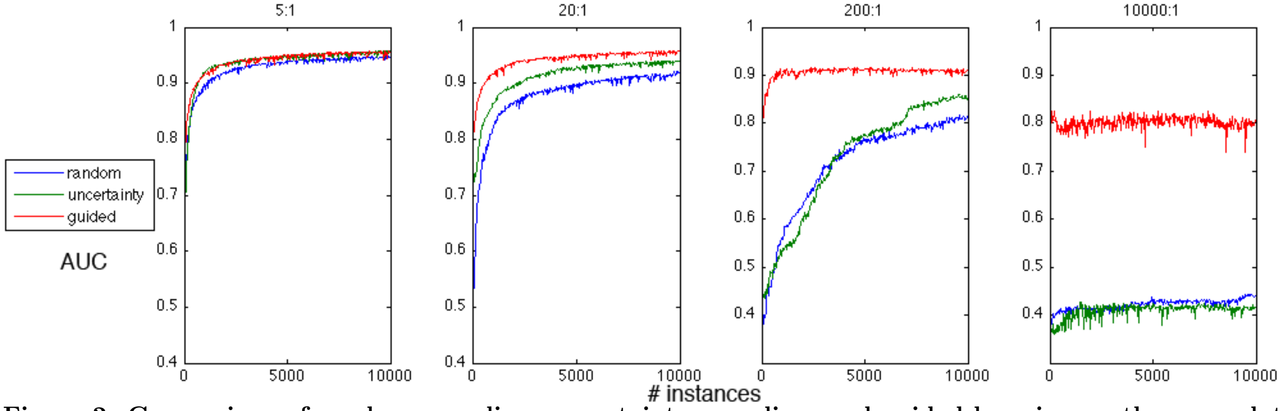


Figure 2: Comparison of random sampling, uncertainty sampling, and guided learning on the same data set with induced skews ranging from $5:1$ to $10,000:1$.

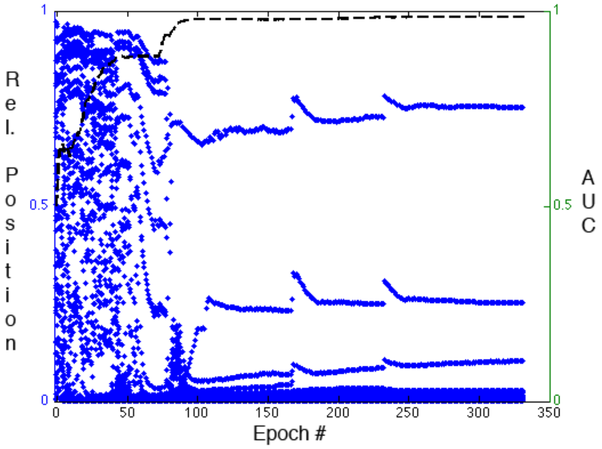


Figure 3: A comparison of the learned model's ordering of the example pool, along with the quality of the cross-validated AUC.

improve upon uncertainty sampling in this domain, when we have just provided some support for our intuition that the concepts are disjunctive, and thus one would expect a density-oriented technique to be appropriate. Unfortunately, for these domains—and we conjecture that this is typical of domains with extreme class imbalance—the *majority* class is even more disjunctive than the minority class. For example, in 20-News-Groups, Science indeed has four very different subclasses. However, non-Science has 16 (with much more variety). We expect the same thing to be true for safe advertising: there are certainly very different sorts of adult content, of hate speech, etc. But there are many more subclasses of non-objectionable content (the whole rest of the ad-supported web). Techniques that (for example) try to find as-of-yet unexplored clusters in the example space are likely to just get lost in vast and varied majority class.

7. COST-SENSITIVE GUIDED LEARNING AND HYBRID STRATEGIES

The per-instance cost for a guided learning strategy is likely to differ from that for strictly label-based active learning. Searching for an example of an obscure class may require more effort than simply identifying if a given sample belongs to the class of interest. Alternatively, using tools like web search engines, clear-cut examples may be readily found, whereas labeling would require time-consuming analysis of each case. The relative costs of guided learning and instance labeling vary from setting to setting, and in this section we seek to investigate the behavior in a variety of cost scenarios.

Figure 4 compares various instantiations of our guided learning approach with uncertainty sampling on the three data sets, where the curves show the increase in generalization performance as a function of investment in human effort (labeling or search). The horizontal axis shows the total cost expended by each strategy; to normalize, we fix the cost of labeling to be 1. For active learning, we only report for uncertainty sampling. The different instantiations of guided learning vary the relative cost of search (γ) from $\gamma = 0.5$ to $\gamma = 16$, doubling each step.

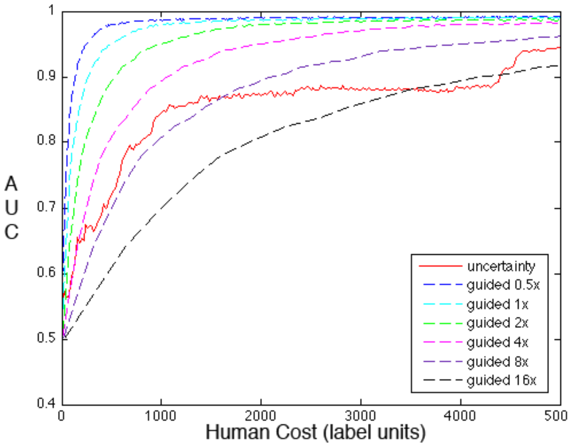


Figure 4: Comparison of guided learning and active learning under different relative costs for search and labeling (20-News-Groups data set). Horizontal axis shows total cost, normalized to 1 for acquiring one label. Here guided learning and active learning have equivalent performance-per-unit-cost when search is about 8 times more expensive than labeling.

By construction, the performance-per-unit-cost of guided learning declines gradually as the cost is increased. These results show how one can judge the relative value of applying human resources for search and for labeling. For example, for 20-News-Groups, in terms of performance-per-unit cost uncertainty sampling seems to be approximately equivalent to guided learning when search is approximately $\gamma = 8$ times more expensive than labeling.

As discussed, guided learning seems to be most appropriate in cases where the class priors are extremely unbalanced, and the cost structure is skewed in the opposite direction: discovering missing instances from the minority class is more expensive than missing examples from the majority class. The results presented up to this point have assumed that one would have to choose either guided learning or active learning. In practice, it might be better to mix strategies. For example, one might bootstrap the active learning process by first searching for good training data, potentially at a higher cost-per-example; alternatively, if one suspects that the active strategies have reached a plateau (as in Figure 1(f)), search may be used to inject additional information.

Given a budget, B, a data set, D, and a cost structure, C, policies for guided/active learning will allocate B to some combination of guided search and instance labeling. The goal of this of this section is to illustrate that hybrid guided/active strategies can be designed to offer superior performance for a given B than either strategy would be capable of in isolation.

While such a hybrid strategy could take many functional forms, here we propose a switching strategy inspired by the $DUAL$ technique used by Donmez et al [12]. We rely on our background knowledge that guided learning excels at finding different examples of the minority class, while many traditional active learning techniques are better at fine-tuning the decision boundary of the base model.

First, the technique estimates the benefit to model generalization of a purely guided selection strategy as a function of the cost of human effort. When the returns for further guided selection are sufficiently low, it switches to a purely active strategy. Note that almost any conventional active learning strategy may be employed for this phase of the hybrid approach. Due to its simplicity, documented success, and ubiquity, we chose uncertainty sampling as the active learning heuristic used in this second phase.

More succinctly, given a certain cost structure representing the cost-per-query to an oracle performing guided learning, we perform guided learning by selecting instances from both classes in proportion ρ. After each phase of guided learning, we estimate the performance, A, and use this performance estimate to construct a learning curve. When the expected gain for performing additional guided learning as a function of cost is sufficiently low, $\frac{\partial A}{\partial c} \leq \tau$, we switch from guided learning to a more traditional active learning strategy that requires only choosing examples from the pool for which to request labels.

In order to determine when to switch phases in our proposed switching strategy, we must understand how the performance of a model is changing under a given selection scheme, as a function of that scheme's cost, $\frac{\partial A}{\partial c}$. This requires careful estimation of the model's performance at each epoch. To accomplish this, we compute x-validated accuracy of the current model on the available pool of instances. Progress of learning curve is estimated empirically [20]. Here we use LOESS regression in order to smooth the variances in estimated learning rates at each epoch. More succinctly, the learning rate at any point is estimated by computing the slope of a least-squares linear regressor fit to performance estimates local to that point (in this case x-validated accuracy; cf.[20]). When the slope of accuracy as a function of cost drops below some threshold, τ, we change strategies from guided learning to active learning.

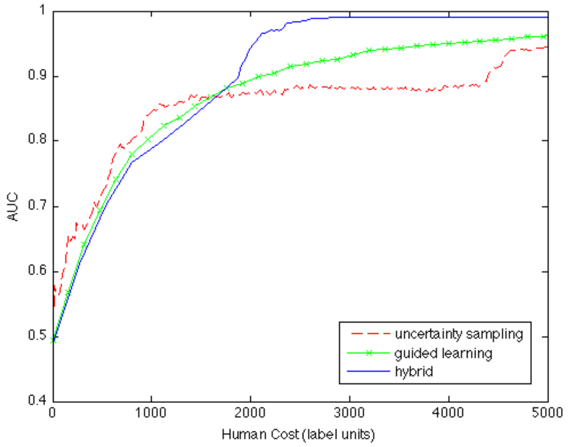


Figure 5: Comparison of our hybrid strategy with guided learning and uncertainty sampling.

The learning curves traced by the proposed hybrid technique on the 20-News-Group data set are presented in Figure 5, using the approximate break even cost of $\gamma = 8$. We see that under this cost setting, a switch from guided learning to active learning does indeed improve the learning rate beyond what is achieved by either component technique in isolation. Note that as the cost approaches approximately $2,000$, the slope of the learning curve increases drastically as the selection strategy switches from guided learning to active learning. This reveals an interesting limitation of the typical learning-curve evaluation. Before cost=2000, the learning curves seem to show fairly equivalent performance (with a slight advantage to active learning here at $\gamma = 8$). However, the hybrid strategy illustrates that the models likely are actually very different. After cost= 2000 both active learning and the hybrid are using the same acquisition strategy (uncertainty sampling). However, the hybrid performs *much* better. The difference is that the hybrid's model has benefitted from a strong exploratory phase, randomly sampling instances from across both classes, leaving it in a state very amenable to refinement by active learning.

8. GUIDED LEARNING IN THE WILD

Through out the preceding sections, we have used class-conditional random sampling as a simulation for a true guided learning system. This section presents results from a real guided learning system, AdSafe Media's Web Hound, where class-exemplary urls are collected to facilitate the swift production of statistical models. We provide a first experimental verification that the conclusions made through simulation throughout this work also hold for a realistic system.

AdSafe Media's Web Hound is a production system applying micro-outsourcing resources towards the construction of statistical models for use in a safe advertising system. A worker in a micro-outsourcing system is provided the definition of a class under consideration and tasked with finding examples of this class, using all tools available. Responses are checked for duplication, and optionally passed through an explicit labeling system to ensure correctness, thereby reducing noise and spam. The resulting urls are then passed

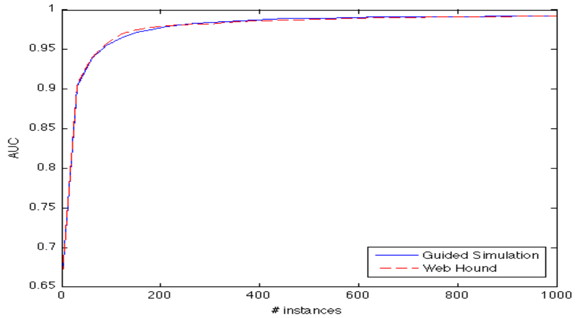


Figure 6: Comparison of our guided learning simulation with AdSafe Media's Web Hound system.

to a machine learning system where model induction is performed.

In order to determine if the results seen throughout the preceding sections indeed hold for such a production setting, we task users with finding examples of adult content and non-adult content in equal proportions. These results are held in a pool where training instances are drawn to build models and produce learning curves. The models induced by Web Hound are then compared to models created through the guided learning simulator accessing the hand labeled data set, *Safe-Adult*. In order to ensure that neither data set has an unfair advantage, we choose a third data set for evaluation. This test set consists of random web pages from the adult and non-adult portions of the DMOZ taxonomy. We perform ten folds of cross-validation evaluated on this external set and present the results in Figure 6. The results are surprisingly similar—both the simulation the real system are able to find examples that produce very accurate classification models, with very little human effort.

A system like Web Hound, based on the explicit selection class examples, while extremely flexible, is but one way to implement a guided learner. Depending on the details of one's application, it may be preferable to ask for key word queries, that, when posed to a search engine, are highly likely to return class-representative examples. Alternately, in certain settings, it may be easy to ask for directory pages, for instance sub-sets of DMOZ likely to contain examples of interest. However, the fine-granularity facilitated through Web Hound allows the explicit exploration of disjuncts- when portions of a class are poorly represented, instructions can be altered to seek more examples from these portions of the problem space. Finally, we point out that in the most skewed settings, it may be possible to select random unlabeled examples from the pool and just assume a negative example, rather than seeking majority instances explicitly. Depending on the base rate, the number of mistaken labels that result from such a strategy may be far lower that the typical human error resulting from a human labeling system.

The implementation choices made in a guided learning system have obvious impacts on the both the costs per instance, and on the distribution of the instances returned. Anecdotally, for our particular instantiation, we have observed per-instance costs between two and five times the cost per label.

9. CONCLUSION & LIMITATIONS

The main result of this paper—that guided learning can dominate active learning so strongly—raises the possibly contentious question of: when should we be doing active learning at all? The analysis of the hybrid techniques shows such situations exist, for instance, under low class skews and relatively high search cost. However, the more general results (Figure 4) make it clear that the question warrants further investigation.

This is not a trivial point. Research on active learning almost always makes one (or both) of two assumptions: (i) that labeling via initial random sampling is going to produce a model sufficiently accurate to do active learning, or (ii) that there is some "cold start" labeled-instance-acquisition process that provides the system with a small initial set of labeled examples to prime the process with a model of sufficient accuracy for active learning to be effective. With even moderately high skews, assumption (i) very often does not hold. With even $999 : 1$ skew, a labeler would have to label $30,000$ examples just to get 30 minority examples.

Most research starts with assumption (ii) that some (usually unspecified) labeled-instance-acquisition process has produced a small labeled training set (often balanced). The results of this paper raise the question: why not just continue with that labeled-instance-acquisition process?! Why do active learning at all? It is possible that the acquisition process was an unrepeatable (historic) stroke of good fortune.[6]

Otherwise, these results suggest that we may want to put more research emphasis and development investment on/into such processes. What is the relative cost as compared to active learning? Under what conditions does it make sense to continue with this process, versus switching to active learning, versus applying some combination of both? Clearly more sophisticated hybrid approaches will be designed than the simple switching strategy we use. For example, we would like to be able to judge the relative benefit-per-unit-cost of different human deployment strategies and allocate resources accordingly. Moreover, perhaps we should be investing more effort in reducing the cost of search for examples.[7] What mechanism design issues arise in building systems for cost-effective guided learning, for example using micro-outsourcing systems [24]?

While many of the analyses presented in this paper are simulations using class-conditional random sampling, we justify many of our conclusions by showing similar acquisition performance for both our simulator and the AdSafe Web Hound. This one experiment does not explore issues surrounding a production guided learning system—any implementation is likely to have some biases induced by the method of selection, be that DMOZ, a search engine, or a user's imagination. As a result, the instances returned from a guided learning system may differ significantly from a class-conditioned uniform random selection in certain cases.

[6] For example, in our safe advertising example, an advertiser complained about a particular set of web pages on which her ad appeared, which then become labeled training data.

[7] An initial reaction to the question of quantifying the relative cost of search versus labeling often is: search *has* to be more expensive than labeling, since with search one must find the examples as well as label them. However, with a search engine and a human brain, it may be less costly to envision what a case might look like and find it, than to examine a presented case in detail to be sure. For example, one may be able to search and find examples of all manner of hate speech on the web more efficiently than reading carefully through a borderline web page to determine whether or not it contains some form of hate speech.

This could bias the results above in either way: they could be overly pessimistic if initial instances returned would in some sense be the best, most informative representatives of that particular class. They could be overly optimistic if the internal biases of the search engine or the human user made the selection of certain examples unlikely. Research has shown repeatedly that classifiers' errors are concentrated in the "small disjuncts"—the model's representations of the rare subclasses [29]. In light of the disjunct-oriented results presented in Section 6 and our preliminary discouraging results using density-sensitive active learning, the problem of finding rare subclasses both with guided and active learning requires further investigation.

Relatedly, above we mentioned that in practice extreme imbalance often is reduced by orders of magnitude via filters on examples. Such filters are essentially codified search procedures for guided learning. These filters also suffer from potential bias problems, and in our experience the bias is substantially more extreme than with human searchers, because of the inflexibility. These filters bias the data largely toward particular disjuncts of the interesting category. In practice, this introduces a particularly insidious problem for data mining: the testing data are biased as well!

Thus, it is necessary to provide tools that simultaneously reduce the cost of search, and challenge humans to explore the "far reaches" of the category. This suggests a different integration with active learning: guiding the human search away from cases already "known" by the model to be members of the class, and (somehow) toward as-of-yet unexplored disjuncts.

Our simulation has other important differences from a real-world implementation of guided learning, particularly in the case of document search on the web. First, the extreme size of the web is likely to yield a near-unlimited number of instances belonging to almost any category. Thus, unlike with our experiments, the guided learning curves would not run out of minority class examples (which is often the apparent reason they "knee over" so sharply). On the other hand, many instances in the web's long tail may be increasingly difficult to find. This could mean that the constant-cost model is simplistic; search probably incurs increasing cost as the number of requested instances increases, and as the exploration of the class probes the smaller disjuncts.

This paper does not explore the possibility that guided learning might focus solely on finding minority classes. In cases of extreme skew, one likely could just presume a randomly selected case to be a majority class, and deal with the small amount of resultant noise. This would effectively halve the cost of guided learning as we present it.

There still is much to do to understand the best ways to employ human resources for some combination of search and labeling, to produce the best models per unit cost in training data. We hope that this paper has made useful headway.

10. REFERENCES

[1] `http://www.dmoz.org/`.

[2] `https://www.mturk.com/mturk/welcome`.

[3] Content category definitions & rating guidelines. New York, NY, USA, January. AdSafe Media, Inc. `http://www.adsafemedia.com/adsafe_rating_guide.php`.

[4] Beyond the grey areas: Transparency, brand safety and the future of online advertising. New York, NY, USA, April 2010. Winterberry Group LLC.

[5] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[6] J. Atteberg and F. Provost. Toward optimal allocation of human resources for active learning with application to safe advertising. In *Winter Conf. on Business Intel.*, 2010.

[7] J. Attenberg and F. Provost. Toward optimal allocation of human resources for active learning with application to safe on-line advertising. Stern working paper ceder-09-07, 2009.

[8] 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 '09*, 2009.

[9] N. V. Chawla, K. W. Bowyer, and P. W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[10] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *In Proc. of the Fifth Intl. Conf. on Knowledge Discovery and Data Mining*, 1999.

[11] P. Donmez and J. Carbonell. Paired Sampling in Density-Sensitive Active Learning. In *Proc. of the 10 Intl. Symp. on Artificial Intelligence and Mathematics*, 2008.

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

[13] S. Ertekin, J. Huang, L. Bottou, and L. Giles. Learning on the border: active learning in imbalanced data classification. In *CIKM '07: Proc. of the sixteenth ACM Conf. on information and knowledge management*, 2007.

[14] T. Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. Technical report hpl-2003-4, 2003.

[15] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR '94: Proc. of the 17th Annual Intl. ACM SIGIR Conf on Research and development in Info. Retreival*, 1994.

[16] X. Y. Liu, J. Wu, and Z. H. Zhou. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, 39, 2009.

[17] R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and M. Friedl. Active class selection. In *ECML '07: Proc. of the 18th European conference on Machine Learning*, pages 640–647, Berlin, Heidelberg, 2007. Springer-Verlag.

[18] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Beyond blacklists: learning to detect malicious web sites from suspicious urls. In *KDD '09*, 2009.

[19] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML '04: Proc. of the 21st Intl. Conf. on Machine learning*, New York, NY, USA, 2004.

[20] F. J. Provost, D. Jensen, and T. Oates. Efficient progressive sampling. In *KDD'99*, 1999.

[21] S. Rajan, D. Yankov, S. Gaffney, and A. Ratnaparkhi. A large-scale active learning system for topical categorization on the web. In *WWW*, 2010.

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

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

[24] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD '08: Proc. of the 14th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, 2008.

[25] K. Tomanek and U. Hahn. Reducing class imbalance during active learning for named entity annotation. In *K-CAP '09: Proc. of the 5th Intl. Conf. on Knowledge capture*, 2009.

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

[27] K. Weinberger, A. Dasgupta, J. Attenberg, J. Langford, and A. Smola. Feature hashing for large scale multitask learning. In *ICML '09*, 2009.

[28] G. Weiss and F. Provost. The effect of class distribution on classifier learning. Rutgers technical report ml-tr-44, 2001.

[29] G. M. Weiss. The impact of small disjuncts on classifier learning. In *Data Mining*, volume 8, pages 193–226. 2010.

[30] G. M. Weiss and F. Provost. Learning when training data are costly: the effect of class distribution on tree induction. *J. Artif. Int. Res.*, 19(1):315–354, 2003.

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

[32] 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.