Guided Feature Labeling for Budget-Sensitive Learning Under Extreme Class Imbalance

Josh Attenberg JOSH@CIS.POLY.EDU
Polytechnic Institute of NYU, Brooklyn, NY 11201

Prem Melville PMELVIL@US.IBM.COM
IBM Research, Yorktown Heights, NY 10598

Foster Provost FPROVOST@STERN.NYU.EDU
NYC Stern School of Business, New York, NY 10012

Abstract

Extreme class skew is a hurdle in many machine learning tasks. In such skewed settings, traditional methods for procuring labeled examples, including random sampling and active learning, are often ineffective—they struggle to find representative minority examples. The framework of Dual Supervision, which incorporates feature-based background information into traditional supervised learning, provides one avenue to combat this problem. However, active learning for feature information (feature labeling), like active learning, is often not resilient to extreme class skew. In this work, we present an alternative to active feature labeling, *Guided Feature Labeling*. In this paradigm, human domain experts are tasked with finding class-indicative features given a description of a class. This work explores different data acquisition costs, and demonstrates that under certain conditions, Guided Feature Labeling does indeed offer high performance models at a far lower budget than complementary active labeling approaches.

1. Introduction

This paper provides empirical support for the efficacy of alternative techniques for gathering and incorporating human resources during the data acquisition phase of classifier induction. The general class of techniques presented herein, *Guided Feature Labeling*, are motivated by classification problems where one class occurs in far greater numbers than the other. While the underlying results and techniques can be applied to a wide range domains, classifier functions, and feature vectors, we motivate this work with 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. For this paper, we will consider only the textual html source for each page, however, the ideas should 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 amount and quality of label information used when performing model induction. For a given budget, some subset of web pages

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

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

can be examined by humans—potentially at very low cost using a micro-outsourcing system (Sheng et al., 2008) such as Amazon's Mechanical Turk[2]—to produce training data for a supervised learning model.

Traditionally, in order to reduce the necessary cost for building high quality models, active learning is employed as a mechanism for selecting only those instances for labeling that the model perceives as most beneficial for training. Typically this learning occurs in epochs: at each epoch the current model computes some utility score for all known instances. Those instances with the highest utility are labeled and subsequently incorporated into the training data. The model is then retrained and the process repeats until the budget is exhausted.

Attenberg and Provost 2010 demonstrate empirically that traditional active learning techniques fail in such extreme class skew settings, such as those faced in safe advertising, proposing *guided learning* (referred to as guided instance labeling in this work to avoid confusion) as a solution. The guided instance labeling paradigm presents human oracles with a concise description of a class being considered. The oracles are then tasked with *seeking* instances that represent this class, using their background knowledge of the problem and whatever tools they are comfortable using, such as a search engine. Attenberg and Provost assigned a variety of costs associated with this guided seeking, showing that in many high skew settings, guided instance labeling can completely eclipse the performance of an active learner for a given budget, even when the costs-per-guided-example are several times the costs per label.

As a supplement to traditional supervised learning, human background information regarding the class polarity of individual features can be incorporated into predictive systems via dual supervised models (Melville et al., 2009; Sindhwani & Melville, 2008). In the case of our motivating example, this information would take the form of classes associated with given terms (feature labels); however, any type of background feature associations can potentially incorporated into a suitable model. In order to efficiently allocate a limited budget in a dual-supervised setting, prior work has investigated active dual supervision, interleaving the predictions of both components of a dual supervised model in order to select informative instances and features for labeling at each epoch (Melville & Sindhwani, 2009a; Sindhwani et al., 2009). While dual supervision provides an interesting alternative path for building machine learning models, this strategy too has difficulties adequately exploring complex, highly skewed problem spaces, as we see in Figure 1.

As an analogue to guided instance labeling for example labels, this work introduces *Guided Feature Labeling* to provide an alternate source for incorporating human knowledge into the model building process. Guided Feature Labeling tasks human oracles with finding features (or feature values) that are likely to indicate membership in the class of interest. For instance, in the problem of safe advertising, an oracle may provide terms that tend to appear in certain types of offensive content. Guided Feature Labeling can be contrasted with feature labeling, where oracles are presented a feature with an unknown class affiliation, and asked to provide a label. As seen in Figure 1, our Guided Feature Labeling can offer superior generalization performance for a given number of examples than active learning, active feature labeling, and even guided instance labeling in settings with moderate class skew. We use this potential performance advantage to motivate further investigation into Guided Feature Labeling.

The remainder of this work proceeds as follows: in Section 2 we offer an explanation of Guided Feature Labeling and present a review of guided instance labeling. Section 3 covers the datasets and details relevant to the experiments performed throughout this work. Section 4 explores different acquisition costs for the different strategies investigated herein, showing that Guided Feature Labeling is competitive even when costing several times as much as explicit labeling. Section 5 provides a detailed review of prior work, and Section 6 provides a conclusion and gives directions for future work.

2. Guided Instance Labeling and Guided Feature Labeling

Guided learning (Attenberg & Provost, 2010) 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 that paper, the basic guided instance labeling task is straightforward: find examples representing the different classes in some proportion, ρ. These instances are provided as input to classifier induction.

Humans, using tools such as web search engines combined with their own background knowledge on the criteria defining the task, can often find informative examples with an efficiency far exceeding that which is possible by a model-based active learner. This is par-

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

ticularly true in settings with highly skewed, disjunctive concepts, and in the early stages of active learning where the model does not have a refined knowledge of the input space. In these settings, models induced through a guided instance labeling process are often able to achieve a far greater level of generalization performance at a given cost than is possible through many active learning schemes, including active learning schemes designed specifically for high skew settings, e.g. (Tomanek & Hahn, 2009; Bloodgood & Shanker, 2009).

Active feature labeling and active dual supervision have been proposed as alternative applications of human resources towards the construction of classification models (Melville & Sindhwani, 2009b). At each epoch in this setting the base classifier selects features (or features and instances in the case of active dual supervision) that it believes will be most informative. These features are presented to oracles for labeling, and the resulting feature labels, representing class associations, are incorporated into the subsequent classifier. As we show, in settings where the base learner does not have at least a rudimentary understanding of the problem space, active dual supervision, like active learning, suffers from poor selections. This may often be the case in highly skewed problems, where those features that tend to indicate the minority class may not have been discovered yet. In this case, the label selection technique likely degrades to random sampling of features to label, implying that any difficulties performing active feature labeling would be exacerbated by high dimensional representations of the problem space.

Guided Feature Labeling is a technique analogous to guided instance labeling for the task of providing feature-class associations. Rather than having human oracles seek class representative examples, Guided Feature Labeling tasks laborers with finding discriminative (polar) features associated with the class of interest. In the case of our motivating example, safe advertising, this would include polar terms that may indicate membership in the class of offensive pages. However, other domain-specific information is equally applicable, be that certain symptoms or physical traits being associated with certain rare disease, or customer characteristics that may indicate an increased likelihood to buy a product. This acquisition approach is fundamentally different from active learning or active feature labeling; rather than have a learner select informative instances or features for an oracle to *label*, oracles are asked to *seek* features that represent a class. In a dual supervision setting, this feature label information can supplement the instance-supervised portion of the model, providing a potential bootstrap to help an active learner (or active dual learner) overcome the cold start problem that plagues high-skew model construction.

3. Experimental Setup and Results

All experiments presented below are conducted on a set of $35,000$ web pages extracted from a stream of real ad impressions. Each url has been hand labeled as to the presence or absence of adult content. In the setting considered here, positive instances are deemed unsafe for advertising. This dataset has a class skew of roughly 80 to 1.

3.1. Experimental Framework

Dual Supervision techniques can be divided into two categories—*early fusion* and *late fusion*. Early fusion approaches, such as (Melville et al., 2009; Sindhwani & Melville, 2008), integrate instance and feature supervision to produce a single composite model. In contrast, a late fusion approach builds separate models based on labeled instances and labeled features, and combines the outputs of these models in order to classify a new instance. The advantage of a late fusion approach is that any existing supervised classifier can be used to learn from labeled instances, without requiring to adapt it to incorporate labeled features. We explore one such late fusion approach here.

In particular, we estimate the probability that instance x_j belongs to class y_i by the convex combination of instance-based and feature-based probability estimates:

$$P(y_i|x_j) = \alpha P_e(y_i|x_j) + (1-\alpha)P_f(y_i|x_j)$$

Here $P_e(y_i|x_j)$ and $P_f(y_i|x_j)$ represent the probabilities generated by the instance based and feature-label based models respectively. The parameter α is the weight for combining these two probability estimates, representing the confidence placed in each predictor. While clever selection of α could no doubt improve predictive performance, we chose to fix its value at 0.5 for the purpose of this work, in order to isolate our results from influence of weight selection.

The instance-based component, $P_e(y_i|x_j)$, is computed through a logistic regression classifier: $P_e(y_i|x_j) = \frac{1}{1+e^{-w^T x_j}}$. For our motivating example, x_j is a vector space representation of each document, with each entry corresponding to the encountered frequency of a certain term in a document. For the feature-based component model, we follow (Melville et al., 2009), assuming that the feature/class associations provided by

the oracle are the conclusion of many instance observations. We seek a Naïve Bayes probability estimator that is likely to generate these instances. Given:

$\mathcal{V}$ — the "vocabulary", i.e., the set of features in the problem domain
$\mathcal{P}$ — the set of features with positive class associations
$\mathcal{N}$ — the set of features with negative class associations
$\mathcal{U}$ — the set of unknown words: $\mathcal{V} - (\mathcal{P} \cup \mathcal{N})$
m — the number of features in the domain: $|\mathcal{V}|$
p — the number of positive features: $|\mathcal{P}|$
n — the number of negative features: $|\mathcal{N}|$

We segregate the feature space into three classes: positive, negative, and unknown to reflect the feature/class associations. Features belonging to these three classes are denoted f_p, f_n, and f_u respectively. The likelihood components of the Naïve Bayes model are given by:

$$P(f_p|y_p) = P(f_n|y_n) = \frac{1}{p+n}$$

$$P(f_p|y_n) = P(f_n|y_p) = \frac{1}{r}\frac{1}{p+n}$$

$$P(f_u|y_p) = \frac{n(1-1/r)}{(p+n)(m-p-n)}$$

$$P(f_u|y_n) = \frac{p(1-1/r)}{(p+n)(m-p-n)}$$

The class probability for a given $x = \vec{f}$ is then:

$$P_f(y|x) = \frac{P(y)}{P(x)} \prod_{f \in x} P(f|y)$$

Throughout this work, we use $r = 100$. The derivation for this probability estimator can be found in (Melville et al., 2009). While the instance-trained and background-trained components of this late-fusion dual-supervised model provide efficient, reasonable predictions, the strategy presented in this paper does not depend on the specific functional form of classifier used.

Each document is encoded with the standard vector space representation by storing the frequencies of the 4,096 terms with the highest global information gain. All experiments are averages across ten folds of cross validation. Due to the skew involved in our experiments, at each epoch, we report the area under the ROC curve. Each epoch consists of thirty additional instances or features being revealed to the model.

As with most active learning research, we simulate how a real learning system would behave in a realistic setting. In the case of guided instance labeling and Guided Feature Labeling, we rely on class-conditional random sampling. For Guided Feature Labeling, this entails choosing features uniformly at random according to some proportion, ρ; $\rho|b|$ of the features gathered at each epoch have minority class associations, and $(1-\rho)|b|$ are randomly selected from those features with majority class associations. Here, $\rho \in [0, 1]$, and $|b|$ is the number of selections made at each epoch. A similar simulation is used in the guided instance labeling setting; class-conditional random selection is used in accordance to some mixing parameter. Experimental validation in (Attenberg & Provost, 2010) shows that models trained using actual explicit human gathering behave similarly to those trained through guided instance labeling.

The feature polarity oracle is simulated by assigning labels to the 500 terms in the data set with the highest information gain. The label assigned to each of these terms corresponds to the class with the highest likelihood, e.g. $\arg\max_y \frac{p(y|f)}{p(y)}$ for class y and term f. The remaining terms in the vector space receive an "unknown" label not denoting any class polarity. In our setting, all features are treated as unknown by default; therefore, a returned label of unknown does not change the state of the feature-based classifier.

3.2. Experimental Results

Figure 1 presents a comparison of guided instance labeling, Guided Feature Labeling, traditional uncertainty sampling on instances, certainty sampling on features[3] as in (Melville & Sindhwani, 2009b), and uniform random sampling from both features and instances. For this particular experiment, we assume an equal cost of data acquisition for labeled instances, features, and for the guided selection of both instances and class-indicative features. We see that given this equal cost assumption, the guided techniques clearly dominate random sampling and uncertainty sampling for our example problem setting. Note that while certainty sampling on features may waste many queries on features with "unknown" class associations, thereby not offering any improvement to the subsequent classifier, Guided Feature Labeling requests result in a feature with a definitive class association. This greatly improves the chances that a model constructed under Guided Feature Labeling is able to see gains in generalization performance over successive epochs.

[3]Sampling most *certain* features was found to perform better than sampling uncertain features.

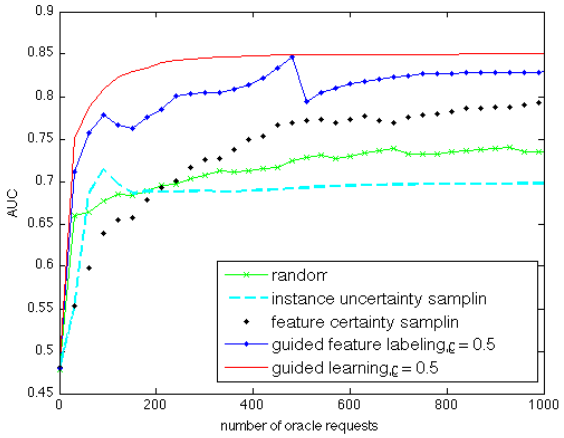

Figure 1. Comparison of data acquisition strategies. The vertical axis shows the generalization performance of the models, the horizontal axis shows the total cost incurred for labeling.

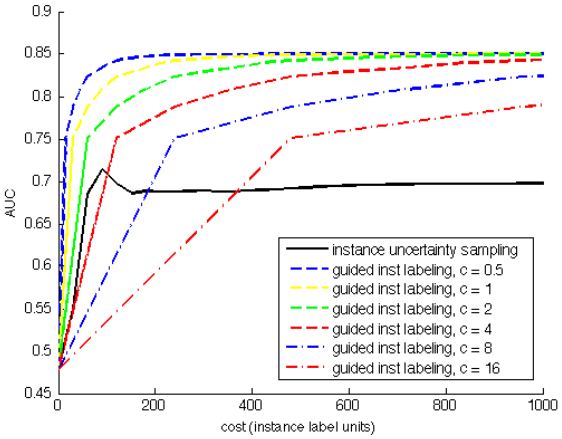

Figure 2. Comparison of different guided instance selection acquisition costs in comparison to traditional uncertainty sampling.

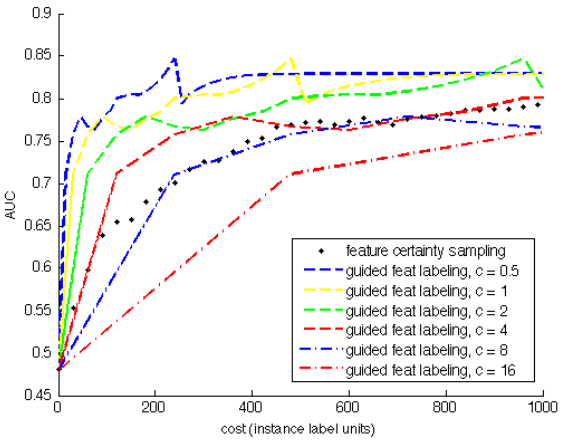

Figure 3. Comparison of different guided feature labeling acquisition costs in comparison to feature certainty sampling.

4. Budget-Sensitive Data Acquisition

Figure 1 clearly shows the potential efficacy of guided selection techniques over conventional learning techniques in certain settings. However, a uniform cost over the four types of proposed efforts may be unrealistic. The search process necessary for guided techniques likely requires a greater deal of human effort than simple labeling, and therefore may incur a greater cost per instance.

In this section, we seek to explore the generalization performance achievable at a given cost for models trained on different data acquisition strategies. Figures 2 and 3 present different cost structures for performing guided instance labeling and Guided Feature Labeling respectively. Here $c = 1$ indicates that one instance selected by guided instance selection costs as much as one label request. We note from these figures that even at fairly extreme cost settings, guided instance selection is able to achieve much better generalization performance than uncertainty sampling for a given budget. This distinction is less pronounced in the case of Guided Feature Labeling (Figure 3), where a cost of about eight labels per guided request, feature certainty sampling seems preferable in our sample problem.

Guided instance selection excels in cases where human experts can identify representative examples more easily than a machine learning algorithm with little knowledge of the problem space. As the human oracles' difficultly of finding examples from the minority class increases, so is the acquisition cost likely to increase. However, since the problem has been posed to a machine learning system, it is to be expected that system designers or human oracles can provide descriptions of positive instances, *features* that are likely to distinguish the positive class from the bulk of instances. In the dual supervision setting, this background knowledge can be encoded into a dual supervision model, and is the essence of Guided Feature Labeling. Figure 4 explores some possible cost scenarios. We see that as the cost of instance selection rises, it becomes preferable to perform Guided Feature Labeling. Note that a cost-per-feature of $c = 2$, Guided Feature Labeling is very competitive with guided instance labeling, when the latter incurs a cost-per-instance of $c = 8$. At this cost setting, both guided strategies are superior to active learning on either features or instances.

5. Related Work

There has been a great deal of research on active learning in the machine learning community. Often this

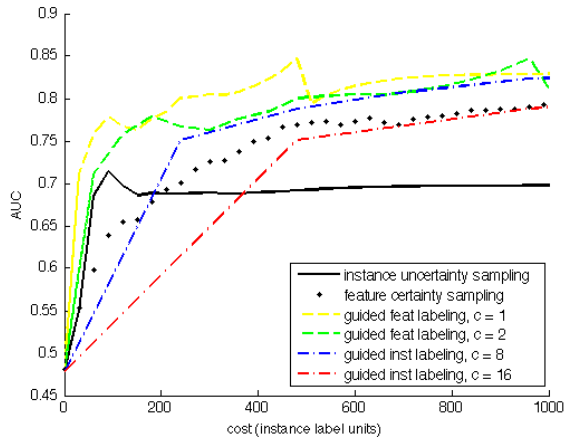

Figure 4. Comparison of different guided feature labeling and guided instance selection facquisition costs in comparison to feature certainty sampling and instance uncertainty sampling.

work assumes the active learner is given some initial set of data upon which initial models can be built. However, the cost of acquiring this initial set is often ignored. Attenberg and Provost 2010 proposed a generalization of this process, guided instance labeling, where class conditional instances can be acquired from an oracle for a certain cost. They demonstrated that under certain cost assumptions, simply continuing the process of having oracles actively acquire data may dramatically outperform active learning, even with significant imbalance in acquisition costs.

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 (Chawla et al., 2002; Liu et al., 2009). 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 (Domingos, 1999).

There has been some work on active learning on skewed data. Tomanek and Hahn 2009 investigate 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 2009 use a similar approach to (Tomanek & Hahn, 2009), 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 2007 investigate active learning in conjunction 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 focus on learning with highly imbalanced data sets. Given a large, imbalanced pool of labeled instances, the authors randomly sub-sample instances, choosing to keep only those that are closely positioned 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. , work extended by Donmez and Carbonell 2008. This work seeks to find "clusters" of distinct content among the unlabeled instances. While these techniques offer greater potential overcoming the cold-start than many common active learning techniques, they still are unlikely to succeed in the extremely skewed case. There is often so much diversity within the majority that such a method will miss the minority instances. Additionally, these complex methods don't scale well to the data sizes necessary to experience an extreme class skew.

Donmez et al 2007 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.

Active learning in the context of dual supervision mod-

els is a new area of research with very little prior work, to the best of our knowledge. Most prior work in active learning has focused on pooled-based techniques, where examples from an unlabeled pool are selected for labeling (Cohn et al., 1994). In contrast, active feature-value acquisition (Melville et al., 2005) and budgeted learning (Lizotte et al., 2003) focus on estimating the value of acquiring missing features, but do not deal with the task of learning from feature *labels*. Raghavan et al. 2007 and Raghavan et al. 2006 study the problem of *tandem learning* where they combine uncertainty sampling for instances along with co-occurence based interactive feature selection. Godbole et al. 2004 propose notions of feature uncertainty and incorporate the acquired feature labels into learning by creating one-term mini-documents. Druck et al. 2009 perform active learning via feature labeling using several uncertainty reduction heuristics. Sindhwani et al. 2009 also study the problem of active dual supervision, applied to a graph-based dual supervision method. They explore various heuristic approaches to active learning for instances and features separately. In order to interleave selections from both instances and features, they randomly probe an active instance learner or an active feature learner for the next query. In contrast, we take a holistic approach to active dual supervision, where by estimating the potential value of features and instances on the same scale, we select the type of acquisition that is most likely to benefit our classifier.

Learning from labeled examples and features via dual supervision is itself a new area of research. Sindhwani et al. 2008 use a kernel-based framework to build dual supervision into co-clustering models. Sindhwani and Melville (Sindhwani & Melville, 2008) apply similar ideas for graph-based sentiment analysis. There have also been previous attempts at using only feature supervision, mostly along with unlabeled documents. Much of this work (Schapire et al., 2002; Wu & Srihari, 2004; Liu et al., 2004; Dayanik et al., 2006) has focused on using labeled features to generate pseudo-labeled examples that are then used with well-known models. In contrast, Druck et al. 2008 constrain the outputs of a multinomial logistic regression model to match certain reference distributions associated with labeled features. In a similar vein, Liang et al. 2009 learn from labeled examples and constrains on model predictions.

6. Conclusions and Future Work

Guided feature labeling provides an alternative strategy for taking advantage of human resources for improving supervised learning: humans are tasked specifically with providing (finding) class-indicative features. In combination with a dual-supervision system, which allows class-polarity information about features to be taken as input, guided feature labeling allows humans quickly to prime the modeling procedure based on their background knowledge of the domain. In addition, humans can continue to add value if it is possible to search for class-indicative features with the aid of search engines or other tools. This paper demonstrates that for certain settings, using guided feature labeling can result in generalization performance far exceeding that obtainable using traditional instance- or feature-based active learning in a dual-supervision setting.

This work is a first foray into the guided acquisition of feature information, and is meant to serve as a motivation for future work. This future work includes intelligent mixing of guided instance selection, guided feature labeling, active learning, and active dual supervision. Such work would be particularly beneficial when the estimated costs of each strategy can be estimated, and then can be used to guide choices made at each epoch. Additionally, more sophisticated active learning strategies may effectively interleave the selection of instances and features for labeling. Such interleaving is also likely to benefit any guided selection strategy. A thorough study of the latest active dual supervision strategies in concert with guided instance labeling is a promising direction for future research. We also conjecture that guided feature labeling is an ideal candidate to help alleviate the bootstrapping, or "cold-start," problem for active learning. Finally, we would like to evaluate real human oracles' abilities at performing guided feature labeling, comparing this to our simulation strategy, thereby assessing the viability of our proposal in real data mining settings.

References

Attenberg, Josh and Provost, Foster. Why label when you can search? strategies for applying human resources to build classification models under extreme class imbalance. In *KDD '10*, 2010.

Bloodgood, Michael and Shanker, K. Vijay. 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.

Chawla, Nitesh V., Bowyer, Kevin W., and Kegelmeyer, Philip W. Smote: Synthetic minor-

ity over-sampling technique. *J. Artif. Int. Res.*, 16: 321–357, 2002.

Cohn, D., Atlas, L., and Ladner, R. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Dayanik, A., Lewis, D., Madigan, D., Menkov, V., and Genkin, A. Constructing informative prior distributions from domain knowledge in text classification. In *SIGIR*, 2006.

Domingos, Pedro. Metacost: A general method for making classifiers cost-sensitive. In *KDD '09*, 1999.

Donmez, P. and Carbonell, J. Paired Sampling in Density-Sensitive Active Learning. In *Proc.* 10^{th} *Int. Symp on Artificial Intel. and Mathematics*, 2008.

Donmez, Pinar, Carbonell, Jaime G., and Bennett, Paul N. Dual strategy active learning. In *ECML '07*, 2007.

Druck, G., Mann, G., and McCallum, A. Learning from labeled features using generalized expectation criteria. In *SIGIR*, 2008.

Druck, G., Settles, B., and McCallum, A. Active learning by labeling features. In *EMNLP '09*, pp. 81–90. Association for Computational Linguistics, 2009.

Ertekin, Seyda, Huang, Jian, Bottou, Leon, and Giles, Lee. Learning on the border: active learning in imbalanced data classification. In *CIKM '07*.

Godbole, S., Harpale, A., Sarawagi, S., and Chakrabarti, S. Document classification through interactive supervision of document and term labels. In *PKDD*, 2004.

Liang, Percy, Jordan, Michael I., and Klein, Dan. Learning from measurements in exponential families. In *ICML*, 2009.

Liu, Bing, Li, Xiaoli, Lee, Wee Sun, and Yu, Philip. Text classification by labeling words. In *AAAI*, 2004.

Liu, X. Y., Wu, J., and Zhou, Z. H. Exploratory undersampling for class-imbalance learning. 2009.

Lizotte, D., Madani, O., and Greiner, R. Budgeted learning of naive-Bayes classifiers. In *UAI*, 2003.

Melville, Prem and Sindhwani, Vikas. Active dual supervision: Reducing the cost of annotating examples and features. In *NAACL HLT 2009*, 2009a.

Melville, Prem and Sindhwani, Vikas. Active dual supervision: reducing the cost of annotating examples and features. In *HLT '09*, 2009b.

Melville, Prem, Saar-Tsechansky, Maytal, Provost, Foster, and Mooney, Raymond. An expected utility approach to active feature-value acquisition. In *ICDM*, 2005.

Melville, Prem, Gryc, Wojciech, and Lawrence, R. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD*, 2009.

Raghavan, H., Madani, O., and Jones, R. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *SIGIR*, 2007.

Raghavan, Hema, Madani, Omid, and Jones, Rosie. Active learning with feedback on features and instances. *J. Mach. Learn. Res.*, 7, 2006.

Schapire, Robert E., Rochery, Marie, Rahim, Mazin G., and Gupta, Narendra. Incorporating prior knowledge into boosting. In *ICML*, 2002.

Sheng, Victor S., Provost, Foster, and Ipeirotis, Panagiotis G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD '08*, 2008.

Sindhwani, Vikas and Melville, Prem. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, 2008.

Sindhwani, Vikas, Hu, Jianying, and Mojsilovic, Alexandra. Regularized co-clustering with dual supervision. In *NIPS*, 2008.

Sindhwani, Vikas, Melville, Prem, and Lawrence, Richard. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, 2009.

Tomanek, Katrin and Hahn, Udo. Reducing class imbalance during active learning for named entity annotation. In *K-CAP '09*, 2009.

Wu, Xiaoyun and Srihari, Rohini. Incorporating prior knowledge with weighted margin support vector machines. In *KDD*, 2004.

Zhu, Jingbo and Hovy, Eduard. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *EMNLP-CoNLL 2007*, 2007.

Zhu, Jingbo, Wang, Huizhen, Yao, Tianshun, and Tsou, Benjamin K. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING '08*.