JMLR: Workshop and Conference Proceedings 16 (2011) 1–18 Workshop on Active Learning and Experimental Design

From Theories to Queries: Active Learning in Practice

Burr Settles BSETTLES@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh PA 15213 USA

Editor: I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov

Abstract

This article surveys recent work in *active learning* aimed at making it more practical for real-world use. In general, active learning systems aim to make machine learning more economical, since they can participate in the acquisition of their own training data. An active learner might iteratively select informative *query* instances to be labeled by an *oracle*, for example. Work over the last two decades has shown that such approaches are effective at maintaining accuracy while reducing training set size in many machine learning applications. However, as we begin to deploy active learning in real ongoing learning systems and data annotation projects, we are encountering unexpected problems—due in part to practical realities that violate the basic assumptions of earlier foundational work. I review some of these issues, and discuss recent work being done to address the challenges.
Keywords: Active Learning, Applied Machine Learning, Human-Computer Interaction.

1. Introduction

It is fairly well established now that *active learning*—a family of machine learning methods which may *query* the data instances to be labeled for training by an *oracle* (e.g., a human annotator)—can achieve higher accuracy with fewer labeled examples than passive learning. Historically, most active learning research has focused on mechanisms for (and the benefit of) selecting queries from the learner's perspective. In essence, this body of work addresses the question, "can machines learn with fewer labeled training instances if they are allowed to ask questions?" By and large, the answer to this question is "yes," with encouraging results that have been demonstrated for a variety of problem settings and domains.

For example, query algorithms (sometimes called "utility measures") based on *uncertainty sampling* select query instances which have the least label certainty under the current trained model. This simple approach is no more computationally expensive than performing inference, and has been shown to work well in a variety of applications (e.g., Lewis and Gale, 1994; Tong and Chang, 2001; Tür et al., 2005; Settles and Craven, 2008). Similarly, algorithms based on *query-by-committee* aim to minimize the version space of the model, and satisfying theoretical bounds on label complexity have been established for these and related methods (Freund et al., 1997; Dasgupta, 2004; Hanneke, 2009). For a more detailed overview of active learning algorithms—containing many example references—see Settles (2009). In addition to all these published accounts, consider that software companies and large-scale research projects such as CiteSeer, Google, IBM, Microsoft, and Siemens are in-

creasingly using active learning in the applications they are building[1]. Published results and increased industry adoption seem to indicate that active learning methods have matured to the point of usefulness in many real situations.

However, there are still plenty of open problems when it comes to using active learning in practice. In a recent survey of annotation projects for natural language processing tasks (Tomanek and Olsson, 2009), only 20% of the respondents said they had ever decided to use active learning. The authors even suspect that this is an over-estimate, since it was advertised as a survey on the use of active learning and thus biased towards those familiar with it. Of the large majority who chose *not* to use active learning, 21% were convinced that it would not work well, with some stating that "while they believed [it] would reduce the amount of instances to be annotated, it would probably not reduce the overall annotation time." Furthermore, recent empirical studies—some employing live active learning with real annotators "in the loop"—have found puzzlingly mixed or negative results (Schein and Ungar, 2007; Guo and Schuurmans, 2008; Settles et al., 2008a; Baldridge and Palmer, 2009). Consider also that implementing query selection methods for certain more sophisticated learning algorithms can require significant software engineering overhead. Given the disconnect between the prevailing message in the literature and these mixed results in practice, coupled with high development costs, it is not surprising that researchers are hesitant to use active learning in live and ongoing projects.

I conjecture that the wealth of positive results in the literature (and there are few negative results to go by due to the publication bias) can be accounted for by the many simplifying assumptions made in previous work. For example, we have often assumed that there is a single infallible annotator whose labels can be trusted, or that the cost of labeling each query is uniformly expensive. Most of these assumptions were made to facilitate controlled experiments, where researchers often use gold-standard labeled data but pretend they are unlabeled until queried and labeled by a simulated oracle. In many real-world situations, though, these and other common assumptions do not hold. As a result, the research question has shifted over the last few years to "can machines learn *more economically* if they are allowed to ask questions?" While this seems related, it is a fundamentally different question. This new way of thinking removes emphasis from the learner and merely reducing the size of its training set, and begins to incorporate all aspect of the problem: annotators, costs, label noise, etc. This is a centrally important direction in active learning research, and the focus of this article.

2. Six Practical Challenges

In this section, we will survey six main research directions which address problems for active learning in practice. Each of the subsections that follow describes a common assumption from the literature, and summarizes ongoing research (mostly from the last three or four years) aimed at solving the problems that arise when these assumptions are violated.

1. Based on personal communication with (respectively): C. Lee Giles, David Cohn, Prem Melville, Eric Horvitz, and Balaji Krishnapuram.

2.1. Querying in Batches

In most active learning experiments, queries are selected in *serial* (one at a time), as opposed to *batches* (several to be labeled at once). It is typically assumed that the learner may inspect a large pool of unlabeled data $\mathcal{U}$ and select the single most informative instance to be labeled by the oracle (e.g., a human annotator). This setting is called "pool-based" active learning. Once the query has been labeled and added to the training set $\mathcal{L}$, the learner re-trains using this newly-acquired knowledge and selects another single query, given all the previously labeled instances, and the process repeats.

However, this is not always a realistic setting. For many applications, the process of inducing a model from training data may slow or expensive, which is often the case with state-of-the-art methods like large ensemble algorithms, or the graphical models used for structured-prediction tasks. In such cases, it is an inefficient use of labeling resources (e.g., a human annotator's time) to wait for the model to re-train before querying for the next label. Consider also that for some applications, it may be more natural to acquire labels for many different instances at once. A good example of this is high-throughput biology, where scientists are willing to wait a long time between experiments, but need to acquire measurements (to be used as training labels) for hundreds or thousands of molecules in a single screen. It is also the case that distributed, parallel labeling environments are increasingly available (some examples and additional challenges are discussed in Section 2.2), allowing multiple annotators to work on different queries at different workstations simultaneously over a network.

(a) serial (b) batch-mode

Figure 1: Serial vs. batch-mode active learning. When querying in batches, the instances should be diverse (to avoid redundancy) as well as informative to the learner.

In such settings, we wish to select a set of queries $\mathcal{Q} \subseteq \mathcal{U}$ to be labeled concurrently with model re-training, or in parallel if that is supported by the experiment or annotation environment. A naïve approach to constructing this batch is to simply evaluate all the potential query instances, and select the "Q-best" as ranked by some utility measure. Unfortunately this is a myopic strategy, and generally does not work well since it fails to consider the overlap in information content among all the "best" instances. In other words, the best two queries might be so highly ranked because they are virtually identical, in which case labeling both is probably wasted effort. Instead, the instances in $\mathcal{Q}$ need to be both informative *and* diverse to make the best use of labeling resources.

To accomplish this, a few batch-mode active learning algorithms have been proposed. These approaches fall into three main categories. The first is to explicitly incorporate a density measure, e.g., by ranking the most informative instances, and then clustering those that are most highly ranked (Brinker, 2003; Kim et al., 2006; Xu et al., 2007). Then the

batch can be constructed by querying representative instances (e.g., the centroids) from each cluster. A second approach views the task as a set optimization problem, where the utility function for any batch $\mathcal{Q}$ is the expected joint reduction in uncertainty of the model using Bayesian experimental design techniques (Hoi et al., 2006a,b). While these approaches use greedy heuristics, Hoi et al. (2006b) exploit the properties of a *submodular* functions (Nemhauser et al., 1978) to find a batch that is guaranteed to be near-optimal. A third approach (Guo and Schuurmans, 2008) attempts to construct a batch by treating the pool of candidate instances $\mathcal{U}$ as a bit vector (with 1's corresponding to the elements included in $\mathcal{Q}$), and use gradient methods to approximate the best query-set vector by numerically optimizing a discriminative expected information gain measure.

For the most part, these *batch-mode* approaches have been shown to be more economical (in terms of accuracy vs. the number of labeled batches) than passively selecting instances for a batch, which in turn is generally better than a myopic "Q-best" method. However, on some data sets a passive (random) batch-construction approach can still outperform the active methods (Guo and Schuurmans, 2008). Thus, there is still work to be done in characterizing the cases in which batch-mode active learning is likely to help, and in making further improvements to the state of the art.

2.2. Noisy Oracles

Another strong assumption in most active learning research is that the quality of labeled data from the oracle is high. In theory, of course, an "oracle" is by definition an infallible authority or guide. However, if labels come from an empirical experiment (e.g., in biological, chemical, or clinical studies), then one can usually expect some noise to result from the instrumentation or experimental setting. Even if labels come from human experts, they may not always be reliable: (*i*) some instances are implicitly difficult for both people and machines, and (*ii*) people can become distracted or fatigued over time, which introduces variability in the quality of their annotations. The recent introduction of Internet-based "crowd-sourcing" tools, such as Mechanical Turk[2] and the clever use of online games[3], have enabled researchers to attempt to "average out" some of this noise by cheaply obtaining labels from multiple non-experts. Such approaches have been used to produce gold-standard quality training sets (Snow et al., 2008) and to evaluate learning algorithms on tasks for which no gold-standard labels exist (Mintz et al., 2009; Carlson et al., 2010).

How to use non-experts (or even noisy experts) as oracles in active learning is still a matter of ongoing investigation. One way of thinking about the problem is *agnostic* active learning (Balcan et al., 2006; Dasgupta et al., 2008), a framework which relaxes the assumption that the oracle's labels are trustworthy, yet still has positive theoretical results. Other recent work assumes that a learner may repeat queries to be labeled by multiple annotators. This introduces another interesting research issues. When should the learner decide to query for the (potentially noisy) label of a *new* unlabeled instance, versus querying for repeated labels to de-noise an *existing* but suspicious training instance? How can the learner even decide that the quality of a label is suspect? Sheng et al. (2008) study this problem using several heuristics that take into account estimates of both oracle and model

2. http://www.mturk.com
3. http://www.gwap.com

uncertainty, and show that data can be improved by selective repeated labeling. However, their analysis assumes that (*i*) annotation is a noisy process over some underlying true label (in other words, there must not be any inherently difficult or ambiguous instances from the oracle's perspective), and (*ii*) all annotators are equally and consistently noisy. To my knowledge, no one has addressed the first problem. However, Donmez et al. (2009) address the second issue by allowing annotators to have different levels of accuracy in their annotations, and show that both true instance labels and individual oracle qualities can be estimated, so long as they do not change over time. Donmez et al. (2010) further relax these assumptions to allow for time-varying noise levels among annotators, and adaptively query different labelers based on the current estimate of their labeling quality.

There are still many open research opportunities along these lines. In particular, how might the effect of payment influence annotation quality (i.e., if you pay a non-expert twice as much, are they sufficiently motivated to be more accurate)? What if some instances are inherently ambiguous regardless of which annotator is used, so repeated labeling is not likely to improve matters? In most crowd-sourcing environments, the users are not necessarily available "on demand," thus accurate estimates of annotator quality may be difficult to achieve in the first place, and might possibly never be applicable again since the model has no real choice over which to use. Finally, most work in this area has been based on theoretical results or experimental simulations, and it would be helpful to see verification of these claims in practice.

2.3. Variable Labeling Costs

For many applications, variance shows up not only in label quality from one instance to another, but also in the *cost* of obtaining these labels. If our goal in active learning is to minimize the overall cost of training an accurate model, then simply reducing the number of labeled instances does not necessarily guarantee a reduction in overall labeling cost.

One proposed approach for reducing annotation effort in active learning involves using the current trained model to assist in the annotation of queries by pre-labeling them in structured learning tasks like parsing (Baldridge and Osborne, 2004) or information extraction (Culotta and McCallum, 2005). However, such methods do not actually represent or reason about labeling costs. Instead, they attempt to reduce cost indirectly by minimizing the number of annotation actions required for a query that has already been selected.

2.3.1. Known Labeling Costs

Alternatively, *cost-sensitive active learning* approaches explicitly account for varying labeling costs while selecting queries (usually under the assumption the costs are known). Kapoor et al. (2007) propose a decision-theoretic approach that takes into account both labeling costs and misclassification costs. In this setting, each candidate query is evaluated by summing its labeling cost with the future misclassification costs that are expected to be incurred if the instance were added to the training set. They make the somewhat reasonable assumption that the cost of labeling an instances is a linear function of its length (e.g., \$0.01 per second for voicemail messages). The approach also requires labeling and misclassification costs to be mapped into the same currency (e.g., \$10 per error), which may not be appropriate or straightforward for some applications. King et al. (2004) use a similar

decision-theoretic approach to reduce real labeling costs. They describe a "robot scientist" which can execute a series of autonomous biological experiments to discover metabolic pathways in yeast, with the objective of minimizing the cost of materials used (i.e., the cost of an experiment plus the expected total cost of future experiments until the correct hypothesis is found). Here again, the cost of materials for each experiment is fixed and known to the learner at the time of the experiment selection.

2.3.2. Unknown Labeling Costs

In the settings above, and indeed in much of the cost-sensitive active learning literature (e.g., Margineantu, 2005; Tomanek et al., 2007), the cost of annotating an instance is still assumed to be fixed and known to the learner before querying. Settles et al. (2008a) propose a novel approach to cost-sensitive active learning in settings where annotation costs are variable and *not* known. For example, if cost is a function of annotation time and we do not know in advance how long the annotator or experiment will take. In this approach, one can learn a regression cost-model (alongside the active task-model) which tries to predict the real, unknown annotation cost based on a few simple "meta features" on the instances. An analysis of four data sets using real-world human annotation costs reveals the following (Settles et al., 2008a):

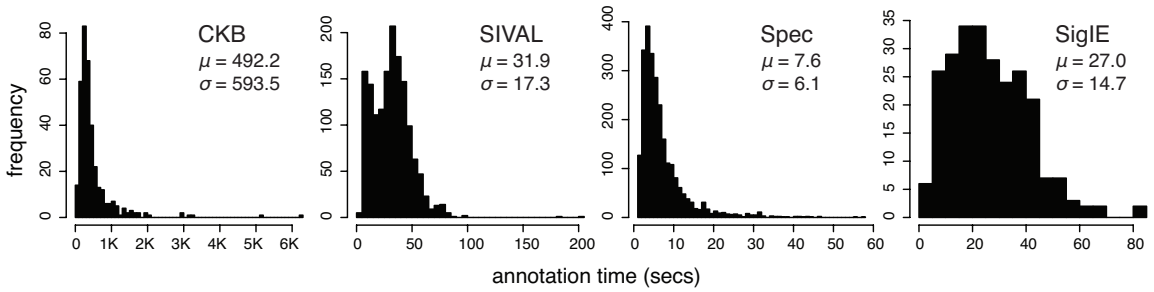


Figure 2: Histograms illustrating the distribution of annotation times for the data sets reported in Settles et al. (2008a). The mean annotation time μ and standard deviation σ for each data set is also reported.

- As shown in Figure 2, annotation costs are not approximately constant across instances, and can vary considerably in some domains. This result is supported by the subsequent findings of others working on different learning tasks (Arora et al., 2009; Vijayanarasimhan and Grauman, 2009a).
- Consequently, active learning approaches which ignore cost may perform no better than random selection (i.e., passive learning).
- As shown in Figure 3, the cost of annotating an instance may not be intrinsic, but may instead vary based on the person doing the annotation. This result is also supported by the findings of Ringger et al. (2008) and Arora et al. (2009).

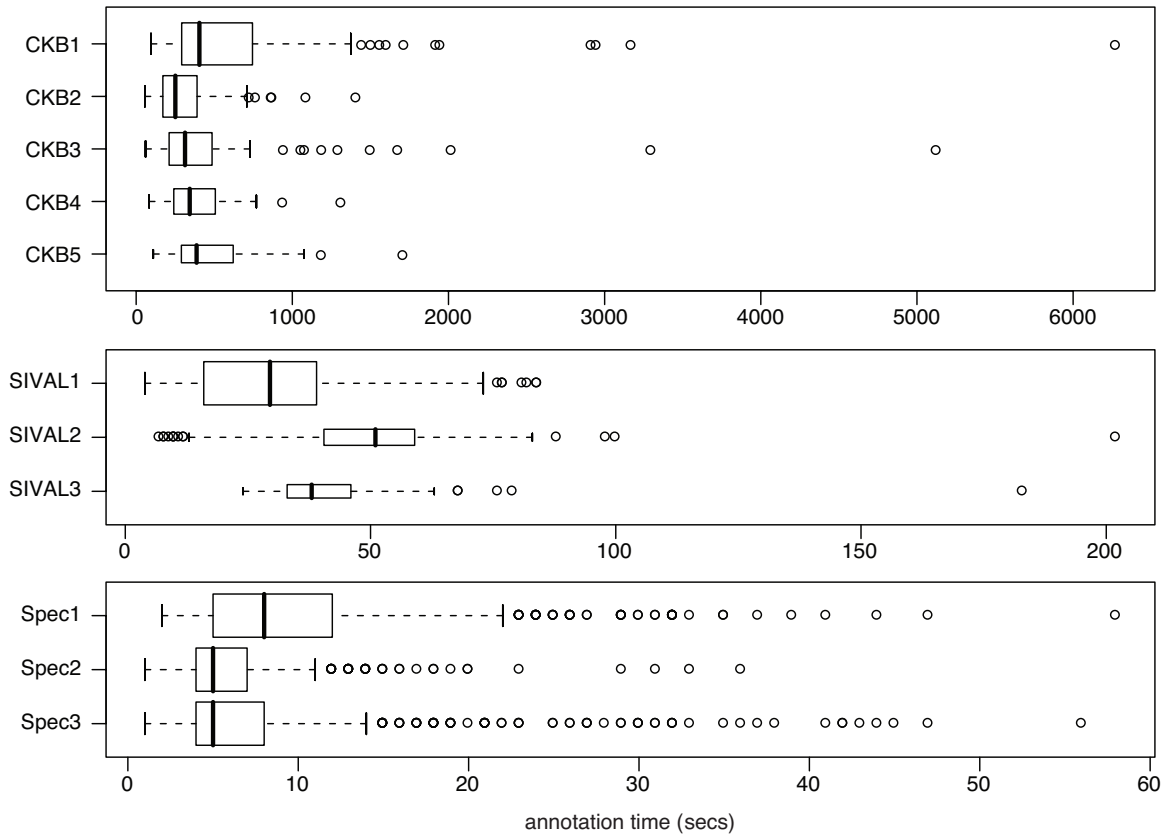


Figure 3: Box plots showing labeling time distributions for different human annotators on several data sets (Settles et al., 2008a). A box represents the middle 50% of annotation times, and the median is marked with a thick black line. Box heights are scaled in proportion to the number of instances labeled. Whiskers on either side span the first and last quartiles of each distribution; circles indicate possible outliers. Note that the range of the horizontal axis varies across data sets.

- The measured cost for an annotation may include stochastic components. In particular, there are at least two types of noise which affect annotation speed: *jitter* (minor variations due to annotator fatigue, latency, etc.) and *pause* (major variations due to interruptions, that should be shorter under normal circumstances).

- Unknown annotation costs can *sometimes* be accurately predicted, even after seeing only a few training instances. This result is also supported by the findings of Vijayanarasimhan and Grauman (2009a). Moreover, these learned cost-models are significantly better than simpler cost heuristics (e.g., a linear function of length).

While empirical experiments show that learned cost-models can be trained to predict annotation times fairly well, further work is warranted to determine how such approximate, predicted labeling costs can be utilized effectively by cost-sensitive active learning systems.

Settles et al. experimented with a simple heuristic that divides the utility measure (e.g., entropy-based uncertainty sampling) by the predicted cost of the instances, but found that, despite fairly good cost predictions, this did not produce better learning curves in multiple natural language tasks when compared to random sampling (In fact, this was sometimes the case when *true* costs are known). Such degradation suggests that uncertainty and cost are correlated, but further investigation is needed. On the other hand, results from Haertel et al. (2008) suggest that this heuristic, which they call *return on investment* (ROI), can be effective for part-of-speech tagging, although they use a fixed heuristic cost model rather than a dynamic one trained in parallel with the task model. Vijayanarasimhan and Grauman (2009a) demonstrated potential cost savings with active learning using predicted annotation costs for computer vision. It is unclear whether these disparities are intrinsic, task-specific, or simply a result of differing experimental settings.

Even among methods that do not explicitly reason about annotation cost, several authors have found that alternative query types (such as labeling features rather than instances, see Section 2.4) can lead to reduced annotation costs for human oracles (Raghavan et al., 2006; Druck et al., 2009; Vijayanarasimhan and Grauman, 2009a). Interestingly, Baldridge and Palmer (2009) used active learning for morpheme annotation in a rare-language documentation study, using two live human oracles (one expert and one novice) interactively "in the loop." They found that the most cost-saving strategy differed between the two annotators, in terms of reducing both labeled corpus size and annotation time. The domain expert was a more efficient oracle with an uncertainty-based active learner, but semi-automated annotations—intended to assist in the labeling process—were of little help. The novice, however, was more efficient with a passive learner (selecting passages at random), but semi-automated annotations were in this case beneficial. There is also some preliminary evidence that for complex annotation tasks, the design of the user interface can have as much or more impact on reducing costs as the active learning strategy (Druck et al., 2009). Continued work along these lines could also prove to be beneficial.

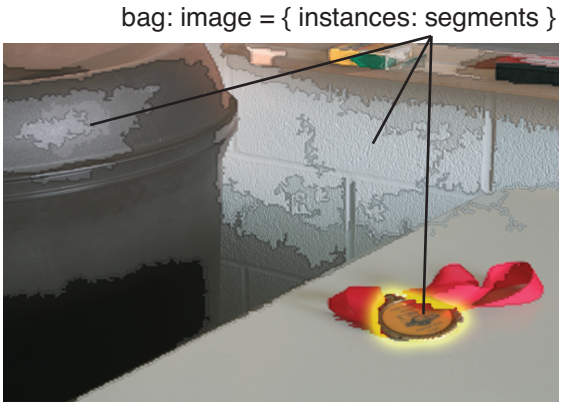
2.4. Alternative Query Types

Most work in active learning assumes that a "query unit" is of the same type as the target concept to be learned. In other words, if the task is to assign class labels to text documents, the learner must query a document and the oracle provides its label. While Angluin (1988) outlines several potential query types in a theoretical analysis of active learning, only the commonly-used *membership query* has been deemed appropriate for most real-world applications. Nevertheless, recent advances have considered other types of queries for learning scenarios that can support them.

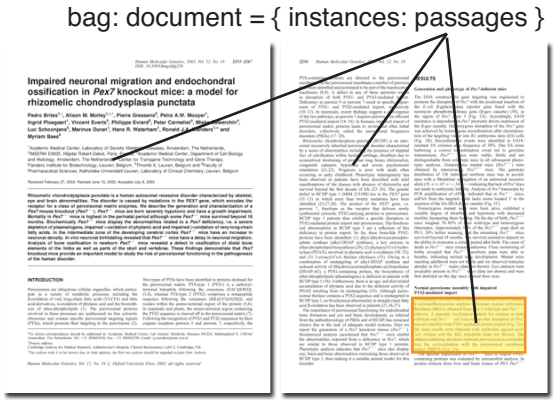
2.4.1. Multiple-Instance Active Learning

Settles et al. (2008b) introduce an alternative querying scenario called *multiple-instance active learning*, which allows the learner to query for labels at various levels of granularity. In the multiple-instance (MI) learning framework, instances are grouped into *bags* (i.e., multi-sets), and it is the bags—rather than instances—that are labeled for training. A bag is labeled negative if and only if all of its instances are negative. A bag is labeled positive, however, if at least one of its instances is positive (note that positive bags may also contain

negative instances). A naïve approach to MI learning is to view it as supervised learning with one-sided noise (i.e., all negative instances are truly negative, but some positives are actually negative). However, special MI learning algorithms have been developed to learn from labeled bags despite this ambiguity. The MI setting was formalized by Dietterich et al. (1997) in the context of drug activity prediction, and has since been applied to a wide variety of tasks including content-based image retrieval (Maron and Lozano-Perez, 1998; Andrews et al., 2003; Rahmani and Goldman, 2006) and text classification (Andrews et al., 2003; Ray and Craven, 2005).



(a) content-based image retrieval



(b) text classification

Figure 4: Multiple-instance active learning. In content-based image retrieval, images are represented as bags and instances correspond to segmented regions. An MI active learner may query which segments belong to the object of interest, such as the gold medal shown in this image. In text classification, documents are bags and the instances represent passages of text. In MI active learning, the learner may query specific passages to determine if they belong to the positive class.

Figure 4 illustrates how the MI representation can be applied to (a) content-based image retrieval (CBIR) and to (b) text classification. For the CBIR task, images are represented as bags and instances correspond to segmented regions of the image. A bag representing

a given image is labeled positive if the image contains some object of interest. The MI paradigm is well-suited to this task because only a few regions of an image may represent the object of interest, such as the gold medal in Figure 4(*a*). An advantage of the MI representation here is that it is significantly easier to label an entire image than it is to label each segment, or even a subset of the image segments. For the text classification task, documents can be represented as bags and instances correspond to short passages (e.g., paragraphs) that comprise each document. The MI representation is compelling for classification tasks for which document labels are freely available or cheaply obtained (e.g., from hyperlinks, indexes, or databases on the Internet), but the target concept is represented by only a few passages.

A traditional active learning approach for these tasks would be to query bags (i.e., images or documents) because that is the unit of classification. For MI learning tasks such as these, however, it is possible to obtain labels both at the bag level and directly at the instance level. Fully labeling all instances is expensive; often the rationale for formulating the learning task as an MI problem is that it allows us to take advantage of coarse labelings that may be available at low cost (or even for free). Fortunately, in MI active learning the learner may selectively query for only the *informative* labels at a finer granularity, e.g., salient passages rather than entire documents, or segmented image regions rather than entire images. Settles et al. (2008b) focus on this type of mixed-granularity active learning with a multiple-instance generalization of logistic regression, and show that it is helpful to incorporate the MI bias directly into the query selection strategy. Vijayanarasimhan and Grauman (2009a,b) have extended the idea to SVMs for the image retrieval task, and also explore an approach that interleaves queries at varying levels of granularity and cost.

2.4.2. Querying Features

Another alternative setting is to query on *features* rather than (or in addition to) instances. Raghavan et al. (2006) was the first to explore this idea with an approach called *tandem learning*, which incorporates feature feedback into traditional classification problems. In their work, a text classifier may interleave typical instance-label queries with feature-salience queries (e.g., "is the word *puck* a discriminative feature if hockey is one of the class labels?"). The values for salient features are then artificially amplified in instance feature vectors to reflect their relative importance. The authors reported that interleaving such queries is very effective for text classification, and also found that words (the features in this case) are often much easier for human annotators to label in empirical user studies, requiring a fifth of the time. Note, however, that answers to these feature queries only imply their discriminative value and do not tie features to class labels directly.

In recent years, several new methods have been developed for incorporating feature-based domain knowledge into supervised and semi-supervised learning (e.g., Haghighi and Klein, 2006; Druck et al., 2008; Melville et al., 2009). In this line of work, users may supply domain knowledge in the form of feature-label constraints, e.g., "the word *puck* indicates class label hockey." Mann and McCallum (2010) describe a semi-supervised method of combining such constraints with unlabeled data in exponential models, and Melville et al. (2009) combine this domain knowledge with labeled examples for naïve Bayes by pooling multinomials. When combined with labeled data instances, this is sometimes called *dual supervision*. Interestingly, Mann and McCallum determined that specifying many

imprecisely-estimated constraints is generally more effective than using a few more precise ones, suggesting that human-specified feature labels (however noisy) are useful if there are enough of them. This begs the question of how to *actively* solicit such feature-based domain knowledge.

Druck et al. (2009) propose and evaluate a variety of active query strategies aimed at gathering useful feature-label constraints for two information extraction tasks. They show that active feature labeling is more effective than either "passive" feature labeling (using a variety of strong baselines) or instance-labeling (both passive and active) for two information extraction tasks. These results held true for both simulated and interactive human-annotator experiments. Liang et al. (2009) present a more principled approach to the problem grounded in Bayesian experimental design, however, they also resort to heuristics in practice due to intractability. Melville and Sindhwani (2009) have explored interleaving instance and feature label queries for sentiment classification in blogs using the pooling multinomials naïve Bayes approach, and Sindhwani et al. (2009) consider a similar query setting for a semi-supervised graph/kernel-based text classifier.

2.5. Multi-Task Active Learning

Most active learning settings assume that there is only one learner trying to solve a single task. In many real-world problems, however, the same data instances may be labeled in multiple ways for different subtasks. In such cases, it is probably more economical to label a single instance for all subtasks simultaneously. Therefore, *multi-task active learning* algorithms assume that a single query will be labeled for multiple tasks, and attempt to assess the "informativeness" of an instance with respect to all the learners involved.

Consider a database of film reviews, which might be used to build a system that (*i*) extracts the names of key actors and production crew, (*ii*) classifies the film by genre, and (*iii*) predicts a five-star rating based on the text. Such a system would probably employ three independent learners: a sequence model for entity extraction, a classifier for genres, and a regression model to predict ratings. Effectively selecting queries that benefit all three of these learners is still an open and promising direction in active learning.

Along these lines, Reichart et al. (2008) study a two-task active learning scenario for natural language parsing and named entity recognition (NER), a form of information extraction. They propose two methods for actively learning both tasks simultaneously. The first is *alternating selection*, which allows the parser to query sentences in one iteration, and then the NER system to query instances in the next. The second is *rank combination*, in which both learners rank the query candidates in the pool by expected utility, and the instances with the highest combined rank are selected for labeling. In both cases, uncertainty sampling is used as the base selection strategy for each learner, but other frameworks could be used as well. As one might expect, these methods outperform passive learning for both subtasks, while learning curves for each individual subtask are not as good as they would have been in a single-task active learning setting.

Qi et al. (2008) study a different multi-task active learning scenario, in which images may be labeled for several binary classification tasks in parallel. For example, an image might be labeled as containing a beach, sunset, mountain, field, etc., which are not all mutually exclusive; however, they are not entirely independent, either. The beach and sunset labels

may be highly correlated in the data, for example, so a simple rank combination might over-estimate the informativeness of some instances. They propose and evaluate a new approach which takes into account the mutual information among labels.

2.6. Changing (or Unknown) Model Classes

An important side-effect of active learning that the resulting labeled training set $\mathcal{L}$ is not an i.i.d. sample of the data, but is rather a biased distribution which is implicitly tied to the model used in selecting the queries. Most work in active learning has assumed that the appropriate model class for the task is already known, so this is not generally a problem. However, it can become problematic if we wish to re-use the training data with a model of a different type—which is common when the state of the art advances—or if we do not even know the appropriate model class (or feature set) for the task to begin with.

Fortunately, this change of or uncertainty about the model is not always an issue. Lewis and Catlett (1994) showed that decision tree classifiers can still benefit significantly from a training set constructed by an active naïve Bayes learner using uncertainty sampling. Tomanek et al. (2007) also showed that information extraction data gathered by a maximum entropy model using the query-by-committee algorithm can be effectively re-used to train more sophisticated conditional random fields (CRFs), maintaining cost savings compared with random sampling. Hwa (2001) successfully re-used natural language parsing data queried by one type of parser to train other types of parsers.

However, Baldridge and Osborne (2004) reported the exact opposite problem when re-using data queried by one parsing model to train a variety of other parsers. As an alternative, they perform active learning using a heterogeneous ensemble composed of different parser types, and also use semi-automated labeling to cut down on human annotation effort. This approach helped to reduce the number of training examples required for each parser type compared with passive learning. Similarly, Lu et al. (2010) employed active learning with a heterogeneous ensemble of neural networks and decision trees, when the more appropriate model class for the learning task was not known in advance. Their ensemble approach was able to simultaneously select informative instances for the overall ensemble, and bias the distribution of constituent weak learners toward the most appropriate model as more training data was gathered. Sugiyama and Rubens (2008) have experimented with an ensemble of linear regression models that used differing feature sets, to study cases in which the appropriate feature set is not yet decided upon.

This brings up a very important point for active learning in practice. If the appropriate model class and feature set happen to be known in advance—or if these are not likely to change much in the future—then active learning can probably be safely used. Otherwise, random sampling (at least for pilot studies, until the task can be better understood) may be more advisable than taking one's chances on active learning with the "wrong" learning algorithm. A viable alternative for active learning seems to be the use of heterogeneous ensembles in selecting queries, but there is still much work to be done in this direction.

3. Conclusion

This article surveys the main challenges currently facing the use of active learning in practice. While many of these issues are nontrivial and well beyond the current state of the

art, I am optimistic that the research community will find pragmatic solutions that are of general use. After all, we have overcome comparable challenges in the past.

Over two decades ago, some exciting theoretical results for active learning (Sammut and Banerji, 1986; Angluin, 1988) led to a body of work applying these early ideas in neural networks. For the most part, these methods assumed that the learner may synthesize arbitrary query instances de novo, and applications studied only simple or artificial learning tasks (e.g., geometrical shapes on a 2D plane). In an interesting early attempt at a real-world task, Lang and Baum (1992) employed active learning with a human oracle to train a classifier for handwritten characters and digits. They encountered an unexpected problem: many of the query images generated by the learner contained no recognizable symbols, only artificial hybrid characters that had no semantic meaning. This negative result did not discourage progress, however, but helped to motivate and justify the selective sampling and pool-based active learning scenarios commonly used today, since they guarantee that query instances are sensible because come from an underlying natural distribution.

I think we find ourselves in a similar situation today. While the past few decades have established active learning as a widely applicable tool for a variety of problem domains, these results are subject to assumptions which focus on the utility of a query to the learner, and not its cost to the teachers or other aspects of the problem as a whole. Rather than quell progress, though, I believe these practical challenges are leading to innovations which draw us closer to methods for effective interactive learning systems.

References

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 15, pages 561–568. MIT Press, 2003.

D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

S. Arora, E. Nyberg, and C.P. Rosé. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 18–26. ACL Press, 2009.

M.F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 65–72. ACM Press, 2006.

J. Baldridge and M. Osborne. Active learning and the total cost of annotation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9–16. ACL Press, 2004.

J. Baldridge and A. Palmer. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 296–305. ACL Press, 2009.

K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 59–66. AAAI Press, 2003.

A. Carlson, J. Betteridge, R. Wang, E.R. Hruschka Jr, and T. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*. ACM Press, 2010.

A. Culotta and A. McCallum. Reducing labeling effort for stuctured prediction tasks. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 746–751. AAAI Press, 2005.

S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, pages 337–344. MIT Press, 2004.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 353–360. MIT Press, 2008.

T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

P. Donmez, J. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268. ACM Press, 2009.

P. Donmez, J. Carbonell, and J. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the SIAM Conference on Data Mining (SDM)*, 2010.

G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602. ACM Press, 2008.

G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 81–90. ACL Press, 2009.

Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective samping using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems (NIPS)*, number 20, pages 593–600. MIT Press, Cambridge, MA, 2008.

R. Haertel, K. Seppi, E. Ringger, and J. Carroll. Return on investment for active learning. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008.

A. Haghighi and D. Klein. Prototype-driven learning for sequence models. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 320–327. ACL Press, 2006.

S. Hanneke. *Theoretical Foundations of Active Learning.* PhD thesis, Carnegie Mellon University, 2009.

S.C.H. Hoi, R. Jin, and M.R. Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the International Conference on the World Wide Web*, pages 633–642. ACM Press, 2006a.

S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 417–424. ACM Press, 2006b.

R. Hwa. On minimizing training corpus for parser acquisition. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 1–6. ACL Press, 2001.

A. Kapoor, E. Horvitz, and S. Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning,. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 877–882. AAAI Press, 2007.

S. Kim, Y. Song, K. Kim, J.W. Cha, and G.G. Lee. MMR-based active machine learning for bio named entity recognition. In *Proceedings of Human Language Technology and the North American Association for Computational Linguistics (HLT-NAACL)*, pages 69–72. ACL Press, 2006.

R.D. King, K.E. Whelan, F.M. Jones, P.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–52, 2004.

K. Lang and E. Baum. Query learning can work poorly when a human oracle is used. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 335–340. IEEE Press, 1992.

D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 148–156. Morgan Kaufmann, 1994.

D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. ACM/Springer, 1994.

P. Liang, M.I. Jordan, and D. Klein. Learning from measurements in exponential families. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 641–648. ACM Press, 2009.

Z. Lu, X. Wu, and J. Bongard. Adaptive informative sampling for active learning. In *Proceedings of SIAM Conference on Data Mining (SDM)*, 2010.

G.S. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984, 2010.

D. Margineantu. Active cost-sensitive learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1622–1623. AAAI Press, 2005.

O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, pages 570–576. MIT Press, 1998.

P. Melville and V. Sindhwani. Active dual supervision: Reducing the cost of annotating examples and features. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 49–57. ACL Press, 2009.

P. Melville, W. Gryc, and R.D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1275–1284. ACM Press, 2009.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2009.

G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

G.J. Qi, X.S. Hua, Y. Rui, J. Tang, and H.J. Zhang. Two-dimensional active learning for image classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on both features and instances. *Journal of Machine Learning Research*, 7:1655–1686, 2006.

R. Rahmani and S.A. Goldman. MISSL: Multiple-instance semi-supervised learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 705–712. ACM Press, 2006.

S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 697–704. ACM Press, 2005.

R. Reichart, K. Tomanek, U. Hahn, and A Rappoport. Multi-task active learning for linguistic annotations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 861–869. ACL Press, 2008.

E. Ringger, M. Carmen, R. Haertel, K. Seppi, D. Lonsdale, P. McClanahan, J. Carroll, and N. Ellison. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2008.

C. Sammut and R. Banerji. Learning concepts by asking questions. In *Machine Learning: An Artificial Intelligence Approach*, volume 2. Morgan Kaufmann, 1986.

A.I. Schein and L.H. Ungar. Active learning for logistic regression: An evaluation. *Machine Learning*, 68(3):235–265, 2007.

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

B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1069–1078. ACL Press, 2008.

B. Settles, M. Craven, and L. Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10, 2008a.

B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 1289–1296. MIT Press, 2008b.

V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, 2008.

V. Sindhwani, P. Melville, and R.D. Lawrence. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 953–960. ACM Press, 2009.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263. ACM Press, 2008.

M. Sugiyama and N. Rubens. Active learning with model selection in linear regression. In *Proceedings of the SIAM International Conference on Data Mining*, pages 518–529. SIAM, 2008.

K. Tomanek and F. Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 45–48. ACL Press, 2009.

K. Tomanek, J. Wermter, and U. Hahn. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 486–495. ACL Press, 2007.

S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pages 107–118. ACM Press, 2001.

G. Tür, D. Hakkani-Tür, and R.E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, 2005.

S. Vijayanarasimhan and K. Grauman. What's it going to cost you? Predicting effort vs. informativeness for multi-label image annotations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Press, 2009a.

S. Vijayanarasimhan and K. Grauman. Multi-level active prediction of useful image annotations for recognition. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1705–1712. MIT Press, 2009b.

Z. Xu, R. Akella, and Y. Zhang. Incorporating diversity and density in active learning for relevance feedback. In *Proceedings of the European Conference on IR Research (ECIR)*, pages 246–257. Springer-Verlag, 2007.