Active Learning Literature Survey

Burr Settles

Computer Sciences Technical Report 1648
University of Wisconsin–Madison
Updated on: January 26, 2010

Abstract

The key idea behind *active learning* is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. An active learner may pose *queries*, usually in the form of unlabeled data instances to be labeled by an *oracle* (e.g., a human annotator). Active learning is well-motivated in many modern machine learning problems, where unlabeled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain.

This report provides a general introduction to active learning and a survey of the literature. This includes a discussion of the scenarios in which queries can be formulated, and an overview of the query strategy frameworks proposed in the literature to date. An analysis of the empirical and theoretical evidence for successful active learning, a summary of problem setting variants and practical issues, and a discussion of related topics in machine learning research are also presented.

Contents

1 Introduction

This report provides a general review of the literature on active learning. There have been a host of algorithms and applications for learning with queries over the years, and this document is an attempt to distill the core ideas, methods, and applications that have been considered by the machine learning community. To make this survey more useful in the long term, an online version will be updated and maintained indefinitely at:

> http://active-learning.net/

When referring to this document, I recommend using the following citation:

> Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

An appropriate BibTeX entry is:

```
@techreport{settles.tr09,
    Author = {Burr Settles},
    Institution = {University of Wisconsin--Madison},
    Number = {1648},
    Title = {Active Learning Literature Survey},
    Type = {Computer Sciences Technical Report},
    Year = {2009},
}
```

This document is written for a machine learning audience, and assumes the reader has a working knowledge of supervised learning algorithms (particularly statistical methods). For a good introduction to general machine learning, I recommend Mitchell (1997) or Duda et al. (2001). I have strived to make this review as comprehensive as possible, but it is by no means complete. My own research deals primarily with applications in natural language processing and bioinformatics, thus much of the empirical active learning work I am familiar with is in these areas. Active learning (like so many subfields in computer science) is rapidly growing and evolving in a myriad of directions, so it is difficult for one person to provide an exhaustive summary. I apologize for any oversights or inaccuracies, and encourage interested readers to submit additions, comments, and corrections to me at: bsettles@cs.cmu.edu.

1.1 What is Active Learning?

Active learning (sometimes called "query learning" or "optimal experimental design" in the statistics literature) is a subfield of machine learning and, more generally, artificial intelligence. The key hypothesis is that if the learning algorithm is allowed to choose the data from which it learns—to be "curious," if you will—it will perform better with less training. Why is this a desirable property for learning algorithms to have? Consider that, for any supervised learning system to perform well, it must often be trained on hundreds (even thousands) of labeled instances. Sometimes these labels come at little or no cost, such as the the "spam" flag you mark on unwanted email messages, or the five-star rating you might give to films on a social networking website. Learning systems use these flags and ratings to better filter your junk email and suggest movies you might enjoy. In these cases you provide such labels for free, but for many other more sophisticated supervised 复杂的 learning tasks, labeled instances are very difficult, time-consuming, or expensive to obtain. Here are a few examples:

- *Speech recognition*. Accurate labeling of speech utterances is extremely time consuming and requires trained linguists. Zhu (2005a) reports that annotation at the word level can take ten times longer than the actual audio (e.g., one minute of speech takes ten minutes to label), and annotating phonemes can take 400 times as long (e.g., nearly seven hours). The problem is compounded for rare languages or dialects. 为稀有语言或方言合成的。

- *Information extraction*. Good information extraction systems must be trained using labeled documents with detailed annotations. Users highlight entities or relations of interest in text, such as person and organization names, or whether a person works for a particular organization. Locating entities and relations can take a half-hour or more for even simple newswire stories (Settles et al., 2008a). Annotations for other knowledge domains may require additional expertise, e.g., annotating gene and disease mentions for biomedical information extraction usually requires PhD-level biologists.

- *Classification and filtering*. Learning to classify documents (e.g., articles or web pages) or any other kind of media (e.g., image, audio, and video files) requires that users label each document or media file with particular labels, like "relevant" or "not relevant." Having to annotate thousands of these instances can be tedious and even redundant.

Active learning systems attempt to overcome the labeling bottleneck by asking *queries* in the form of unlabeled instances to be labeled by an *oracle* (e.g., a human annotator). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. Active learning is well-motivated in many modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain. Note that this kind of active learning is related in spirit, though not to be confused, with the family of instructional techniques by the same name in the education literature (Bonwell and Eison, 1991).

1.2 Active Learning Examples

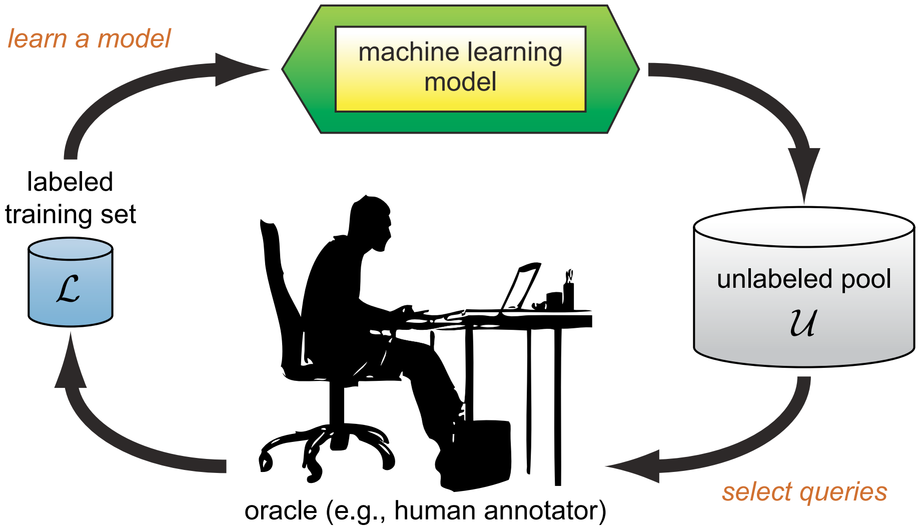
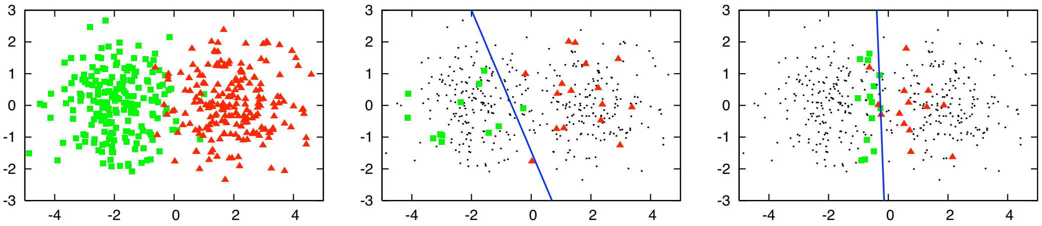


Figure 1: The pool-based active learning cycle.

There are several scenarios in which active learners may pose queries, and there are also several different query strategies that have been used to decide which instances are most informative. In this section, I present two illustrative examples in the *pool-based* active learning setting (in which queries are selected from a large pool of unlabeled instances $\mathcal{U}$) using an *uncertainty sampling* query strategy (which selects the instance in the pool about which the model is least certain how to label). Sections 2 and 3 describe all the active learning scenarios and query strategy frameworks in more detail.



(a) (b) (c)

Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (70% accuracy). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (90%).

Figure 1 illustrates the pool-based *active learning cycle*. A learner may begin with a small number of instances in the labeled training set $\mathcal{L}$, request labels for one or more carefully selected instances, learn from the query results, and then leverage its new knowledge to choose which instances to query next. Once a query has been made, there are usually no additional assumptions on the part of the learning algorithm. The new labeled instance is simply added to the labeled set $\mathcal{L}$, and the learner proceeds from there in a standard supervised way. There are a few exceptions to this, such as when the learner is allowed to make alternative types of queries (Section 6.4), or when active learning is combined with semi-supervised learning (Section 7.1).

Figure 2 shows the potential of active learning in a way that is easy to visualize. This is a toy data set generated from two Gaussians centered at (-2,0) and (2,0) with standard deviation $\sigma = 1$, each representing a different class distribution. Figure 2(a) shows the resulting data set after 400 instances are sampled (200 from each class); instances are represented as points in a 2D feature space. In a real-world setting these instances may be available, but their labels usually are not. Figure 2(b) illustrates the traditional supervised learning approach after randomly selecting 30 instances for labeling, drawn i.i.d. from the unlabeled pool $\mathcal{U}$. The line shows the linear decision boundary of a logistic regression model (i.e., where the posterior equals 0.5) trained using these 30 points. Notice that most of the labeled instances in this training set are far from zero on the horizontal

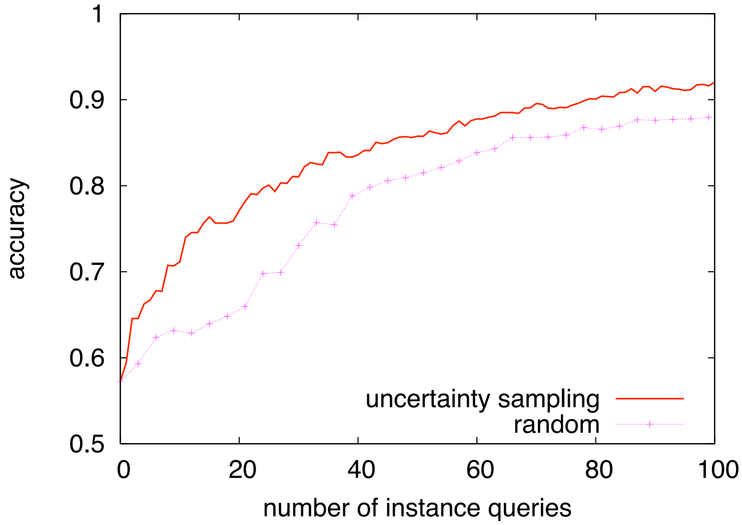


Figure 3: Learning curves for text classification: baseball vs. hockey. Curves plot classification accuracy as a function of the number of documents queried for two selection strategies: uncertainty sampling (active learning) and random sampling (passive learning). We can see that the active learning approach is superior here because its learning curve dominates that of random sampling.

axis, which is where the Bayes optimal decision boundary should probably be. As a result, this classifier only achieves 70% accuracy on the remaining unlabeled points. Figure 2(c), however, tells a different story. The active learner uses uncertainty sampling to focus on instances closest to its decision boundary, assuming it can adequately explain those in other parts of the input space characterized by $\mathcal{U}$. As a result, it avoids requesting labels for redundant or irrelevant instances, and achieves 90% accuracy with a mere 30 labeled instances.

Now let us consider active learning for a real-world learning task: text classification. In this example, a learner must distinguish between baseball and hockey documents from the 20 Newsgroups corpus (Lang, 1995), which consists of 2,000 Usenet documents evenly divided between the two classes. Active learning algorithms are generally evaluated by constructing *learning curves*, which plot the evaluation measure of interest (e.g., accuracy) as a function of the number of new instance queries that are labeled and added to $\mathcal{L}$. Figure 3 presents learning curves for the first 100 instances labeled using uncertainty sampling and random sampling. The reported results are for a logistic regression model averaged over ten folds using cross-validation. After labeling 30 new instances, the accuracy of uncertainty sampling is 81%, while the random baseline is only 73%. As can be seen, the active learning curve dominates the baseline curve for all of the points shown in this figure. We can conclude that an active learning algorithm is superior to some other approach (e.g., a random baseline like traditional passive supervised learning) if it dominates the other for most or all of the points along their learning curves.

1.3 Further Reading

This is the first large-scale survey of the active learning literature. One way to view this document is as a heavily annotated bibliography of the field, and the citations within a particular section or subsection of interest serve as good starting points for further investigation. There have also been a few PhD theses over the years dedicated to active learning, with rich related work sections. In fact, this report originated as a chapter in my PhD thesis (Settles, 2008), which focuses on active learning with structured instances and potentially varied annotation costs. Also of interest may be the related work chapters of Tong (2001), which considers active learning for support vector machines and Bayesian networks, Monteleoni (2006), which considers more theoretical aspects of active learning for classification, and Olsson (2008), which focuses on active learning for named entity recognition (a type of information extraction). Fredrick Olsson has also written a survey of active learning specifically within the scope of the natural language processing (NLP) literature (Olsson, 2009).

2 Scenarios

There are several different problem scenarios in which the learner may be able to ask queries. The three main settings that have been considered in the literature are (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based sampling. Figure 4 illustrates the differences among these three scenarios, which are explained in more detail in this section. Note that all these scenarios (and the lion's share of active learning work to date) assume that queries take the form of unlabeled instances to be labeled by the oracle. Sections 6 and 5 discuss some alternatives to this setting.

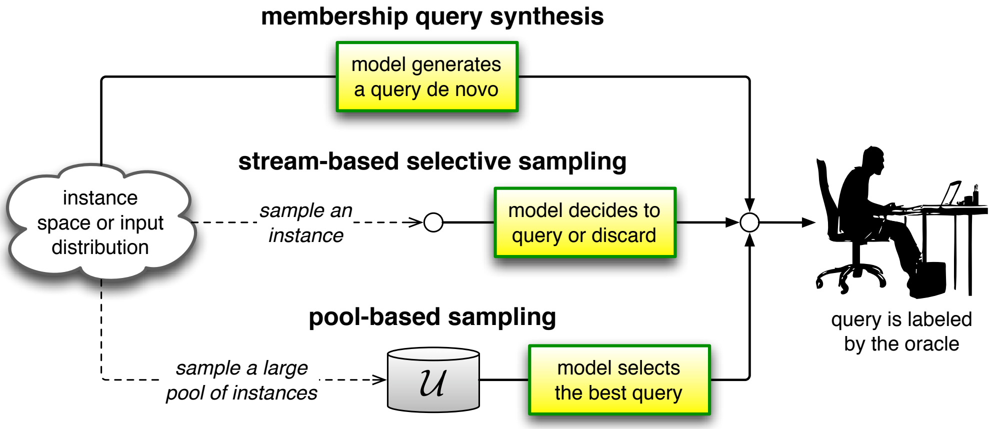

membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
pool-based sampling
sample a large pool of instances
$\mathcal{U}$
model selects the best query
query is labeled by the oracle

Figure 4: Diagram illustrating the three main active learning scenarios.

2.1 Membership Query Synthesis

One of the first active learning scenarios to be investigated is learning with *membership queries* (Angluin, 1988). In this setting, the learner may request labels for any unlabeled instance in the input space, including (and typically assuming) queries that the learner generates de novo, rather than those sampled from some underlying natural distribution. Efficient query synthesis is often tractable and 易处理的 efficient for finite problem domains (Angluin, 2001). The idea of synthesizing queries has also been extended to regression learning tasks, such as learning to predict the absolute coordinates of a robot hand given the joint angles of its mechanical arm as inputs (Cohn et al., 1996).

Query synthesis is reasonable for many problems, but labeling such arbitrary 任意的 instances can be awkward if the oracle is a human annotator. For example, Lang and Baum (1992) employed membership query learning with human oracles to train a neural network to classify handwritten characters. 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 natural semantic meaning. Similarly, one could imagine that membership queries for natural language processing tasks might create streams of text or speech that amount to gibberish. The stream-based and pool-based scenarios (described in the next sections) have been proposed to address these limitations.

However, King et al. (2004, 2009) describe an innovative and promising real-world application of the membership query scenario. They employ a "robot scien-

机器设计实验，验证结果，主动学习

tist” which can execute a series of autonomous biological experiments to discover metabolic pathways in the yeast *Saccharomyces cerevisiae*. Here, an instance is a mixture of chemical solutions that constitute a growth medium, as well as a particular yeast mutant. A label, then, is whether or not the mutant thrived in the growth medium. All experiments are autonomously synthesized using an active learning approach based on inductive logic programming, and physically performed using a laboratory robot. This active method results in a three-fold decrease in the cost of experimental materials compared to naïvely running the least expensive experiment, and a 100-fold decrease in cost compared to randomly generated experiments. In domains where labels come not from human annotators, but from experiments such as this, query synthesis may be a promising direction for automated scientific discovery.

它可以执行一系列自主生物实验，以发现酵母的代谢途径。这里，实例是a

构成生长介质的化学溶液的混合物，以及特定的酵母突变体。一个标签，那么，是突变体是否在生长培养基中茁壮成长。所有实验都是使用基于归纳逻辑编程的主动学习方法自主合成的，并使用实验室机器人进行物理操作。与运行最便宜的实验na ï vely相比，这种主动方法的结果是实验材料成本降低了三倍，与随机生成的实验相比，成本降低了100倍。在标签不是来自人类注释者，而是来自这样的实验的领域，查询合成可能是auto的一个有前途的方向

2.2 Stream-Based Selective Sampling

An alternative to synthesizing queries is *selective sampling* (Cohn et al., 1990, 1994). The key assumption is that obtaining an unlabeled instance is free (or inexpensive), so it can first be sampled from the actual distribution, and then the learner can decide whether or not to request its label. This approach is sometimes called *stream-based* or *sequential* active learning, as each unlabeled instance is typically drawn one at a time from the data source, and the learner must decide whether to query or discard it. If the input distribution is uniform, selective sampling may well behave like membership query learning. However, if the distribution is non-uniform and (more importantly) unknown, we are guaranteed that queries will still be sensible, since they come from a real underlying distribution.

先根据真实分布采样

基于流或顺序的主动学习

The decision whether or not to query an instance can be framed several ways. One approach is to evaluate samples using some “informativeness measure” or “query strategy” (see Section 3 for examples) and make a biased random decision, such that more informative instances are more likely to be queried (Dagan and Engelson, 1995). Another approach is to compute an explicit *region of uncertainty* (Cohn et al., 1994), i.e., the part of the instance space that is still ambiguous to the learner, and only query instances that fall within it. A naïve way of doing this is to set a minimum threshold on an informativeness measure which defines the region. Instances whose evaluation is above this threshold are then queried. Another more principled approach is to define the region that is still unknown to the overall model class, i.e., to the set of hypotheses consistent with the current labeled training set called the *version space* (Mitchell, 1982). In other words, if any two models of the same model class (but different parameter settings) agree on all

明确的 不确定区域

需要更多实体来训练以减少不确定性

与现有标签训练结果一致的假设集，称之为版本空间

the labeled data, but disagree on some unlabeled instance, then that instance lies within the region of uncertainty. Calculating this region completely and explicitly is computationally expensive, however, and it must be maintained after each new query. As a result, approximations are used in practice (Seung et al., 1992; Cohn et al., 1994; Dasgupta et al., 2008).

The stream-based scenario has been studied in several real-world tasks, including part-of-speech tagging (Dagan and Engelson, 1995), sensor scheduling (Krishnamurthy, 2002), and learning ranking functions for information retrieval (Yu, 2005). Fujii et al. (1998) employ selective sampling for active learning in word sense disambiguation, e.g., determining if the word “bank” means land alongside a river or a financial institution in a given context (only they study Japanese words in their work). The approach not only reduces annotation effort, but also limits the size of the database used in nearest-neighbor learning, which in turn expedites the classification algorithm.

2005)。Fujii等人(1998)在词义消歧中采用选择性抽样进行主动学习，例如，确定单词“ bank ”在给定的上下文中是指河边的土地还是指金融机构(只有他们在工作中研究日语单词)。这种方法不仅减少了注释工作，而且还限制了注释最近邻学习中使用的数据库的大小，这反过来又加快了分类算法。

It is worth noting that some authors (e.g., Thompson et al., 1999; Moskovitch et al., 2007) use “selective sampling” to refer to the pool-based scenario described in the next section. Under this interpretation, the term merely signifies that queries are made with a select set of instances sampled from a real data distribution. However, in most of the literature selective sampling refers to the stream-based scenario described here.

2.3 Pool-Based Sampling

For many real-world learning problems, large collections of unlabeled data can be gathered at once. This motivates *pool-based sampling* (Lewis and Gale, 1994), which assumes that there is a small set of labeled data $\mathcal{L}$ and a large pool of unlabeled data $\mathcal{U}$ available. Queries are selectively drawn from the pool, which is usually assumed to be closed (i.e., static or non-changing), although this is not strictly necessary. Typically, instances are queried in a greedy fashion, according to an informativeness measure used to evaluate all instances in the pool (or, perhaps if $\mathcal{U}$ is very large, some subsample thereof). The examples from Section 1.2 use this active learning setting.

The pool-based scenario has been studied for many real-world problem domains in machine learning, such as text classification (Lewis and Gale, 1994; McCallum and Nigam, 1998; Tong and Koller, 2000; Hoi et al., 2006a), information extraction (Thompson et al., 1999; Settles and Craven, 2008), image classification and retrieval (Tong and Chang, 2001; Zhang and Chen, 2002), video classification and retrieval (Yan et al., 2003; Hauptmann et al., 2006), speech recognition (Tür et al., 2005), and cancer diagnosis (Liu, 2004) to name a few.

The main difference between stream-based and pool-based active learning is that the former scans through the data sequentially and makes query decisions individually, whereas the latter evaluates and ranks the entire collection before selecting the best query. While the pool-based scenario appears to be much more common among application papers, one can imagine settings where the stream-based approach is more appropriate. For example, when memory or processing power may be limited, as with mobile and embedded devices.

3 Query Strategy Frameworks

All active learning scenarios involve evaluating the informativeness of unlabeled instances, which can either be generated de novo or sampled from a given distribution. There have been many proposed ways of formulating such *query strategies* in the literature. This section provides an overview of the general frameworks that are used. From this point on, I use the notation x^*_A to refer to the most informative instance (i.e., the best query) according to some query selection algorithm A.

3.1 Uncertainty Sampling

Perhaps the simplest and most commonly used query framework is *uncertainty sampling* (Lewis and Gale, 1994). In this framework, an active learner queries the instances about which it is least certain how to label. This approach is often straightforward for probabilistic learning models. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5 (Lewis and Gale, 1994; Lewis and Catlett, 1994).

For problems with three or more class labels, a more general uncertainty sampling variant might query the instance whose prediction is the *least confident*:

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x),$$

最低置信度策略：最接近0.5的

where $\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$, or the class label with the highest posterior probability under the model θ. One way to interpret this uncertainty measure is the expected 0/1-loss, i.e., the model's belief that it will mislabel x. This sort of strategy has been popular, for example, with statistical sequence models in information extraction tasks (Culotta and McCallum, 2005; Settles and Craven, 2008). This is because the most likely label sequence (and its associated likelihood) can be efficiently computed using dynamic programming.

However, the criterion for the least confident strategy only considers information about the most probable label. Thus, it effectively "throws away" information about the remaining label distribution. To correct for this, some researchers use a different multi-class uncertainty sampling variant called *margin sampling* (Scheffer et al., 2001):

$$x^*_M = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x),$$

边界采样：第一可能标签概率-第二可能标签概率

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable class labels under the model, respectively. Margin sampling aims to correct for a shortcoming in least confident strategy, by incorporating the posterior of the second most likely label. Intuitively, instances with large margins are easy, since the classifier has little doubt in differentiating between the two most likely class labels. Instances with small margins are more ambiguous, thus knowing the true label would help the model discriminate more effectively between them. However, for problems with very large label sets, the margin approach still ignores much of the output distribution for the remaining classes.

A more general uncertainty sampling strategy (and possibly the most popular) uses *entropy* (Shannon, 1948) as an uncertainty measure:

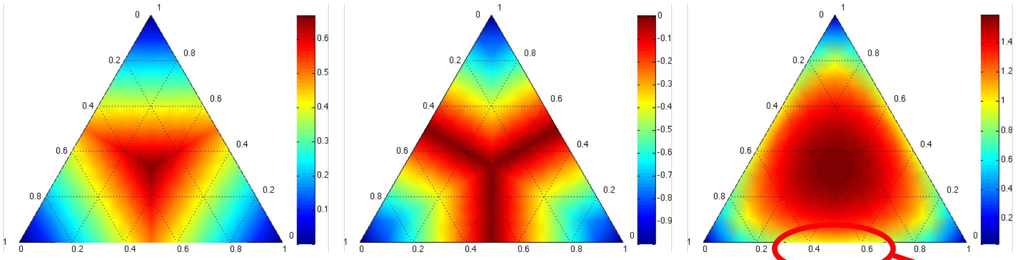
$$x^*_H = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x),$$

计算熵：直接计算信息量

where y_i ranges over all possible labelings. Entropy is an information-theoretic measure that represents the amount of information needed to "encode" a distribution. As such, it is often thought of as a measure of uncertainty or impurity in machine learning. For binary classification, entropy-based sampling reduces to the margin and least confident strategies above; in fact all three are equivalent to querying the instance with a class posterior closest to 0.5. However, the entropy-based approach generalizes easily to probabilistic multi-label classifiers and probabilistic models for more complex structured instances, such as sequences (Settles and Craven, 2008) and trees (Hwa, 2004).

熵：编码一个 分布/事件
所需要的最小理论信息量

Figure 5 visualizes the implicit relationship among these uncertainty measures. In all cases, the *most* informative instance would lie at the center of the triangle, because this represents where the posterior label distribution is most uniform (and thus least certain under the model). Similarly, the *least* informative instances are at the three corners, where one of the classes has extremely high



(a) least confident (b) margin (c) entropy

不会优先选中只有一个标签非常不可能的情况

Figure 5: Heatmaps illustrating the query behavior of common uncertainty measures in a three-label classification problem. Simplex corners indicate where one label has very high probability, with the opposite edge showing the probability range for the *other* two classes when that label has very low probability. Simplex centers represent a uniform posterior distribution. The most informative query region for each strategy is shown in dark red, radiating from the centers.

probability (and thus little model uncertainty). The main differences lie in the rest of the probability space. For example, the entropy measure does not favor instances where only one of the labels is highly *un*likely (i.e., along the outer side edges), because the model is fairly certain that it is *not* the true label. The least confident and margin measures, on the other hand, consider such instances to be useful if the model cannot distinguish between the remaining two classes. Empirical comparisons of these measures (e.g., Körner and Wrobel, 2006; Schein and Ungar, 2007; Settles and Craven, 2008) have yielded mixed results, suggesting that the best strategy may be application-dependent (note that all strategies still generally outperform passive baselines). Intuitively, though, entropy seems appropriate if the objective function is to minimize log-loss, while the other two (particularly margin) are more appropriate if we aim to reduce classification error, since they prefer instances that would help the model better discriminate among specific classes.

实证比较

概率(因此模型的不确定性很小)。主要的区别在于概率空间的其余部分。例如，熵度量不支持只有一个标签非常不可能的情况(即沿外侧边缘)，因为模型相当确定它不是真正的标签。另一方面，如果模型不能区分其余两类，则最不自信和边际度量认为这些实例是有用的。

这些措施的实证比较(例如，Körner和Wrobel，2006；Schein and Ungar，2007；Settles和Craven(2008)得出了不同的结果，表明最佳策略可能是应用相关的(注意，所有策略通常仍然优于被动基线)。直观地说，如果目标函数是最小化对数损失，熵似乎是合适的，而如果我们的目标是减少分类错误，其他两个(特别是边际)更合适，因为它们更倾向于帮助模型更好地区分特定类别的实例。

Uncertainty sampling strategies may also be employed with non-probabilistic classifiers. One of the first works to explore uncertainty sampling used a decision tree classifier (Lewis and Catlett, 1994). Similar approaches have been applied to active learning with nearest-neighbor (a.k.a. "memory-based" or "instance-based") classifiers (Fujii et al., 1998; Lindenbaum et al., 2004), by allowing each neighbor to vote on the class label of x, with the proportion of these votes representing the posterior label probability. Tong and Koller (2000) also experiment

不确定性采样策略也可用于非概率分类器。探索不确定性抽样的首批作品之一使用了决策树分类器(Lewis和Catlett，1994)。类似的方法也被应用于基于最近邻的主动学习。“基于记忆”或“基于实例”)分类器(Fujii et al.，1998;Lindenbaum et al.，2004)，通过允许每个邻居对x的类标签进行投票，这些投票的比例表示后验标签概率。Tong和Koller(2000)也进行了实验

with an uncertainty sampling strategy for support vector machines—or SVMs—that involves querying the instance closest to the linear decision boundary. This last approach is analogous to uncertainty sampling with a probabilistic binary linear classifier, such as logistic regression or naïve Bayes.

支持向量机(svm)的不确定性采样策略涉及查询最接近线性决策边界的实例。最后一种方法类似于使用概率二元线性分类器的不确定性抽样，例如逻辑回归或naïve贝叶斯。

So far we have only discussed classification tasks, but uncertainty sampling is also applicable in *regression* problems (i.e., learning tasks where the output variable is a continuous value rather than a set of discrete class labels). In this setting, the learner simply queries the unlabeled instance for which the model has the highest output variance in its prediction. Under a Gaussian assumption, the entropy of a random variable is a monotonic function of its variance, so this approach is very much in same the spirit as entropy-based uncertainty sampling for classification. Closed-form approximations of output variance can be computed for a variety of models, including Gaussian random fields (Cressie, 1991) and neural networks (MacKay, 1992). Active learning for regression problems has a long history in the statistics literature, generally referred to as *optimal experimental design* (Federov, 1972). Such approaches shy away from uncertainty sampling in lieu of more sophisticated strategies, which we will explore further in Section 3.5.

到目前为止，我们只讨论了分类任务，但不确定性采样也适用于回归问题(即，输出变量是连续值而不是一组离散类标签的学习任务)。在这种设置中，学习器只需查询模型在其预测中具有最高输出方差的未标记实例。在高斯假设下，随机变量的熵是其方差的单调函数，因此这种方法与基于熵的不确定性抽样进行分类的精神非常相似。输出方差的封闭近似可以计算各种模型，包括高斯随机场(Cressie，1991)和神经网络(MacKay，1992)。针对回归问题的主动学习在统计文献中有着悠久的历史，通常被称为最优实验设计(Federov，1972)。这种方法回避不确定性抽样，而采用更复杂的策略，我们将在第3.5节进一步探讨。

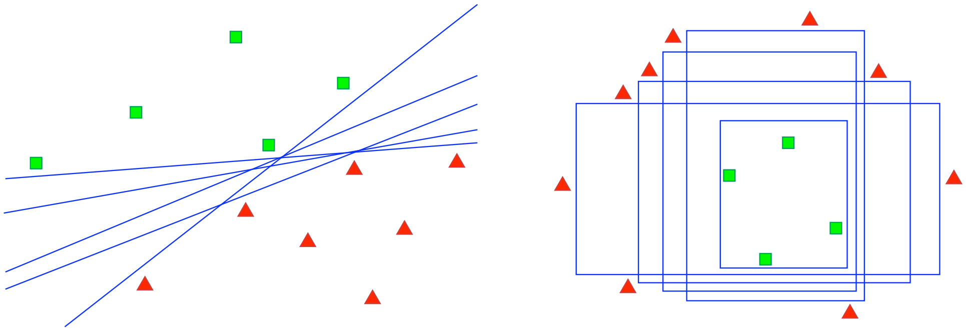
3.2 Query-By-Committee

Another, more theoretically-motivated query selection framework is the *query-by-committee* (QBC) algorithm (Seung et al., 1992). The QBC approach involves maintaining a committee $\mathcal{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ of models which are all trained on the current labeled set $\mathcal{L}$, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.

就是用一组基础数据训练c个模型，然后每个模型给未标记的实体预测标签。选出它们中差异最大的实体进行标记

The fundamental premise behind the QBC framework is minimizing the version space, which is (as described in Section 2.2) the set of hypotheses that are consistent with the current labeled training data $\mathcal{L}$. Figure 6 illustrates the concept of version spaces for (a) linear functions and (b) axis-parallel box classifiers in different binary classification tasks. If we view machine learning as a search for the "best" model within the version space, then our goal in active learning is to constrain the size of this space as much as possible (so that the search can be more precise) with as few labeled instances as possible. This is exactly what QBC aims to do, by querying in controversial regions of the input space. In order to implement a QBC selection algorithm, one must:

P在不同的二值分类任务。如果我们将机器学习视为在版本空间中搜索“最佳”模型，那么我们在主动学习中的目标就是尽可能地限制这个空间的大小(以便搜索可以更精确)，并且尽可能少地标记实例。这正是QBC的目标，通过查询输入空间中有争议的区域。为了实现QBC选择算法，必须：



(a)

(b)

Figure 6: Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in $\mathcal{L}$ (as indicated by shaded polygons), but each represents a different model in the version space.

i. be able to construct a committee of models that represent different regions of the version space, and

ii. have some measure of disagreement among committee members.

1 .能够构建一个代表版本空间不同区域的模型委员会，以及
2委员会成员之间存在一定程度的分歧。

Seung et al. (1992) accomplish the first task simply by sampling a committee of two random hypotheses that are consistent with $\mathcal{L}$. For generative model classes, this can be done more generally by randomly sampling an arbitrary number of models from some posterior distribution $P(\theta|\mathcal{L})$. For example, McCallum and Nigam (1998) do this for naïve Bayes by using the Dirichlet distribution over model parameters, whereas Dagan and Engelson (1995) sample hidden Markov models—or HMMs—by using the Normal distribution. For other model classes, such as discriminative or non-probabilistic models, Abe and Mamitsuka (1998) have proposed *query-by-boosting* and *query-by-bagging*, which employ the well-known ensemble learning methods boosting (Freund and Schapire, 1997) and bagging (Breiman, 1996) to construct committees. Melville and Mooney (2004) propose another ensemble-based method that explicitly encourages diversity among committee members. Muslea et al. (2000) construct a committee of two models by partitioning the feature space. There is no general agreement in the literature on the appropriate committee size to use, which may in fact vary by model class or application. However, even small committee sizes (e.g., two or three) have been shown to work well in practice (Seung et al., 1992; McCallum and Nigam, 1998; Settles and Craven, 2008).

基于后验概率可以构建任意的模型： 如图6

For measuring the level of disagreement, two main approaches have been proposed. The first is *vote entropy* (Dagan and Engelson, 1995):

最大投票熵：选x

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C},$$

where y_i again ranges over all possible labelings, and $V(y_i)$ is the number of "votes" that a label receives from among the committee members' predictions, and C is the committee size. This can be thought of as a QBC generalization of entropy-based uncertainty sampling. Another disagreement measure that has been proposed is average *Kullback-Leibler (KL) divergence* (McCallum and Nigam, 1998):

最大平均KL散度

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}} \| P_{\mathcal{C}}),$$

where:

$$D(P_{\theta^{(c)}} \| P_{\mathcal{C}}) = \sum_i P_{\theta^{(c)}}(y_i|x) \log \frac{P_{\theta^{(c)}}(y_i|x)}{P_{\mathcal{C}}(y_i|x)}.$$

Here $\theta^{(c)}$ represents a particular model in the committee, and $\mathcal{C}$ represents the committee as a whole, thus $P_{\mathcal{C}}(y_i|x) = \frac{1}{C}\sum_{c=1}^{C} P_{\theta^{(c)}}(y_i|x)$ is the "consensus" probability that y_i is the correct label. KL divergence (Kullback and Leibler, 1951) is an information-theoretic measure of the difference between two probability distributions. So this disagreement measure considers the most informative query to be the one with the largest average difference between the label distributions of any one committee member and the consensus. Other information-theoretic approaches like Jensen-Shannon divergence have also been used to measure disagreement (Melville et al., 2005), as well as the other uncertainty sampling measures discussed in Section 3.1, by pooling the model predictions to estimate class posteriors (Körner and Wrobel, 2006). Note also that in the equations above, such posterior estimates are based on committee members that cast "hard" votes for their respective label predictions. They might also cast "soft" votes using their posterior label probabilities, which in turn could be weighted by an estimate of each committee member's accuracy.

“ 共识

KL散度：衡量两个分布差异的度量

根据标签投票的叫 硬票
根据概率投票的为 软票

Aside from the QBC framework, several other query strategies attempt to minimize the version space as well. For example, Cohn et al. (1994) describe a selective sampling algorithm that uses a committee of two neural networks, the "most specific" and "most general" models, which lie at two extremes the version space given the current training set $\mathcal{L}$. Tong and Koller (2000) propose a pool-based margin strategy for SVMs which, as it turns out, attempts to minimize the version space directly. The membership query algorithms of Angluin (1988) and King et al. (2004) can also be interpreted as synthesizing instances de novo that most constrain the size of the version space. However, Haussler (1994) shows that the size of the version space can grow exponentially with the size of $\mathcal{L}$. This means that, in general, the version space of an arbitrary model class cannot be explicitly represented in practice. The QBC framework, rather, uses a committee to serve as a subset approximation.

e明确

版本空间的大小可以随着c的大小呈指数增长，这意味着，通常情况下，在实践中不能显式表示任意模型类的版本空间。相反，QBC框架使用委员会作为子集近似。

QBC can also be employed in regression settings, i.e., by measuring disagreement as the variance among the committee members' output predictions (Burbidge et al., 2007). Note, however, that there is no notion of "version space" for models that produce continuous outputs, so the interpretation of QBC in regression settings is a bit different. We can think of $\mathcal{L}$ as constraining the posterior joint probability of predicted output variables *and* the model parameters, $P(Y, \theta|\mathcal{L})$ (note that this applies for both regression and classification tasks). By integrating over a set of hypotheses and identifying queries that lie in controversial regions of the instance space, the learner attempts to collect data that reduces variance over both the output predictions and the parameters of the model itself (as opposed to uncertainty sampling, which focuses only on the output variance of a single hypothesis).

一个子集近似。QBC也可以用于回归设置，即通过测量委员会成员产出预测之间的差异来衡量分歧(Burbidge et al.， 2007)。但是请注意，对于产生连续输出的模型没有“ 版本空间 ”的概念，因此在回归设置中对QBC的解释有点不同。我们可以将C视为约束预测输出变量和模型参数P(Y,eIC)的后验联合概率(注意，这适用于回归和分类任务)。通过对一组假设进行整合并识别位于实例空间有争议区域的查询，学习器试图收集减少输出预测和模型本身参数方差的数据(与不确定性采样相反，不确定性采样只关注单个假设的输出方差)。

3.3 Expected Model Change

Another general active learning framework uses a decision-theoretic approach, selecting the instance that would impart the greatest change to the current model *if we knew its label.* An example query strategy in this framework is the "expected gradient length" (EGL) approach for discriminative probabilistic model classes. This strategy was introduced by Settles et al. (2008b) for active learning in the multiple-instance setting (see Section 6.4), and has also been applied to probabilistic sequence models like CRFs (Settles and Craven, 2008).

In theory, the EGL strategy can be applied to any learning problem where gradient-based training is used. Since discriminative probabilistic models are usually trained using gradient-based optimization, the "change" imparted to the model can be measured by the length of the training gradient (i.e., the vector used to re-estimate parameter values). In other words, the learner should query the instance x which, if labeled and added to $\mathcal{L}$, would result in the new training gradient of the largest magnitude. Let $\nabla\ell_\theta(\mathcal{L})$ be the gradient of the objective function ℓ with respect to the model parameters θ. Now let $\nabla\ell_\theta(\mathcal{L} \cup \langle x, y\rangle)$ be the new gradient that would be obtained by adding the training tuple $\langle x, y \rangle$ to $\mathcal{L}$. Since the query algorithm does not know the true label y in advance, we must instead calculate the length as an expectation over the possible labelings:

$$x^*_{EGL} = \underset{x}{\operatorname{argmax}} \sum_i P_\theta(y_i|x) \Big\| \nabla \ell_\theta(\mathcal{L} \cup \langle x, y_i \rangle) \Big\|,$$

where $\| \cdot \|$ is, in this case, the Euclidean norm of each resulting gradient vector. Note that, at query time, $\|\nabla \ell_\theta(\mathcal{L})\|$ should be nearly zero since ℓ converged at the previous round of training. Thus, we can approximate $\nabla \ell_\theta(\mathcal{L} \cup \langle x, y_i \rangle) \approx \nabla \ell_\theta(\langle x, y_i \rangle)$ for computational efficiency, because training instances are usually assumed to be independent.

The intuition behind this framework is that it prefers instances that are likely to most influence the model (i.e., have greatest impact on its parameters), regardless of the resulting query label. This approach has been shown to work well in empirical studies, but can be computationally expensive if both the feature space and set of labelings are very large. Furthermore, the EGL approach can be led astray if features are not properly scaled. That is, the informativeness of a given instance may be over-estimated simply because one of its feature values is unusually large, or the corresponding parameter estimate is larger, both resulting in a gradient of high magnitude. Parameter regularization (Chen and Rosenfeld, 2000; Goodman, 2004) can help control this effect somewhat, and it doesn't appear to be a significant problem in practice.

问题在于计算量大
特征需要缩放
可能忽略富含信息的实体

误入歧途

3.4 Expected Error Reduction

Another decision-theoretic approach aims to measure not how much the model is likely to change, but how much its generalization error is likely to be reduced. The idea it to estimate the expected future error of a model trained using $\mathcal{L} \cup \langle x, y \rangle$ on the remaining unlabeled instances in $\mathcal{U}$ (which is assumed to be representative of the test distribution, and used as a sort of validation set), and query the instance with minimal expected future error (sometimes called *risk*). One approach is to minimize the expected 0/1-loss:

$$x^*_{0/1} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(\sum_{u=1}^{U} 1 - P_{\theta^{+\langle x, y_i \rangle}}(\hat{y}|x^{(u)}) \right),$$

where $\theta^{+\langle x, y_i \rangle}$ refers to the the new model after it has been re-trained with the training tuple $\langle x, y_i \rangle$ added to $\mathcal{L}$. Note that, as with EGL in the previous section, we do not know the true label for each query instance, so we approximate using expectation over all possible labels under the current model θ. The objective here is to reduce the expected total number of incorrect predictions. Another, less stringent objective is to minimize the expected log-loss:

$$x^*_{\text{log}} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(- \sum_{u=1}^{U} \sum_j P_{\theta+\langle x,y_i \rangle}(y_j|x^{(u)}) \log P_{\theta+\langle x,y_i \rangle}(y_j|x^{(u)}) \right),$$

which is equivalent to reducing the expected entropy over $\mathcal{U}$. Another interpretation of this strategy is maximizing the expected *information gain* of the query x, or (equivalently) the *mutual information* of the output variables over x and $\mathcal{U}$.

这相当于减少期望熵除以u，另一种解释该策略的目的是最大化查询x的预期信息增益，或者(等价地)输出变量在x和u上的互信息。

Roy and McCallum (2001) first proposed the expected error reduction framework for text classification using naïve Bayes. Zhu et al. (2003) combined this framework with a semi-supervised learning approach (Section 7.1), resulting in a dramatic improvement over random or uncertainty sampling. Guo and Greiner (2007) employ an "optimistic" variant that biases the expectation toward the most likely label for computational convenience, using uncertainty sampling as a fallback strategy when the oracle provides an unexpected labeling. This framework has the dual advantage of being near-optimal and not being dependent on the model class. All that is required is an appropriate objective function and a way to estimate posterior label probabilities. For example, strategies in this framework have been successfully used with a variety of models including naïve Bayes (Roy and McCallum, 2001), Gaussian random fields (Zhu et al., 2003), logistic regression (Guo and Greiner, 2007), and support vector machines (Moskovitch et al., 2007). In theory, the general approach can be employed not only to minimize loss functions, but to optimize any generic performance measure of interest, such as maximizing precision, recall, F_1-measure, or area under the ROC curve.

Roy和McCallum(2001)首先使用naïve贝叶斯提出了文本分类的预期错误减少框架。Zhu等人(2003)将该框架与半监督学习方法(第7.1节)相结合，大大改善了随机或不确定性采样。Guo和Greiner(2007)采用了一种“乐观”变体，为了计算方便，将期望偏向于最可能的标签，当oracle提供意外标签时，使用不确定性采样作为后备策略。该框架具有接近最优和不依赖于模型类的双重优势。所需要的只是一个合适的目标函数和一种估计后验标签概率的方法。例如，该框架中的策略已成功用于各种模型，包括naïve贝叶斯(Roy and McCallum, 2001)、高斯随机场(Zhu et al.， 2003)、逻辑回归(Guo and Greiner, 2007)和支持向量机(Moskovitch et al., 2007)。理论上，一般方法不仅可以用于最小化损失函数，还可以用于优化任何感兴趣的通用性能度量，例如最大化精度、召回率、fi度量或ROC曲线下的面积。

后备策略

双重优势

In most cases, unfortunately, expected error reduction is also the most computationally expensive query framework. Not only does it require estimating the expected future error over $\mathcal{U}$ for each query, but a new model must be incrementally re-trained for each possible query labeling, which in turn iterates over the entire pool. This leads to a drastic increase in computational cost. For nonparametric model classes such as Gaussian random fields (Zhu et al., 2003), the incremental training procedure is efficient and exact, making this approach fairly practical[1]. For a many other model classes, this is not the case. For example, a binary logistic regression model would require $O(ULG)$ time complexity simply

不幸的是，在大多数情况下，期望的减少错误也是计算成本最高的查询框架。它不仅需要估计每个查询在U上的预期未来误差，而且必须为每个可能的查询标记增量地重新训练新模型，这反过来又遍历整个池。这将导致计算成本的急剧增加。对于非参数模型类，如高斯随机场(Zhu et al.，2003)，增量训练过程是高效和精确的，使这种方法相当实用1。对于许多其他模型类，情况并非如此。例如，二元逻辑回归模型只需要O (ULG)的时间复杂度

增量训练

[1] The bottleneck in non-parametric models generally not re-training, but inference.

to choose the next query, where U is the size of the unlabeled pool $\mathcal{U}$, L is the size of the current training set $\mathcal{L}$, and G is the number of gradient computations required by the by optimization procedure until convergence. A classification task with three or more labels using a MaxEnt model (Berger et al., 1996) would require $O(M^2ULG)$ time complexity, where M is the number of class labels. For a sequence labeling task using CRFs, the complexity explodes to $O(TM^{T+2}ULG)$, where T is the length of an input sequence. Because of this, the applications of the expected error reduction framework have mostly only considered simple binary classification tasks. Moreover, because the approach is often still impractical, researchers must resort to Monte Carlo sampling from the pool (Roy and McCallum, 2001) to reduce the U term in the previous analysis, or use approximate training techniques (Guo and Greiner, 2007) to reduce the G term.

选择下一个查询，其中U为未标记池U的大小，L为当前训练集C的大小，G为by优化过程直到收敛所需的梯度计算次数。使用MaxEnt模型(Berger et al.， 1996)具有三个或更多标签的分类任务将需要O (M2ULG)时间复杂度，其中M是类标签的数量。对于使用CRFs的序列标记任务，复杂度爆炸到O (TMT+2ULG)，其中T是输入序列的长度。正因为如此，预期误差减少框架的应用大多只考虑简单的二值分类任务。此外，由于这种方法往往仍然不切实际，研究人员必须采用蒙特卡洛抽样(Roy and McCallum, 2001)来减少先前分析中的U项，或者使用近似训练技术(Guo and Greiner, 2007)来减少G项。

3.5 Variance Reduction

Minimizing the expectation of a loss function directly is expensive, and in general this cannot be done in closed form. However, we can still reduce generalization error *indirectly* by minimizing output variance, which sometimes does have a closed-form solution. Consider a regression problem, where the learning objective is to minimize standard error (i.e., squared-loss). We can take advantage of the result of Geman et al. (1992), showing that a learner's expected future error can be decomposed in the following way:

$$\begin{aligned} E_T\left[(\hat{y}-y)^2|x\right] &= E\left[(y-E[y|x])^2\right] \quad \text{noise} \\ &+ \left(E_{\mathcal{L}}[\hat{y}]-E[y|x]\right)^2 \quad \text{bias} \\ &+ E_{\mathcal{L}}\left[(\hat{y}-E_{\mathcal{L}}[\hat{y}])^2\right], \quad \text{variance} \end{aligned}$$

where $E_{\mathcal{L}}[\cdot]$ is an expectation over the labeled set $\mathcal{L}$, $E[\cdot]$ is an expectation over the conditional density $P(y|x)$, and E_T is an expectation over both. Here also $\hat{y}$ is shorthand for the model's predicted output for a given instance x, while y indicates the true label for that instance.

The first term on the right-hand side of this equation is *noise*, i.e., the variance of the true label y given only x, which does not depend on the model or training data. Such noise may result from stochastic effects of the method used to obtain the labels, for example, or because the feature representation is inadequate. The second term is the *bias*, which represents the error due to the model class itself, e.g., if a linear model is used to learn a function that is only approximately linear. This component of the overall error is invariant given a fixed model class.

The third term is the model's *variance*, which is the remaining component of the learner's squared-loss with respect to the target function. Minimizing the variance, then, is guaranteed to minimize the future generalization error of the model (since the learner itself can do nothing about the noise or bias components).

Cohn (1994) and Cohn et al. (1996) present the first statistical analyses of active learning for regression in the context of a robot arm kinematics problem, using the estimated distribution of the model's output $\sigma^2_{\hat{y}}$. They show that this can be done in closed-form for neural networks, Gaussian mixture models, and locally-weighted linear regression. In particular, for neural networks the output variance for some instance x can be approximated by (MacKay, 1992):

Cohn(1994)和Cohn等人(1996)首次利用模型输出0的估计分布，在机械臂运动学问题的背景下，对主动学习进行回归的统计分析。他们表明，这可以以神经网络、高斯混合模型和局部加权线性回归的封闭形式完成。特别是，对于神经网络，某些实例x的输出方差可以近似为(MacKay，1992)：

$$\sigma^2_{\hat{y}}(x) \approx \left[\frac{\partial \hat{y}}{\partial \theta}\right]^{\mathsf{T}} \left[\frac{\partial^2}{\partial \theta^2} S_\theta(\mathcal{L})\right]^{-1} \left[\frac{\partial \hat{y}}{\partial \theta}\right] \approx \nabla x^{\mathsf{T}} F^{-1} \nabla x,$$

where $S_\theta(\mathcal{L})$ is the squared error of the current model θ on the training set $\mathcal{L}$. In the equation above, the first and last terms are computed using the gradient of the model's predicted output with respect to model parameters, written in shorthand as ∇x. The middle term is the inverse of a covariance matrix representing a second-order expansion around the objective function S with respect to θ, written in shorthand as F. This is also known as the ***Fisher information matrix*** (Schervish, 1995), and will be discussed in more detail later. An expression for $\langle \tilde{\sigma}^2_{\hat{y}} \rangle^{+x}$ can then be derived, which is the estimated mean output variance across the input distribution after the model has been re-trained on query x and its corresponding label. Given the assumptions that the model's prediction for x is fairly good, that ∇x is locally linear (true for most network configurations), and that variance is Gaussian, variance can be estimated efficiently in closed form so that actual model re-training is not required; more gory details are given by Cohn (1994). The *variance reduction* query selection strategy then becomes:

其中So(C)是当前模型0在训练集C上的平方误差。在上面的方程中，第一项和最后一项是使用模型预测输出相对于模型参数的梯度来计算的，简称为Vx。中间项是协方差矩阵的逆，表示围绕目标函数S对θ的二阶展开，简写为f。这也被称为Fisher信息矩阵(Schervish，1995)，稍后将详细讨论。然后可以导出()+的表达式，它是在查询x及其相应标签上重新训练模型后，在整个输入分布上估计的平均输出方差。假设模型对x的预测相当好，Vx是局部线性的(对于大多数网络配置都是如此)，并且方差是高斯的，那么方差可以以封闭形式有效地估计，从而不需要实际的模型重新训练;Cohn(1994)给出了更血腥的细节。方差减少查询选择策略变为：

$$x^*_{VR} = \underset{x}{\operatorname{argmin}} \; \langle \tilde{\sigma}^2_{\hat{y}} \rangle^{+x}.$$

Because this equation represents a smooth function that is differentiable with respect to any query instance x in the input space, gradient methods can be used to search for the best possible query that minimizes output variance, and therefore generalization error. Hence, their approach is an example of query synthesis (Section 2.1), rather than stream-based or pool-based active learning.

因为这个方程表示一个光滑的函数，它对于输入空间中的任何查询实例x都是可微的，所以梯度方法可以用来搜索最小化输出方差的最佳查询，从而最小化泛化误差。因此，他们的方法是查询综合的一个例子(第2.1节)，而不是基于流或基于池的主动学习。

This sort of approach is derived from statistical theories of *optimal experimental design*, or OED (Federov, 1972; Chaloner and Verdinelli, 1995). A key ingredient of these approaches is Fisher information, which is sometimes written $\mathcal{I}(\theta)$ to make its relationship with model parameters explicit. Formally, Fisher information is the variance of the *score*, which is the partial derivative of the log-likelihood function with respect to the model parameters:

这些方法的组成部分是费雪信息，有时写成I(0)，以明确其与模型参数的关系。形式上，Fisher信息是分数的方差，即对数似然函数对模型参数的偏导数:

$$\mathcal{I}(\theta) = N \int_x P(x) \int_y P_\theta(y|x) \frac{\partial^2}{\partial \theta^2} \log P_\theta(y|x),$$

where there are N independent samples drawn from the input distribution. This measure is convenient because its inverse sets a lower bound on the variance of the model's parameter estimates; this result is known as the Cramér-Rao inequality (Cover and Thomas, 2006). In other words, to minimize the variance over its parameter estimates, an active learner should select data that maximizes its Fisher information (or minimizes the inverse thereof). When there is only one parameter in the model, this strategy is straightforward. But for models of K parameters, Fisher information takes the form of a $K \times K$ covariance matrix (denoted earlier as F), and deciding what exactly to optimize is a bit tricky. In the OED literature, there are three types of optimal designs in such cases:

从输入分布中抽取N个独立样本。这种方法很方便，因为它的逆设定了模型参数估计方差的下界;这个结果被称为cram<s:1> - rao不等式(Cover and Thomas, 2006)。换句话说，为了最小化其参数估计的方差，主动学习者应该选择最大化其Fisher信息(或最小化其逆)的数据。当模型中只有一个参数时，这个策略很简单。但是对于K个参数的模型，Fisher信息采用K × K协方差矩阵(前面表示为F)的形式，并且决定优化的具体内容有点棘手。在《牛津英语词典》的文献中，在这种情况下有三种类型的最佳设计:

- *A-optimality* minimizes the *trace* of the inverse information matrix,
- *D-optimality* minimizes the *determinant* of the inverse matrix, and
- *E-optimality* minimizes the maximum *eigenvalue* of the inverse matrix.

a -最优性最小化逆信息矩阵的轨迹，
d -最优性最小化逆矩阵的行列式，并且
e -最优性最小化逆矩阵的最大特征值。

E-optimality doesn't seem to correspond to an obvious utility function, and is not often used in the machine learning literature, though there are some exceptions (Flaherty et al., 2006). *D*-optimality, it turns out, is related to minimizing the expected posterior entropy (Chaloner and Verdinelli, 1995). Since the determinant can be thought of as a measure of volume, the *D*-optimal design criterion essentially aims to minimize the volume of the (noisy) version space, with boundaries estimated via entropy, which makes it somewhat analogous to the query-by-committee algorithm (Section 3.2).

e -最优性似乎并不对应于一个明显的效用函数，并且在机器学习文献中不经常使用，尽管有一些例外(Flaherty et al.， 2006)。事实证明，d -最优性与最小化预期后验熵有关(Chaloner和Verdinelli, 1995)。由于行行式可以被认为是体积的度量，因此d -最优设计准则本质上旨在最小化(有噪声的)版本空间的体积，通过熵来估计边界，这使得它在某种程度上类似于委员会查询算法(第3.2节)。

A-optimal designs are considerably more popular, and aim to reduce the *average* variance of parameter estimates by focusing on values along the diagonal of the information matrix. A common variant of A-optimal design is to minimize $\mathrm{tr}(AF^{-1})$—the trace of the product of A and the inverse of the information matrix F—where A is a square, symmetric "reference" matrix. As a special case, consider a matrix of rank one: $A = \mathbf{c}\mathbf{c}^\mathsf{T}$, where $\mathbf{c}$ is some vector of length

a -最优设计相当流行，其目的是通过关注沿信息矩阵对角线的值来减少参数估计的平均方差。A-最优设计的一个常见变体是最小化tr(AF-1)——A与信息矩阵f的逆积的轨迹，其中A是一个方形的对称“参考”矩阵。作为一种特殊情况，考虑一个秩为1的矩阵:a = ccT，其中c是某个长度为1的向量

K (i.e., the same length as the model's parameter vector). In this case we have $\mathrm{tr}(AF^{-1}) = \mathbf{c}^{\mathsf{T}} F^{-1}\mathbf{c}$, and minimizing this value is sometimes called c-optimality. Note that, if we let $\mathbf{c} = \nabla x$, this criterion results in the equation for output variance $\sigma^2_{\hat{y}}(x)$ in neural networks defined earlier. Minimizing this variance measure can be achieved by simply querying on instance x, so the c-optimal criterion can be viewed as a formalism for uncertainty sampling (Section 3.1).

K(即与模型参数向量相同的长度)。在这种情况下，我们有tr(AF-1) = cTF-1c，最小化这个值有时被称为c-最优性。注意，如果我们让c = Vx，这个准则会得到前面定义的神经网络输出方差o(z)的方程。可以通过简单地查询实例x来最小化这种方差度量，因此c-最优准则可以被视为不确定性抽样的形式(第3.1节)。

Recall that we are interested in reducing variance across the input distribution (not merely for a single point in the instance space), thus the A matrix should encode the whole instance space. MacKay (1992) derived such solutions for regression with neural networks, while Zhang and Oles (2000) and Schein and Ungar (2007) derived similar methods for classification with logistic regression. Consider letting the reference matrix $A = \mathcal{I}_{\mathcal{U}}(\theta)$, i.e., the Fisher information of the unlabeled pool of instances $\mathcal{U}$, and letting $F = \mathcal{I}_x(\theta)$, i.e., the Fisher information of some query instance x. Using A-optimal design, we can derive the *Fisher information ratio* (Zhang and Oles, 2000):

回想一下，我们感兴趣的是减少整个输入分布的方差(不仅仅是实例空间中的单个点)，因此a矩阵应该编码整个实例空间。MacKay(1992)用神经网络推导了这样的回归解，而Zhang和Oles(2000)以及Schein和Ungar(2007)用逻辑回归推导了类似的分类方法。考虑让参考矩阵A = Iu(0)，即未标记实例池U的Fisher信息，让F = i(0)，即某个查询实例x的Fisher信息。利用A-最优设计，我们可以推导出Fisher信息比(Zhang and Oles, 2000):

$$x^*_{FIR} = \underset{x}{\mathrm{argmin}}\ \mathrm{tr}\left(\mathcal{I}_{\mathcal{U}}(\theta)\mathcal{I}_x(\theta)^{-1}\right).$$

The equation above provides us with a ratio given by the inner product of the two matrices, which can be interpreted as the model's output variance across the input distribution (as approximated by $\mathcal{U}$) that is not accounted for by x. Querying the instance which minimizes this ratio is then analogous to minimizing the future output variance once x has been labeled, thus indirectly reducing generalization error (with respect to $\mathcal{U}$). The advantage here over error reduction (Section 3.4) is that the model need not be retrained: the information matrices give us an approximation of output variance that simulates retraining. Zhang and Oles (2000) and Schein and Ungar (2007) applied this sort of approach to text classification using binary logistic regression. Hoi et al. (2006a) extended this to active text classification in the batch-mode setting (Section 6.1) in which a set of queries $\mathcal{Q}$ is selected all at once in an attempt to minimize the ratio between $\mathcal{I}_{\mathcal{U}}(\theta)$ and $\mathcal{I}_{\mathcal{Q}}(\theta)$. Settles and Craven (2008) have also generalized the Fisher information ratio approach to probabilistic sequence models such as CRFs.

用信息矩阵近似再训练的输出方差。模型不需要再训练很多次

上面的方程为我们提供了一个由两个矩阵的内积给出的比率，它可以解释为模型在输入分布(由u近似)上的输出方差，而不是由x来解释。查询最小化该比率的实例类似于在x被标记后最小化未来的输出方差。因此间接地减少了泛化误差(相对于U)。与减少误差(第3.4节)相比，这里的优点是模型不需要再训练:信息矩阵为我们提供了模拟再训练的输出方差的近似值。Zhang和Oles(2000)以及Schein和Ungar(2007)使用二元逻辑回归将这种方法应用于文本分类。Hoi等人(2006a)将其扩展到批处理模式设置下的主动文本分类(第6.1节)，其中一次选择一组查询Q，试图最小化Iu(0)和Zo(θ)之间的比率。Settles和Craven(2008)也将Fisher信息比方法推广到概率序列模型(如CRFs)中。

There are some practical disadvantages to these variance-reduction methods, however, in terms of computational complexity. Estimating output variance requires inverting a $K \times K$ matrix for each new instance, where K is the number of parameters in the model θ, resulting in a time complexity of $O(UK^3)$, where U is the size of the query pool $\mathcal{U}$. This quickly becomes intractable for large K, which is a common occurrence in, say, natural language processing tasks. Paass

然而，就计算复杂性而言，这些方差减少方法存在一些实际缺点。估计输出方差需要为每个新实例反转一个K × K矩阵，其中K是模型中参数的数量θ，导致时间复杂度为O (UK3)，其中U是查询池U的大小。对于大K来说，这很快变得难以处理，这在自然语言处理任务中很常见。paas and Kindermann (1995) propose a sampling approach based on Markov chains to reduce the U term in this analysis. For inverting the Fisher information matrix and reducing the K^3 term, Hoi et al. (2006a) use principal component analysis to reduce the dimensionality of the parameter space. Alternatively, Settles and Craven (2008) approximate the matrix with its diagonal vector, which can be inverted in only $O(K)$ time. However, these methods are still empirically much slower than simpler query strategies like uncertainty sampling.

和Kindermann(1995)提出了一种基于马尔可夫链的抽样方法来减少该分析中的U项。为了反演Fisher信息矩阵并降低K3项，Hoi等人(2006a)使用主成分分析来降低参数空间的维数。另外，Settles和Craven(2008)用其对角向量近似矩阵，该对角向量仅在O(K)时间内即可反转。然而，从经验来看，这些方法仍然比不确定性采样等简单的查询策略慢得多。

3.6 Density-Weighted Methods

A central idea of the estimated error and variance reduction frameworks is that they focus on the entire input space rather than individual instances. Thus, they are less prone to querying outliers than simpler query strategies like uncertainty sampling, QBC, and EGL. Figure 7 illustrates this problem for a binary linear classifier using uncertainty sampling. The least certain instance lies on the classification boundary, but is not "representative" of other instances in the distribution, so knowing its label is unlikely to improve accuracy on the data as a whole. QBC and EGL may exhibit similar behavior, by spending time querying possible outliers simply because they are controversial, or are expected to impart significant change in the model. By utilizing the unlabeled pool $\mathcal{U}$ when estimating future errors and output variances, the estimated error and variance reduction strategies implicitly avoid these problems. We can also overcome these problems by modeling the input distribution explicitly during query selection.

更不容易查询离群值

隐含的避免了这些问题

估计误差和方差减少框架的中心思想是它们关注整个输入空间而不是单个实例。因此，与不确定性采样、QBC和EGL等更简单的查询策略相比，它们更不容易查询离群值。图7说明了二元线性的这个问题

最不确定的实例位于分类边界上，但不“代表”分布中的其他实例，因此知道它的标签不太可能提高整个数据的准确性。QBC和EGL可能表现出类似的行为，因为它们花费时间查询可能的异常值只是因为它们是有争议的，或者期望在模型中传递重要的变化。通过在估计未来误差和输出方差时使用未标记池u，估计误差和方差减小策略隐含地避免了这些问题。我们还可以通过在查询选择期间显式地建模输入分布来克服这些问题。

The *information density* framework described by Settles and Craven (2008), and further analyzed in Chapter 4 of Settles (2008), is a general density-weighting technique. The main idea is that informative instances should not only be those which are uncertain, but also those which are "representative" of the underlying distribution (i.e., inhabit dense regions of the input space). Therefore, we wish to query instances as follows:

$$x^*_{ID} = \underset{x}{\operatorname{argmax}}\, \phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} \operatorname{sim}(x, x^{(u)}) \right)^{\beta}.$$

用简单策略A选择x,乘以与其它实体的相似度，用β调整权重

Here, $\phi_A(x)$ represents the informativeness of x according to some "base" query strategy A, such as an uncertainty sampling or QBC approach. The second term weights the informativeness of x by its average similarity to all other instances in the input distribution (as approximated by $\mathcal{U}$), subject to a parameter β that

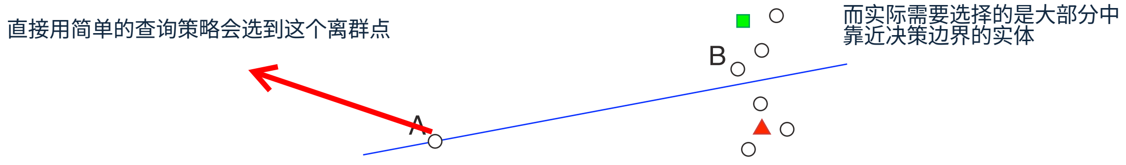

直接用简单的查询策略会选到这个离群点
B
A
而实际需要选择的是大部分中靠近决策边界的实体

Figure 7: An illustration of when uncertainty sampling can be a poor strategy for classification. Shaded polygons represent labeled instances in $\mathcal{L}$, and circles represent unlabeled instances in $\mathcal{U}$. Since A is on the decision boundary, it would be queried as the most uncertain. However, querying B is likely to result in more information about the data distribution as a whole.

图7:不确定性抽样在什么情况下是一种糟糕的分类策略。阴影多边形表示C中标记的实例，圆圈表示u中未标记的实例。由于A位于决策边界上，因此将其作为最不确定的查询。然而，查询B可能会得到关于整个数据分布的更多信息。

controls the relative importance of the density term. A variant of this might first cluster $\mathcal{U}$ and compute average similarity to instances in the same cluster.

This formulation was presented by Settles and Craven (2008), however it is not the only strategy to consider density and representativeness in the literature. McCallum and Nigam (1998) also developed a density-weighted QBC approach for text classification with naïve Bayes, which is a special case of information density. Fujii et al. (1998) considered a query strategy for nearest-neighbor methods that selects queries that are (i) least similar to the labeled instances in $\mathcal{L}$, and (ii) most similar to the unlabeled instances in $\mathcal{U}$. Nguyen and Smeulders (2004) proposed a density-based approach that first clusters instances and tries to avoid querying outliers by propagating label information to instances in the same cluster. Similarly, Xu et al. (2007) use clustering to construct sets of queries for batch-mode active learning (Section 6.1) with SVMs. Reported results in all these approaches are superior to methods that do not consider density or representativeness measures. Furthermore, Settles and Craven (2008) show that if densities can be pre-computed efficiently and cached for later use, the time required to select the next query is essentially no different than the base informativeness measure (e.g., uncertainty sampling). This is advantageous for conducting active learning interactively with oracles in real-time.

控制密度项的相对重要性。这种方法的一种变体可能首先聚类l，并计算与同一集群中实例的平均相似度。这个公式是由Settles和Craven(2008)提出的，然而，它并不是文献中考虑密度和代表性的唯一策略。McCallum和Nigam(1998)还开发了一种密度加权的QBC方法，用于naïve贝叶斯的文本分类，这是信息密度的一种特殊情况。Fujii等人(1998)考虑了一种最近邻方法的查询策略，该方法选择(i)与C中标记的实例最不相似，(ii)与u中未标记的实例最相似的查询。Nguyen和Smeulders(2004)提出了一种基于密度的方法，该方法首先聚类实例，并试图通过将标签信息传播到同一聚类中的实例来避免查询异常值。类似地，Xu等人(2007)使用聚类构建查询集，用于使用支持向量机进行批处理模式主动学习(第6.1节)。所有这些方法的报告结果都优于不考虑密度或代表性措施的方法。此外，Settles和Craven(2008)表明，如果密度可以有效地预先计算并缓存以供以后使用，那么选择下一个查询所需的时间基本上与基本信息度量没有什么不同(例如，不确定度抽样)。这对于实时地与oracle交互进行主动学习是有利的。

4 Analysis of Active Learning

This section discusses some of the empirical and theoretical evidence for how and when active learning approaches can be successful.

4.1 Empirical Analysis

An important question is: "does active learning work?" Most of the empirical results in the published literature suggest that it does (e.g., the majority of papers in the bibliography of this survey). Furthermore, consider that software companies and large-scale research projects such as CiteSeer, Google, IBM, Microsoft, and Siemens are increasingly using active learning technologies in a variety of real-world applications[2]. Numerous published results and increased industry adoption seem to indicate that active learning methods have matured to the point of practical use in many situations.

As usual, however, there are caveats. In particular, consider that a training set built in cooperation with an active learner is inherently tied to the model that was used to generate it (i.e., the class of the model selecting the queries). Therefore, the labeled instances are a biased distribution, not drawn i.i.d. from the underlying natural density. If one were to change model classes—as we often do in machine learning when the state of the art advances—this training set may no longer be as useful to the new model class (see Section 6.6 for more discussion on this topic). Somewhat surprisingly, Schein and Ungar (2007) showed that active learning can sometimes require more labeled instances than passive learning even when using the *same* model class, in their case logistic regression. Guo and Schuurmans (2008) found that off-the-shelf query strategies, when myopically employed in a batch-mode setting (Section 6.1) are often much worse than random sampling. Gasperin (2009) reported negative results for active learning in an anaphora resolution task. Baldridge and Palmer (2009) found a curious inconsistency in how well active learning helps that seems to be correlated with the proficiency of the annotator (specifically, a domain expert was better utilized by an active learner than a domain novice, who was better suited to a passive learner).

然而，像往常一样，也有一些警告。特别是，考虑与主动学习者合作构建的训练集本质上与用于生成它的模型(即，选择查询的模型的类)相关联。因此，标记的实例是有偏分布，而不是从潜在的自然密度中抽取的。如果要改变模型类——就像我们在机器学习中经常做的那样，当技术进步时——这个训练集可能不再对新的模型类有用(关于这个主题的更多讨论，请参阅第6.6节)。有些令人惊讶的是，Schein和Ungar (2007)表明，主动学习有时比被动学习需要更多的标记实例，即使使用相同的模型类，在他们的情况下，逻辑回归。Guo和Schuurmans(2008)发现，在批处理模式设置(第6.1节)中使用现成的查询策略时，通常比随机抽样差得多。Gasperin (2009)报告了主动学习在回指消解任务中的负面结果。Baldridge和Palmer(2009)在主动学习的帮助程度上发现了一个奇怪的不一致，这似乎与注释者的熟练程度有关(具体来说，一个领域专家比一个领域新手更适合主动学习者，而新手更适合被动学习)。

主动学习的注释者在领域内的熟练程度影响主动学习的效果

不过

Nevertheless, active learning does reduce the number of labeled instances required to achieve a given level of accuracy in the majority of reported results (though, admittedly, this may be due to the publication bias). This is often true even for simple query strategies like uncertainty sampling. Tomanek and Olsson (2009) report in a survey that 91% of researchers who used active learning in large-scale annotation projects had their expectations fully or partially met. Despite these findings, the survey also states that 20% of respondents opted *not* to use active learning in such projects, specifically because they were "not convinced that [it] would work well in their scenario." This is likely because other

Son(2009)在一项调查中报告说，在大规模标注项目中使用主动学习的研究人员中，91%的人完全或部分达到了他们的期望。尽管有这些发现，调查还指出，20%的受访者选择不在此类项目中使用主动学习，特别是因为他们“ 不相信[它]会在他们的场景中很好地工作 ”。这可能是因为其他

[2] Based on personal communication with (respectively): C. Lee Giles, David "Pablo" Cohn, Prem Melville, Eric Horvitz, and Balaji Krishnapuram.

subtleties arise when using active learning in practice (implementation overhead among them). Section 6 discusses some of the more problematic issues for real-world active learning.

在实践中使用主动学习时会出现一些微妙之处(其中包括实现开销)。第6节讨论了现实世界中主动学习的一些更有问题的问题。

4.2 Theoretical Analysis

A strong theoretical case for why and when active learning should work remains somewhat elusive, although there have been some recent advances. In particular, it would be nice to have some sort of bound on the number of queries required to learn a sufficiently accurate model for a given task, and theoretical guarantees that this number is less than in the passive supervised setting. Consider the following toy learning task to illustrate the potential of active learning. Suppose that instances are points lying on a one-dimensional line, and our model class is a simple binary thresholding function g parameterized by θ:

尽管最近取得了一些进展，但关于主动学习为什么以及何时应该起作用的强有力的理论案例仍然有些难以捉摸。特别是，对于学习给定任务的足够精确的模型所需的查询数量有某种限制是很好的，并且理论上保证这个数字比被动监督设置中的数字要少。考虑下面的玩具学习任务来说明主动学习的潜力。假设实例是位于一维直线上的点，我们的模型类是一个简单的二元阈值函数g，参数化为θ：

$$g(x;\theta) = \begin{cases} 1 & \text{if } x > \theta \text{, and} \\ 0 & \text{otherwise.} \end{cases}$$

对于有规律的分布，主动学习的性能是可以很强的

According to the *probably approximately correct* (PAC) learning model (Valiant, 1984), if the underlying data distribution can be perfectly classified by some hypothesis θ, then it is enough to draw $O(1/\epsilon)$ random labeled instances, where ϵ is the maximum desired error rate. Now consider a pool-based active learning setting, in which we can acquire the same number of *unlabeled* instances from this distribution for free (or very inexpensively), and only labels incur a cost. If we arrange these points on the real line, their (unknown) labels are a sequence of zeros followed by ones, and our goal is to discover the location at which the transition occurs while paying for as few labels as possible. By conducting a simple binary search through these unlabeled instances, a classifier with error less than ϵ can be achieved with a mere $O(\log 1/\epsilon)$ queries—since all other labels can be inferred—resulting in an exponential reduction in the number of labeled instances. Of course, this is a simple, one-dimensional, noiseless, binary toy learning task. Generalizing this phenomenon to more interesting and realistic problem settings is the focus of much theoretical work in active learning.

根据可能近似正确(probably approximately correct, PAC)学习模型(Valiant, 1984)，如果底层数据分布可以通过某个假设θ完美分类，那么就可以绘制O(1/e)个随机标记实例，其中e为最大期望错误率。现在考虑一个基于池的主动学习设置，在这个设置中，我们可以免费(或非常便宜地)从这个分布中获得相同数量的未标记实例，并且只有标签会产生成本。如果我们将这些点排列在实线上，它们的(未知的)标签是一个0后面跟着1的序列，我们的目标是在花费尽可能少的标签的情况下发现转换发生的位置。通过对这些未标记的实例进行简单的二值搜索，只需O(log 1/e)次查询就可以得到一个误差小于e的分类器——因为所有其他标记都可以被推断出来——从而使标记实例的数量呈指数级减少。当然，这是一个简单的、一维的、无噪声的、二元玩具式的学习任务。将这种现象推广到更有趣、更现实的问题设置中，是主动学习中许多理论工作的重点。

There have been some fairly strong results for the membership query scenario, in which the learner is allowed to create query instances de novo and acquire their labels (Angluin, 1988, 2001). However, such instances can be difficult for humans to annotate (Lang and Baum, 1992) and may result in querying outliers, since they are not created according to the data's underlying natural density. A great many

对于成员查询场景，已经有了一些相当强大的结果，在这个场景中，学习者被允许从头创建查询实例并获取它们标签(Angluin, 1988,2001)。然而，这种情况对人类来说可能很困难(Lang and Baum, 1992)并且可能导致查询异常值，因为它们不是根据数据的潜在自然密度创建的。很多

applications for active learning assume that unlabeled data (drawn from a real distribution) are available, so these results also have limited practical impact.

A stronger early theoretical result in the stream-based and pool-based scenarios is an analysis of the query-by-committee (QBC) algorithm by Freund et al. (1997). They show that, under a Bayesian assumption, it is possible to achieve generalization error ϵ after seeing $O(d/\epsilon)$ unlabeled instances, where d is the *Vapnik-Chervonenkis (VC) dimension* (Vapnik and Chervonenkis, 1971) of the model space, and requesting only $O(d \log 1/\epsilon)$ labels. This, like the toy example above, is an exponential improvement over the typical $O(d/\epsilon)$ sample complexity of the supervised setting. This result can tempered somewhat by the computational complexity of the QBC algorithm in certain practical situations, but Gilad-Bachrach et al. (2006) offer some improvements by limiting the version space via kernel functions.

Dasgupta et al. (2005) propose a variant of the perceptron update rule which can achieve the same label complexity bounds as reported for QBC. Interestingly, they show that a *standard* perceptron makes a poor active learner in general, requiring $O(1/\epsilon^2)$ labels as a lower bound. The modified training update rule—originally proposed in a non-active setting by Blum et al. (1996)—is key in achieving the exponential savings. The two main differences between QBC and their approach are that (i) QBC is more limited, requiring a Bayesian assumption for the theoretical analysis, and (ii) QBC can be computationally prohibitive, whereas the modified perceptron algorithm is much more lightweight and efficient, even suitable for online learning.

Dasgupta等人(2005)提出了感知器更新规则的一种变体，该规则可以实现与QBC相同的标签复杂性界限。有趣的是，他们表明，标准感知器通常是一个糟糕的主动学习者，需要O(1/e2)个标签作为下界。Blum等人(1996)最初在非活动设置中提出的改进的训练更新规则是实现指数节省的关键。QBC和他们的方法之间的两个主要区别是:(i) QBC更有限，需要一个贝叶斯假设来进行理论分析，(ii) QBC可能在计算上是禁止的，而改进的感知器算法更轻量和高效，甚至适合在线学习。

In earlier work, Dasgupta (2004) also provided a variety of theoretical upper and lower bounds for active learning in the more general pool-based setting. In particular, if using linear classifiers the sample complexity can grow to $O(1/\epsilon)$ in the worst case, which offers no improvement over standard supervised learning, but is also no worse. Encouragingly, Balcan et al. (2008) also show that, asymptotically, certain active learning strategies should always better than supervised learning in the limit.

在早期的工作中，Dasgupta (2004)也为更普遍的基于池的环境下的主动学习提供了各种理论上的上限和下限。特别是，如果使用线性分类器，在最坏的情况下，样本复杂度可以增长到O(1/e)，这与标准的监督学习相比没有改进，但也没有更差。令人鼓舞的是，Balcan等人(2008)也表明，在极限情况下，某些主动学习策略应该总是优于监督学习。

Most of these results have used theoretical frameworks similar to the standard PAC model, and necessarily assume that the learner knows the correct concept class in advance. Put another way, they assume that *some* model in our hypothesis class can perfectly classify the instances, and that the data are also noise-free. To address these limitations, there has been some recent theoretical work in *agnostic active learning* (Balcan et al., 2006), which only requires that unlabeled instances are drawn i.i.d. from a fixed distribution, and even noisy distributions are allowed. Hanneke (2007) extends this work by providing upper bounds on

不可知论主动学习

这些结果大多使用了类似于标准PAC模型的理论框架，并且必须假设学习者事先知道正确的概念类。换句话说，他们假设我们的假设类中的某个模型可以完美地对实例进行分类，并且数据也是无噪声的。为了解决这些限制，最近有一些关于不可知论主动学习的理论工作(Balcan et al.， 2006)，它只要求从固定分布中抽取未标记的实例，甚至允许有噪声的分布。Hanneke(2007)扩展了这项工作，提供了

query complexity for the agnostic setting. Dasgupta et al. (2008) propose a somewhat more efficient query selection algorithm, by presenting a polynomial-time reduction from active learning to supervised learning for arbitrary input distributions and model classes. These agnostic active learning approaches explicitly use complexity bounds to determine which hypotheses still "look viable," so to speak, and queries can be assessed by how valuable they are in distinguishing among these viable hypotheses. Methods such as these have attractive PAC-style convergence guarantees and complexity bounds that are, in many cases, significantly better than passive learning.

不可知论设置的查询复杂度。Dasgupta等人(2008)提出了一种更有效的查询选择算法，通过对任意输入分布和模型类提出了从主动学习到监督学习的多项式时间缩减。这些不可知论的主动学习方法明确地使用复杂性界限来确定哪些假设仍然“ 看起来可行 ”，可以这么说，并且可以通过区分这些可行假设的价值来评估查询。诸如此类的方法具有吸引人的pac风格的收敛保证和复杂性界限，在许多情况下，它们明显优于被动学习。

However, most positive theoretical results to date have been based on intractable algorithms, or methods otherwise too prohibitively complex and particular to be used in practice. The few analyses performed on efficient algorithms have assumed uniform or near-uniform input distributions (Balcan et al., 2006; Dasgupta et al., 2005), or severely restricted hypothesis spaces. Furthermore, these studies have largely only been for simple classification problems. In fact, most are limited to binary classification with the goal of minimizing 0/1-loss, and are not easily adapted to other objective functions that may be more appropriate for many applications. Furthermore, some of these methods require an explicit enumeration over the version space, which is not only often intractable (see the discussion at the end of Section 3.2), but difficult to even consider for complex learning models (e.g., heterogeneous ensembles or structured prediction models for sequences, trees, and graphs). However, some recent theoretical work has begun to address these issues, coupled with promising empirical results (Dasgupta and Hsu, 2008; Beygelzimer et al., 2009).

然而，迄今为止，大多数积极的理论结果都是基于难以处理的算法，或者基于过于复杂和特殊而无法在实践中使用的方法。对高效算法进行的少数分析假设了均匀或接近均匀的输入分布(Balcan et al., 2006;Dasgupta et al., 2005)，或者严格限制假设空间。此外，这些研究在很大程度上只针对简单的分类问题。事实上，大多数都局限于以最小化0/1损失为目标的二值分类，并且不容易适应于可能更适合许多应用程序的其他目标函数。此外，其中一些方法需要对版本空间进行显式枚举，这不仅通常难以处理(参见3.2节末尾的讨论)，而且甚至难以考虑复杂的学习模型(例如，异构集成或序列、树和图的结构化预测模型)。然而，最近的一些理论工作已经开始解决这些问题，加上有希望的实证结果(Dasgupta和Hsu，2008; Beygelzimer et al., 2009)。

5 Problem Setting Variants

5问题设置变体本节讨论了传统主动学习工作在不同问题设置中的一些概括和扩展。

This section discusses some of the generalizations and extensions of traditional active learning work into different problem settings.

5.1 Active Learning for Structured Outputs

结构化输出，即序列化输出，如语言之类的输出问题

Active learning for classification tasks has been widely studied (e.g., Cohn et al., 1994; Zhang and Oles, 2000; Guo and Greiner, 2007). However, many important learning problems involve predicting structured outputs on instances, such as sequences and trees. Figure 8 illustrates how, for example, an information extraction problem can be viewed as a sequence labeling task. Let $\mathbf{x} = \langle x_1, \ldots, x_T \rangle$

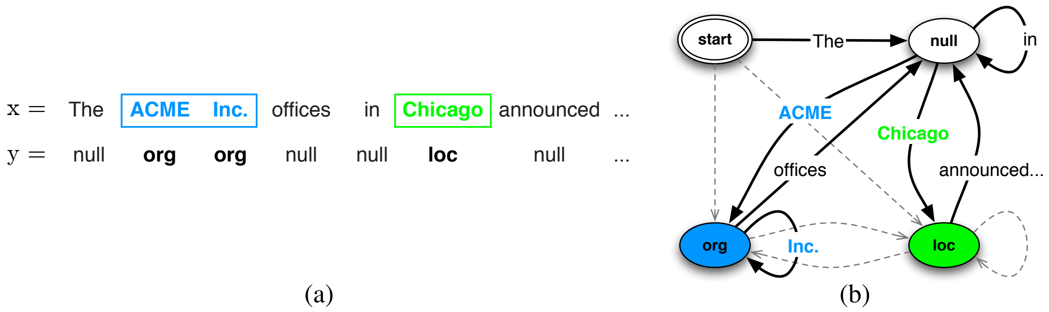

x = The ACME Inc. offices in Chicago announced ...
y = null org org null null loc null ...
(a)
start
The
null
in
ACME
Chicago
offices
announced...
org
Inc.
loc
(b)

Figure 8: An information extraction example viewed as a sequence labeling task. (a) A sample input sequence $\mathbf{x}$ and corresponding label sequence $\mathbf{y}$. (b) A sequence model represented as a finite state machine, illustrating the path of $\langle \mathbf{x}, \mathbf{y} \rangle$ through the model.

be an observation sequence of length T with a corresponding label sequence $\mathbf{y} = \langle y_1, \ldots, y_T \rangle$. Words in a sentence correspond to *tokens* in the input sequence $\mathbf{x}$, which are mapped to labels in $\mathbf{y}$. Figure 8(a) presents an example $\langle \mathbf{x}, \mathbf{y} \rangle$ pair. The labels indicate whether a given word belongs to a particular entity class of interest (org and loc in this case, for "organization" and "location," respectively) or not (null).

Unlike simpler classification tasks, each instance $\mathbf{x}$ in this setting is not represented by a single feature vector, but rather a structured sequence of feature vectors: one for each token (i.e., word). For example, the word "Madison" might be described by the features WORD=*Madison* and CAPITALIZED. However, it can variously correspond to the labels person ("The fourth U.S. President James Madison..."), loc ("The city of Madison, Wisconsin..."), and org ("Madison defeated St. Cloud in yesterday's hockey match..."). The appropriate label for a token often depends on its context in the sequence. For sequence-labeling problems like information extraction, labels are typically predicted by a *sequence model* based on a probabilistic finite state machine, such as CRFs or HMMs. An example sequence model is shown in Figure 8(b).

Settles and Craven (2008) present and evaluate a large number of active learning algorithms for sequence labeling tasks using probabilistic sequence models like CRFs. Most of these algorithms can be generalized to other probabilistic sequence models, such as HMMs (Dagan and Engelson, 1995; Scheffer et al., 2001) and probabilistic context-free grammars (Baldridge and Osborne, 2004;

Hwa, 2004). Thompson et al. (1999) also propose query strategies for structured output tasks like semantic parsing and information extraction using inductive logic programming methods.

5.2 Active Feature Acquisition and Classification

5.2主动特征获取与分类

In some learning domains, instances may have incomplete feature descriptions. For example, many data mining tasks in modern business are characterized by naturally incomplete customer data, due to reasons such as data ownership, client disclosure, or technological limitations. Consider a credit card company that wishes to model its most profitable customers; the company has access to data on client transactions using their own cards, but no data on transactions using cards from other companies. Here, the task of the model is to classify a customer using incomplete purchase information as the feature set. Similarly, consider a learning model used in medical diagnosis which has access to some patient symptom information, but not other data that require complex, expensive, or risky medical procedures. Here, the task of the model is to suggest a diagnosis using incomplete patient information as the feature set.

在某些学习领域中，实例可能具有不完整的特征描述。例如，由于数据所有权、客户披露或技术限制等原因，现代商业中的许多数据挖掘任务的特点是客户数据自然不完整。考虑一家信用卡公司，它希望以最赚钱的客户为模型;该公司可以访问使用自己的卡的客户交易数据，但无法访问使用其他公司的卡的交易数据。这里，模型的任务是使用不完整的购买信息作为特征集对客户进行分类。类似地，考虑医疗诊断中使用的学习模型，它可以访问一些患者症状信息，但不能访问需要复杂、昂贵或有风险的医疗过程的其他数据。在这里，模型的任务是使用不完整的患者信息作为特征集来建议诊断。

特征残缺的情况下，用模型推断特征，用查询策略选择需要请求特征值的实体

In these domains, *active feature acquisition* seeks to alleviate these problems by allowing the learner to request more complete feature information. The assumption is that additional features can be obtained at a cost, such as leasing transaction records from other credit card companies, or running additional diagnostic procedures. The goal in active feature acquisition is to select the most informative features to obtain during training, rather than randomly or exhaustively acquiring all new features for all training instances. Zheng and Padmanabhan (2002) proposed two “single-pass” approaches for this problem. In the first approach, they attempt to impute the missing values, and then acquire the ones about which the model has least confidence. As an alternative, they also consider imputing these values, training a classifiers on the imputed training instances, and only acquiring feature values for the instances which are misclassified. In contrast, incremental active feature acquisition may acquire values for a few salient features at a time, either by selecting a small batch of misclassified examples (Melville et al., 2004), or by taking a decision-theoretic approach and acquiring the feature values which are expected to maximize some utility function (Saar-Tsechansky et al., 2009).

在这些领域中，主动特征获取旨在通过允许学习者请求更完整的特征信息来缓解这些问题。假设额外的功能可以通过一定的成本获得，比如从其他信用卡公司租赁交易记录，或者运行额外的诊断过程。主动特征获取的目标是在训练过程中选择信息量最大的特征，而不是随机地或穷尽地获取所有训练实例的所有新特征。Zheng和Padmanabhan(2002)针对这一问题提出了两种“单通道”方法。在第一种方法中，他们试图推断缺失的值，然后获得模型置信度最低的值。作为一种选择，他们还考虑输入这些值，在输入的训练实例上训练分类器，并且只获取错误分类实例的特征值。相比之下，增量主动特征获取可能一次获取几个显著特征的值，要么通过选择一小批错误分类的示例(Melville et al.， 2004)，要么通过采用决策理论方法获取期望最大化某些效用函数的特征值(Saar-Tsechansky et al.， 2009)。

Similarly, work in *active classification* considers the case in which missing feature values may be obtained during classification (test time) rather than during training. Greiner et al. (2002) introduced this setting and provided a PAC-style theoretical analysis of learning such classifiers given a fixed budget. Variants of

类似地，主动分类中的工作考虑的是在分类(测试时间)而不是在训练期间获得缺失特征值的情况。Greiner et al. (2002)介绍了这种设置，并提供了pac式的理论分析，在给定固定预算的情况下学习这种分类器。变异的

在测试阶段，通过分类结果的模糊程度，来决定请求哪些实体的特征

naïve Bayes (Ling et al., 2004) and decision tree classifiers (Chai et al., 2004; Esmeir and Markovitch, 2008) have also been proposed to minimize costs at classification time. Typically, these are evaluated in terms of their total cost (feature acquisition plus misclassification, which must be converted into the same currency) as a function of the number of missing values. This is often flexible enough to incorporate other types of costs, such as delays between query time and value acquisition (Sheng and Ling, 2006). Another approach is to model the feature acquisition task as a sequence of decisions to either acquire more information or to terminate and make a prediction, using an HMM (Ji and Carin, 2007).

naïve贝叶斯(Ling et al., 2004)和决策树分类器(Chai et al., 2004;Esmeir和Markovitch(2008)也提出了最小化分类时的成本。通常，这些都是根据它们的总成本(特征)来评估的获取加上错误分类(必须转换成相同的货币)作为缺失价值数量的函数。这通常足够灵活，可以纳入其他类型的成本，例如查询时间和价值获取之间的延迟(Sheng and Ling, 2006)。另一种方法是使用HMM将特征获取任务建模为获取更多信息或终止并进行预测的一系列决策(Ji和Carin, 2007)。

The difference between these learning settings and typical active learning is that the "oracle" provides salient feature values rather than training labels. Since feature values can be highly variable in their acquisition costs (e.g., running two different medical tests might provide roughly the same predictive power, while one is half the cost of the other), some of these approaches are related in spirit to cost-sensitive active learning (see Section 6.3).

这些学习设置与典型主动学习的区别在于，“oracle”提供显著的特征值，而不是训练标签。由于特征值的获取成本变化很大(例如，运行两种不同的医学测试可能提供大致相同的预测能力，而其中一种的成本是另一种的一半)，因此其中一些方法在精神上与成本敏感型主动学习有关(见第6.3节)。

5.3 Active Class Selection

Active learning assumes that instances are freely or inexpensively obtained, and it is the *labeling* process that incurs a cost. Imagine the opposite scenario, however, where a learner is allowed to query a known class label, and obtaining each *instance* incurs a cost. This fairly new problem setting is known as *active class selection*. Lomasky et al. (2007) propose several active class selection query algorithms for an "artificial nose" task, in which a machine learns to discriminate between different vapor types (the class labels) which must be chemically synthesized (to generate the instances). Some of their approaches show significant gains over uniform class sampling, the "passive" learning equivalent.

通过label反推实体

5.4 Active Clustering

For most of this survey, we assume that the learner to be "activized" is *supervised*, i.e., the task of the learner is to induce a function that accurately predicts a label y for some new instance x. In contrast, a learning algorithm is called *unsupervised* if its job is simply to organize a large amount of unlabeled data in a meaningful way. The main difference is that supervised learners try to map instances into a pre-defined vocabulary of labels, while unsupervised learners exploit latent structure

5.4主集群在大多数调查中，我们假设被“激活”的学习者是被监督的，也就是说，学习器的任务是归纳出一个能准确预测标签y的函数相反，如果一个学习算法的工作只是简单地以一种有意义的方式组织大量未标记的数据，那么它就被称为无监督算法。主要区别在于有监督学习器尝试将实例映射到预定义的标签词汇表中，而无监督学习器则利用潜在结构

in the data alone to find meaningful patterns[3]. *Clustering* algorithms are probably the most common examples of unsupervised learning (e.g., see Chapter 10 of Duda et al., 2001).

Since active learning generally aims to select data that will reduce the model's classification error or label uncertainty, unsupervised active learning may seem a bit counter-intuitive. Nevertheless, Hofmann and Buhmann (1998) have proposed an active clustering algorithm for proximity data, based on an expected value of information criterion. The idea is to generate (or subsample) the unlabeled instances in such a way that they self-organize into groupings with less overlap or noise than for clusters induced using random sampling. The authors demonstrate improved clusterings in computer vision and text retrieval tasks.

由于主动学习通常旨在选择能够减少模型分类误差或标签不确定性的数据，因此无监督主动学习可能看起来有点违反直觉。然而，Hofmann和Buhmann(1998)提出了一种基于信息准则期望值的邻近数据主动聚类算法。我们的想法是生成(或子样本)未标记的实例，使它们自组织成分组，与使用随机抽样诱导的集群相比，重叠或噪声更少。作者在计算机视觉和文本检索任务中演示了改进的聚类。

Some clustering algorithms operate under certain constraints, e.g., a user can specify a priori that two instances *must* belong to the same cluster, or that two others *cannot*. Grira et al. (2005) Have explored an active variant of this approach for image databases, where queries take the form of such “must-link” and “cannot-link” constraints on similar or dissimilar images. Huang and Mitchell (2006) experiment with interactively-obtained clustering constraints on both instances and features, and Andrzejewski et al. (2009) address the analogous problem of incorporating constraints on features in *topic modeling* (Steyvers and Griffiths, 2007), another popular unsupervised learning technique. Although these last two works do not solicit constraints in an active manner, one can easily imagine extending them to do so. Active variants for these unsupervised methods are akin to the work on active learning by labeling features discussed in Section 6.4, with the subtle difference that constraints in the (semi-)supervised case are links between features and *labels*, rather than features (or instances) with one another.

一些聚类算法在一定的约束下运行，例如，用户可以预先指定两个实例必须属于同一个集群，或者其他两个实例不能属于同一个集群。Grira等人(2005)探索了这种方法在图像数据库中的一种主动变体，其中查询采用类似或不相似图像上的“必须链接”和“不能链接”约束的形式。Huang和Mitchell(2006)对实例和特征进行了交互获得聚类约束的实验，Andrzejewski等人(2009)解决了在主题建模中纳入特征约束的类似问题(Steyvers和Griffiths，2007)，这是另一种流行的无监督学习技术。虽然后两件作品并没有主动地寻求约束，但我们可以很容易地想象将它们扩展到这样做。这些无监督方法的主动变体类似于第6.4节中讨论的通过标记特征进行主动学习的工作，细微的区别是(半)监督情况下的约束是特征和标签之间的联系，而不是特征(或实例)之间的联系。

6 Practical Considerations

Until very recently, most active learning research has focused on mechanisms for choosing queries from the learner's perspective. In essence, this body of work addressed the question, “can machines learn with fewer training instances if they ask questions?” By and large, the answer to this question is “yes,” subject to some assumptions. For example, we often assume that there is a single oracle, or that the oracle is always correct, or that the cost for labeling queries is either free or uniformly expensive.

直到最近，大多数主动学习研究都集中在从学习者的角度选择问题的机制上。从本质上讲，这项工作解决了这样一个问题：“如果机器提出问题，它们能在更少的训练实例下学习吗？”总的来说，这个问题的答案是“是”，但要有一些假设。例如，我们经常假设只有一个oracle，或者oracle总是正确的，或者标记查询的成本要么是免费的，要么都是昂贵的。

[3] Note that *semi-supervised learning* (Section 7.1) also tries to exploit the latent structure of unlabeled data, but with the specific goal of improving label predictions.

In many real-world situations these assumptions do not hold. As a result, the research question for active learning has shifted in recent years to "can machines learn *more economically* if they ask questions?" This section describes several of the challenges for active learning in practice, and summarizes some the research that has addressed these issues to date.

在许多现实世界的情况下，这些假设并不成立。因此，近年来主动学习的研究问题已经转变为“如果机器提出问题，它们能更经济地学习吗?”本节描述了主动学习在实践中的几个挑战，并总结了迄今为止解决这些问题的一些研究。

6.1 Batch-Mode Active Learning

不再一个一个选，而是一批一批选

In most active learning research, queries are selected in *serial*, i.e., one at a time. However, sometimes the time required to induce a model is slow or expensive, as with large ensemble methods and many structured prediction tasks (see Section 5.1). Consider also that sometimes a distributed, parallel labeling environment may be available, e.g., multiple annotators working on different labeling workstations at the same time on a network. In both of these cases, selecting queries in serial may be inefficient. By contrast, *batch-mode* active learning allows the learner to query instances in groups, which is better suited to parallel labeling environments or models with slow training procedures.

在大多数主动学习研究中，查询是连续选择的，即一次选择一个。然而，有时建立模型所需的时间很慢或很昂贵，如大型集成方法和许多结构化预测任务(参见第5.1节)。还要考虑到有时可能有一个分布式的、并行的标注环境，例如，多个注释者同时在网络上的不同标注工作站上工作。在这两种情况下，按顺序选择查询可能效率很低。相比之下，批处理模式主动学习允许学习者以组为单位查询实例，这更适合并行标记环境或训练过程缓慢的模型。

The challenge in batch-mode active learning is how to properly assemble the optimal query set $\mathcal{Q}$. Myopically querying the "$\mathcal{Q}$-best" queries according to some instance-level query strategy often does not work well, since it fails to consider the overlap in information content among the "best" instances. To address this, a few batch-mode active learning algorithms have been proposed. Brinker (2003) considers an approach for SVMs that explicitly incorporates diversity among instances in the batch. Xu et al. (2007) propose a similar approach for SVM active learning, which also incorporates a density measure (Section 3.6). Specifically, they query the centroids of clusters of instances that lie closest to the decision boundary. Hoi et al. (2006a,b) extend the Fisher information framework (Section 3.5) to the batch-mode setting for binary logistic regression. Most of these approaches use greedy heuristics to ensure that instances in the batch are both diverse and informative, although Hoi et al. (2006b) exploit the properties of submodular functions (see Section 7.3) to find batches that are guaranteed to be near-optimal. Alternatively, Guo and Schuurmans (2008) treat batch construction for logistic regression as a discriminative optimization problem, and attempt to construct the most informative batch directly. For the most part, these approaches show improvements over random batch sampling, which in turn is generally better than simple "$\mathcal{Q}$-best" batch construction.

批处理模式主动学习面临的挑战是如何正确地组合最优查询集q。根据某些实例级查询策略，短视地查询“Q-best”查询往往不能很好地工作，因为它没有考虑“best”实例之间信息内容的重叠。为了解决这个问题，已经提出了一些批处理模式主动学习算法。Brinker(2003)考虑了一种支持向量机的方法，该方法显式地结合了批处理中实例之间的多样性。Xu等人(2007)提出了类似的SVM主动学习方法，该方法也包含了密度度量(第3.6节)。具体来说，它们查询离决策边界最近的实例簇的质心。Hoi等人(2006a,b)将Fisher信息框架(第3.5节)扩展到二元逻辑回归的批处理模式设置。尽管Hoi等人(2006b)利用子模块函数的属性(参见第7.3节)来找到保证接近最优的批，但大多数方法使用贪婪启发式来确保批中的实例既多样又有信息，(2006b)。另外，Guo和Schuurmans(2008)将逻辑回归的批构建视为判别优化问题，并试图直接构建信息量最大的批。在大多数情况下，这些方法比随机批抽样有改进，而随机批抽样通常比简单的“Q-best”批构造更好。

6.2 Noisy Oracles

现实中标记往往不是永远准确的。有实验误差（噪声），也有人类误差

Another strong assumption in most active learning work is that the quality of labeled data is high. 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 of experimental setting. Even if labels come from human experts, they may not always be reliable, for several reasons. First, some instances are implicitly difficult for people and machines, and second, people can become distracted or fatigued over time, introducing variability in the quality of their annotations. The recent introduction of Internet-based "crowdsourcing" tools such as Amazon's Mechanical Turk[4] and the clever use of online annotation games[5] have enabled some 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 also to evaluate learning algorithms on data for which no gold-standard labelings exist (Mintz et al., 2009; Carlson et al., 2010).

在大多数主动学习工作中，另一个强有力的假设是标记数据的质量很高。如果标签来自经验实验(例如，在生物、化学或临床研究中)，那么通常可以预期实验设置的仪器会产生一些噪声。即使标签来自人类专家，出于几个原因，它们也可能并不总是可靠的。首先，有些实例对于人和机器来说都很困难，其次，随着时间的推移，人们可能会分心或疲劳，从而导致注释质量的变化。最近引入的基于互联网的“众包”工具，如亚马逊的土耳其机械，以及在线注释游戏的巧妙使用，使一些研究人员能够通过廉价地从多个非专家那里获得标签，试图“平均”出一些噪音。这些方法已用于生成金标准质量训练集(Snow等人，2008)，也用于评估不存在金标准标签的数据上的学习算法(Mintz等人，2009; Carlson et al.， 2010)。

The question remains about how to use non-experts (or even noisy experts) as oracles in active learning. In particular, 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* training instance that seems a bit off? Sheng et al. (2008) study this problem using several heuristics that take into account estimates of both oracle and model uncertainty, and show that data can be improved by selective repeated labeling. However, their analysis assumes that (i) all oracles are equally and consistently noisy, and (ii) annotation is a noisy process over some underlying true label. Donmez et al. (2009) address the first issue by allowing annotators to have different noise levels, and show that both true instance labels and individual oracle qualities can be estimated (so long as they do not change over time). They take advantage of these estimates by querying only the more reliable annotators in subsequent iterations active learning.

问题仍然是如何在主动学习中使用非专家(甚至是嘈杂的专家)作为预言者。特别是，什么时候学习器应该决定查询一个新的未标记实例的(可能有噪声的)标签，而不是查询重复的标签来消除一个看起来有点偏离的现有训练实例的噪声?Sheng等人(2008)使用几种启发式方法研究了这个问题，这些方法考虑了oracle和模型不确定性的估计，并表明可以通过选择性重复标记来改进数据。然而，他们的分析假设(i)所有的预言机都是相同且一致的噪声，(ii)注释是在一些潜在的真实标签上的一个噪声过程。Donmez等人(2009)通过允许注释器具有不同的噪声级别解决了第一个问题，并表明可以估计真实实例标签和单个oracle质量(只要它们不随时间变化)。它们通过在随后的迭代主动学习中只查询更可靠的注释器来利用这些估计。

There are still many open research questions along these lines. For example, how can active learners deal with noisy oracles whose quality varies over time (e.g., after becoming more familiar with the task, or after becoming fatigued)? How might the effect of payment influence annotation quality (i.e., if you pay a non-expert twice as much, are they likely to try and be more accurate)? What if some instances are inherently noisy regardless of which oracle is used, and repeated labeling is not likely to improve matters? Finally, in most crowdsourcing

在这方面仍有许多悬而未决的研究问题。例如，主动学习者如何处理质量随时间变化的嘈杂的预言(例如，在对任务更加熟悉之后，或者在感到疲劳之后)?付费的效果会如何影响注释的质量(例如，如果你支付给非专家两倍的钱，他们是否可能尝试更准确)?如果无论使用哪种oracle，某些实例都具有固有的噪声，并且重复标记不太可能改善问题，该怎么办?最后，在大多数众包中

[4] http://www.mturk.com

[5] http://www.gwap.com

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 oracles to use. How might the learner continue to make progress?

用户的环境不一定是“按需”可用的，因此对注释器质量的准确估计首先可能很难实现，并且可能永远不会再适用，因为模型对于使用哪个oracle没有真正的选择。学习者如何继续取得进步?

6.3 Variable Labeling Costs

Continuing in the spirit of the previous section, in many applications there is variance not only in label quality from one instance to the next, but also in the *cost* of obtaining that label. 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 labeling of query instances 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.

6.3可变标签成本继续前一节的精神，在许多应用中，不仅在标签质量上存在差异，而且在获得该标签的成本上也存在差异。如果我们在主动学习中的目标是最小化训练一个准确模型的总成本，那么简单地减少标记实例的数量并不一定保证总标记成本的降低。一种减少主动学习中标注工作量的建议方法是使用当前训练好的模型，通过在结构化学习任务中预先标记查询实例，如解析(Baldridge和Osborne，2004)或信息提取(Culotta和McCallum，2005)。然而，这些方法实际上并不能代表或解释标签成本。相反，它们试图通过最小化已经选择的查询所需的注释操作的数量来间接降低成本。

Another group of cost-sensitive active learning approaches explicitly accounts for varying label costs while selecting queries. 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. Instead of using real costs, however, their experiments make the simplifying assumption that the cost of labeling an instances is a linear function of its length (e.g., one cent per second for voicemail messages). Furthermore, labeling and misclassification costs must be mapped into the same currency (e.g., \$0.01 per second of annotation and \$10 per misclassification), which may not be appropriate or straightforward for some applications. King et al. (2004) use a similar decision-theoretic approach to reduce actual labeling costs. They describe a “robot scientist” which can execute a series of autonomous biological experiments to discover metabolic pathways, 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). But here again, the cost of materials is fixed and known at the time of experiment (query) selection.

另一组成本敏感型主动学习方法在选择查询时明确地考虑了不同的标签成本。Kapoor等人(2007)提出了一种同时考虑标签成本和错误分类成本的决策理论方法。在这种设置中，通过将每个候选查询的标记成本与如果将实例添加到训练集中预计将产生的未来错误分类成本相加来评估每个候选查询。然而，他们的实验没有使用实际成本，而是做了一个简化的假设，即标记一个实例的成本是其长度的线性函数(例如，语音邮件信息每秒一美分)。此外，标注和错误分类的成本必须映射成相同的货币(例如，注释每秒0.01美元，错误分类每秒10美元)，这对于某些应用程序可能不合适或不直接。King等人(2004)使用类似的决策理论方法来降低实际的标签成本。他们描述了一个“机器人科学家”，它可以执行一系列自主的生物实验来发现代谢途径，其目标是最大限度地减少所用材料的成本(即，实验的成本加上未来实验的预期总成本，直到找到正确的假设)。但在这里，材料的成本是固定的，并且在实验(查询)选择时已知。

In all the settings above, and indeed in most 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, when the labeling cost is a function of elapsed annotation time. They 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):

在上述所有设置中，实际上在大多数成本敏感型主动学习文献中(例如，Margineantu, 2005;Tomanek et al., 2007)，注释实例的成本仍然假设是固定的，并且在查询之前学习者已经知道。Settles等人(2008a)提出了一种新的成本敏感主动学习方法，该方法适用于标注成本可变且未知的环境，例如，当标注成本是标注时间的函数时。他们学习一个回归成本模型(与活动任务模型一起)，该模型试图基于实例上的一些简单的“元特征”来预测真实的、未知的注释成本。对使用现实世界人类注释成本的四个数据集的分析揭示了以下内容(Settles等人，2008a):

- In some domains, annotation costs are not (approximately) constant across instances, and can vary considerably. This result is also supported by the subsequent findings of others, working on different learning tasks (Arora et al., 2009; Vijayanarasimhan and Grauman, 2009a).

  在某些领域中，跨实例的注释成本(大致)不是恒定的，可能会有很大的差异。这一结果也得到了其他人在不同学习任务上的后续研究结果的支持(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).

  因此，忽略成本的主动学习方法可能并不比随机选择(即被动学习)更好。

- 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).

  注释实例的成本可能不是固有的，而是根据做注释的人而变化。Ringger et al.(2008)和Arora et al.(2009)的研究结果也支持这一结果。

- 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 that should be shorter under normal circumstances).

  一个注释的测量成本可能包含随机成分。特别是，至少有两种类型的噪声会影响注释速度:抖动(由于注释器疲劳、延迟等引起的小变化)和暂停(在正常情况下应该更短的大变化)。

- 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 more accurate than simple cost heuristics (e.g., a linear function of document length).

  有时，即使只看到几个训练实例，也可以准确地预测未知的注释成本。Vijayanarasimhan和Grauman (2009a)的研究结果也支持这一结果。此外，这些学习成本模型比简单的成本启发式(例如，文档长度的线性函数)要准确得多。to

While empirical experiments show that learned cost-models can be trained to predict accurate annotation times, 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. show that simply dividing the informativeness measure (e.g., entropy) by the cost is not necessarily an effective

虽然经验实验表明，学习成本模型可以被训练来预测准确的标注时间，但需要进一步的工作来确定这种近似的、预测的标注成本如何被成本敏感的主动学习系统有效地利用。Settles等人表明，简单地将信息量度量(例如，熵)除以成本不一定是有效的

cost-reducing strategy for several natural language tasks when compared to random sampling (even if *true* costs are known). However, results from Haertel et al. (2008) suggest that this heuristic, which they call *return on investment* (ROI), is sometimes effective for part-of-speech tagging, although like most work they use a fixed heuristic cost model. Vijayanarasimhan and Grauman (2009a) also demonstrate potential cost savings in active learning using predicted annotation costs in a computer vision task using a decision-theoretic approach. It is unclear whether these disparities are intrinsic, task-specific, or simply a result of differing experimental assumptions.

与随机抽样相比，几种自然语言任务的成本降低策略(即使真实成本已知)。然而，Haertel等人(2008)的结果表明，这种他们称之为投资回报率(ROI)的启发式方法有时对词性标注是有效的，尽管像大多数工作一样，他们使用固定的启发式成本模型。Vijayanarasimhan和Grauman (2009a)也展示了在使用决策理论方法的计算机视觉任务中使用预测注释成本的主动学习中潜在的成本节约。目前尚不清楚这些差异是内在的、特定任务的，还是仅仅是不同实验假设的结果。

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 6.4) 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 best query strategy differed between the two annotators, in terms of reducing both labeled corpus size and annotation costs. 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.

即使在没有明确推理标注成本的方法中，一些作者也发现替代查询类型(例如标记特征而不是实例，参见第6.4节)可以降低人类预言机的标注成本(Raghavan et al.， 2006; Druck et al.， 2009; Vijayanarasimhan and Grauman，2009a)。有趣的是，Baldridge和Palmer(2009)在一项稀有语言文档研究中使用主动学习进行语素注释，使用两个活的人类预言者(一个专家和一个新手)交互“在循环中”。他们发现，在减少标记语料库大小和注释成本方面，两种注释器之间的最佳查询策略有所不同。领域专家是一个基于不确定性的主动学习器的更有效的预言者，但是半自动化的注释-旨在帮助标记过程-几乎没有帮助。然而，新手使用被动学习器(随机选择段落)时效率更高，但半自动注释在这种情况下是有益的。

6.4 Alternative Query Types

可选查询类型

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. What other forms might a query take?

主动学习中的大多数工作都假定“查询单元”与要学习的目标概念具有相同的类型。换句话说，如果任务是为文本文档分配类标签，则学习者必须查询文档并由oracle提供其标签。查询还可能有哪些其他形式?

Settles et al. (2008b) introduce an alternative query scenario in the context of *multiple-instance active learning*. In multiple-instance (MI) learning, 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

Settles等人(2008b)在多实例主动学习的背景下引入了另一种查询场景。在多实例(MI)学习中，实例被分组到包中(即多集)，并且是包而不是实例被标记为训练。当且仅当一个包的所有实例都为负时，它被标记为负。然而，如果一个包至少有一个实例是阳性的，则该包被标记为阳性(注意，阳性包也可能包含阴性实例)。人工智能学习的naïve方法是将其视为具有片面噪声的监督学习(即，所有负面实例都是真正的负面，但一些正面实例实际上是负面的)。然而，特殊的人工智能学习算法已经被开发出来

将实体放在更大的实体bag中，标记是否存在

转换了查询的类型

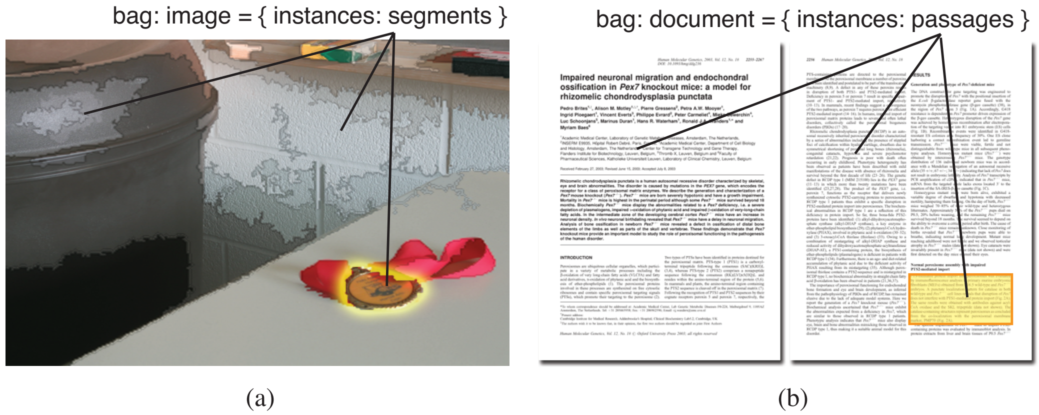

bag: image = { instances: segments }
bag: document = { instances: passages }
(a)
(b)

Figure 9: Multiple-instance active learning. (a) In content-based image retrieval, images are represented as bags and instances correspond to segmented image regions. An active MI learner may query which segments belong to the object of interest, such as the gold medal shown in this image. (b) 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 are representative of the positive class at hand.

多实例主动学习。(a)在基于内容的图像检索中，图像用袋表示，实例对应于分割的图像区域。一个主动的人工智能学习器可以查询哪些部分属于感兴趣的对象，比如这张图中显示的金牌。(b)在文本分类中，文件是袋，实例代表文本段落。在人工智能主动学习中，学习者可以查询特定的段落，以确定它们是否代表了手边的积极类。

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).

从标签袋中学习，尽管模棱两可。Dietterich等人(1997)在药物活性预测的背景下正式确定了MI设置，并已应用于各种各样的任务，包括基于内容的图像检索(Maron和Lozano-Perez, 1998;Andrews et al., 2003;Rahmani and Goldman, 2006)和文本分类(Andrews et al., 2003;雷和克雷文, 2005)。

Figure 9 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 9(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

图9说明了如何将MI表示应用于(a)基于内容的图像检索(CBIR)和(b)文本分类。对于CBIR任务，图像被表示为包，实例对应于图像的分割区域。如果图像包含一些感兴趣的对象，则表示给定图像的袋子被标记为正。MI范例非常适合此任务，因为图像中只有少数区域可能表示感兴趣的对象，例如图9(a)中的金牌。这里MI表示的一个优点是，标记整个图像比标记每个图像段，甚至是图像段的子集要容易得多。对于文本分类任务，文档可以表示为包，实例对应于组成每个文档的短段落(例如，段落)。对于可以免费获得或廉价获得文档标签的分类任务来说，MI表示非常有吸引力

(e.g., from online indexes and databases), but the target concept is represented by only a few passages.

(例如，从在线索引和数据库)，但目标概念仅由少数段落表示。

For MI learning tasks such as these, it is possible to obtain labels both at the bag level and directly at the instance level. Fully labeling all instances, however, 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. In MI active learning, however, the learner is sometimes allowed to query for labels at a finer granularity than the target concept, e.g., querying 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. 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.

对于诸如此类的人工智能学习任务，可以在包级别和直接在实例级别获得标签。然而，对所有实例进行完全标记是非常昂贵的。通常，将学习任务表述为MI问题的基本原理是，它允许我们利用可能以低成本甚至免费获得的粗糙标签。然而，在人工智能主动学习中，有时允许学习者以比目标概念更细的粒度查询标签，例如，查询段落而不是整个文档，或查询分割的图像区域而不是整个图像。Settles等人(2008b)通过逻辑回归的多实例泛化来关注这种类型的混合粒度主动学习。Vijayanarasimhan和Grauman (2009a,b)已经将这一思想扩展到用于图像检索任务的支持向量机，并且还探索了一种在不同粒度和成本水平上交叉查询的方法。

查询特征，而非实体本身

串联学习

Another alternative setting is to query on *features* rather than (or in addition to) instances. Raghavan et al. (2006) have proposed one such approach, *tandem learning*, which can incorporate feature feedback in traditional classification problems. In their work, a text classifier may interleave instance-label queries with feature-salience queries (e.g., "is the word *puck* a discriminative feature for classifying sports documents?"). Values for the salient features are then amplified in instance feature vectors to reflect their relative importance. Raghavan et al. reported that interleaving such queries is very effective for text classification, and also found that words (or features) are often much easier for human annotators to label in empirical user studies. Note, however, that these "feature labels" only imply their discriminative value and do not tie features to class labels directly.

另一个可选的设置是查询特性而不是(或除了)实例。Raghavan等人(2006)提出了一种这样的方法，串联学习，它可以在传统的分类问题中结合特征反馈。在他们的工作中，文本分类器可以将实例标签查询与特征显著性查询(例如，“单词puck是分类体育文档的判别特征吗?”)交织在一起。然后在实例特征向量中放大显著特征的值，以反映它们的相对重要性。Raghavan等人报告说，交错这样的查询对于文本分类非常有效，并且还发现在经验用户研究中，单词(或特征)通常更容易被人类注释者标记。然而，请注意，这些“特征标签”只暗示它们的区别值，而不直接将特征与类标签联系起来。

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). In this line of work, users may specify a set of constraints between features and labels, e.g., "95% of the time, when the word *puck* is observed in a document, the class label is hockey." The learning algorithm then tries to find a set of model parameters that match expected label distributions over the unlabeled pool $\mathcal{U}$ against these user-specified priors (for details, see Druck et al., 2008; Mann and McCallum, 2008). Interestingly, Mann and McCallum found that specifying many imprecise constraints is more effective than fewer 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 these constraints.

近年来，将基于特征的领域知识整合到监督学习和半监督学习中已经发展了几种新的方法(例如，Haghighi and Klein, 2006;Druck et al., 2008)。在这一行工作中，用户可以在特征和标签之间指定一组约束，例如，“95%的情况下，当在文档中观察到单词puck时，类标签是hockey。”然后，学习算法尝试找到一组模型参数，这些参数与用户指定的先验在未标记池U上的预期标签分布相匹配(详细信息，参见Druck et al., 2008;Mann and McCallum, 2008)。有趣的是，Mann和McCallum发现，指定许多不精确的约束比指定更少的更精确的约束更有效，这表明如果有足够多的人指定的特征标签(无论多么嘈杂)是有用的。这就引出了一个问题，即如何积极地寻求这些约束。

Druck et al. (2009) propose and evaluate a variety of active query strategies aimed at gathering useful feature-label constraints. 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 (see Section 3.5). However, this method is intractable for most real-world problems, and they also resort to heuristics in practice. Sindhwani et al. (2009) have also explored interleaving class-label queries for both instances and features, which they refer to as *active dual supervision*, in a semi-supervised graphical model.

Druck等人(2009)提出并评估了各种旨在收集有用特征标签约束的主动查询策略。他们表明，对于两个信息提取任务，主动特征标记比“ 被动 ”特征标记(使用各种强基线)或实例标记(被动和主动)更有效。这些结果适用于模拟和交互人类注释器实验。Liang等人(2009)提出了一种基于贝叶斯实验设计的更具原则性的方法来解决这个问题(见第3.5节)。然而，对于大多数现实世界的问题，这种方法是难以处理的，他们在实践中也诉诸于启发式。Sindhwani等人(2009)也在半监督图形模型中探索了实例和特征的交错类标签查询，他们称之为主动双重监督。

6.5 Multi-Task Active Learning

多个模型公用同一组实体

The typical active learning setting assumes 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 likely most 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 a query with respect to all the learners involved.

典型的主动学习环境假设只有一个学习者试图解决一个任务。然而，在许多实际问题中，相同的数据实例可能会以多种方式标记不同的子任务。在这种情况下，同时为所有子任务标记单个实例可能是最经济的。因此，多任务主动学习算法假设单个查询将被标记为多个任务，并尝试评估与所有涉及的学习者相关的查询的信息性。

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 in tandem. 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 independently, and 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. 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 the single-task active setting.

Reichart等人(2008)研究了自然语言解析和命名实体识别(NER)的双任务主动学习场景，NER是一种信息提取形式。他们提出了两种同时积极学习这两项任务的方法。第一种是交替选择，它允许解析器在一次迭代中查询句子，然后NER系统在下一次迭代中查询实例。第二种是排序组合，即两个学习器对查询池中的候选对象进行独立排序，并选择组合排序最高的实例进行标注。在这两种情况下，不确定性采样被用作每个学习者的基本选择策略。正如人们所期望的那样，这些方法在两个子任务上都优于被动学习，而每个单独子任务的学习曲线不如在单任务主动设置下的学习曲线好。

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, for example, so a simple rank combination might over-estimate the informativeness

Qi等人(2008)研究了一种不同的多任务主动学习场景，在这种场景中，图像可以并行地为几个二元分类任务标记。例如，一张图像可能被标记为包含海滩、日落、山脉、田野等，它们并不都是相互排斥的；然而，它们也不是完全独立的。例如，海滩和日落标签可能高度相关，因此简单的排名组合可能高估了信息量

of some instances. They propose and evaluate a novel Bayesian approach, which takes into account the mutual information among labels.

在某些情况下。他们提出并评估了一种新的贝叶斯方法，该方法考虑了标签之间的互信息。

6.6 Changing (or Unknown) Model Classes

改变模型种类

用一个模型主动学习选出来的数据集，泛用到别的模型上收获了好的结果

As mentioned in Section 4.1, a training set built via active learning comes from a biased distribution, which is implicitly tied to the class of the model used in selecting the queries. This can be an issue if we wish to re-use this training data with models of a different type, or if we do not even know the appropriate model class (or feature set) for the task to begin with. Fortunately, this is not always a problem. For example, 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 MaxEnt model using QBC can be effectively re-used to train CRFs, maintaining cost savings compared with random sampling. Hwa (2001) successfully re-used natural language parsing data selected by one type of parser to train other types of parsers.

如第4.1节所述，通过主动学习构建的训练集来自有偏分布，该分布隐式地与选择查询时使用的模型类联系在一起。如果我们希望用不同类型的模型重用这些训练数据，或者如果我们甚至不知道开始任务的适当模型类(或特征集)，这可能是一个问题。幸运的是，这并不总是一个问题。例如，Lewis和Catlett(1994)表明，决策树分类器仍然可以从使用不确定性采样的主动朴素贝叶斯学习器构建的训练集中获益。Tomanek等人(2007)也表明，使用QBC的MaxEnt模型收集的信息提取数据可以有效地重复用于训练crf，与随机抽样相比，保持了成本节约。Hwa(2001)成功地重用了由一种类型的解析器选择的自然语言解析数据来训练其他类型的解析器。

多个模型选出的复合集也取得了好的结果

However, Baldridge and Osborne (2004) encountered the exact opposite problem when re-using data selected 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 and Bongard (2009) employed active learning with a heterogeneous ensemble of neural networks and decision trees, when the more appropriate model was not known in advance. Their ensemble approach is able to simultaneously select informative instances for the overall model, as well as bias the constituent weak learners toward the more appropriate model class as it learns. Sugiyama and Rubens (2008) have experimented with an ensemble of linear regression models using different feature sets, to study cases in which the appropriate feature set is not yet decided upon.

然而，Baldridge和Osborne(2004)在重用一个解析模型选择的数据来训练各种其他解析器时遇到了完全相反的问题。作为替代方案，它们使用由不同解析器类型组成的异构集成执行主动学习，并且还使用半自动标记来减少人工注释工作。与被动学习相比，这种方法有助于减少每种解析器类型所需的训练示例的数量。同样，Lu和Bongard(2009)在事先不知道更合适的模型的情况下，采用了神经网络和决策树异构集成的主动学习。他们的集成方法能够同时为整个模型选择有信息的实例，并在学习过程中使组成弱学习者偏向于更合适的模型类。Sugiyama和Rubens(2008)使用不同的特征集对线性回归模型的集合进行了实验，以研究尚未确定适当特征集的情况。

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

这一节提出了在实践中主动学习的一个非常重要的问题。如果最好的模型类和特征集恰好是事先已知的，或者如果这些在未来不太可能改变，那么主动学习可能可以安全地使用。否则，随机抽样(至少对于试点研究，直到任务能够被更好地理解)可能比用不合适的学习模型冒险进行主动学习更可取。一种可行的主动方法

seems to be the use of heterogeneous ensembles in selecting queries, but there is still much work to be done in this direction.

似乎是在选择查询时使用异构集成，但是在这个方向上还有很多工作要做。

6.7 Stopping Criteria

A potentially important element of interactive learning applications in general is knowing when to *stop* learning. One way to think about this is the point at which the cost of acquiring new training data is greater than the cost of the errors made by the current model. Another view is how to recognize when the accuracy of a learner has reached a plateau, and acquiring more data is likely a waste of resources. Since active learning is concerned with improving accuracy while remaining sensitive to data acquisition costs, it is natural to think about devising a "stopping criterion" for active learning, i.e., a method by which an active learner may decide to stop asking questions in order to conserve resources.

一般来说，交互式学习应用程序的一个潜在重要元素是知道何时停止学习。考虑这个问题的一种方式是，获取新训练数据的成本大于当前模型所犯错误的成本。另一种观点是如何识别学习者的准确性何时达到了一个平台，获取更多的数据可能是浪费资源。由于主动学习关注的是提高准确性，同时对数据获取成本保持敏感，因此很自然地要考虑为主动学习设计一个“ 停止标准 ”，即主动学习者可能决定停止提问以节省资源的方法。

Several such stopping criteria for active learning have been proposed (Vlachos, 2008; Bloodgood and Shanker, 2009; Olsson and Tomanek, 2009). These methods are all fairly similar, generally based on the notion that there is an intrinsic measure of stability or self-confidence within the learner, and active learning ceases to be useful once that measure begins to level-off or degrade. Such self-stopping methods seem like a good idea, and may be applicable in certain situations. However, in my own experience, the real stopping criterion for practical applications is based on economic or other external factors, which likely come well before an intrinsic learner-decided threshold.

已经提出了几个这样的主动学习停止标准(Vlachos，2008;Bloodgood and Shanker，2009;Olsson和Tomanek，2009)。这些方法都相当相似，通常基于这样一种观念，即学习者有一种内在的稳定或自信的衡量标准，一旦这种衡量标准开始趋于稳定或下降，主动学习就不再有用了。这种自停止方法似乎是一个好主意，可能适用于某些情况。然而，根据我自己的经验，实际应用的真正停止标准是基于经济或其他外部因素，这些因素可能远远早于学习者决定的内在阈值。

7 Related Research Areas

Research in active learning is driven by two key ideas: (i) the learner should not be strictly passive, and (ii) unlabeled data are often readily available or easily obtained. There are a few related research areas with rich literature as well.

7.1 Semi-Supervised Learning

Active learning and *semi-supervised learning* (for a good introduction, see Zhu, 2005b) both traffic in making the most out of unlabeled data. As a result, there are a few conceptual overlaps between the two areas that are worth considering. For example, a very basic semi-supervised technique is self-training (Yarowsky, 1995), in which the learner is first trained with a small amount of labeled data, and

主动学习和半监督学习(有很好的介绍，参见Zhu，2005b)都在最大限度地利用未标记数据。因此，这两个领域之间存在一些值得考虑的概念重叠。例如，一种非常基本的半监督技术是自我训练(Yarowsky，1995)，在这种技术中，学习者首先使用少量标记数据进行训练，并且

then used to classify the unlabeled data. Typically the *most* confident unlabeled instances, together with their predicted labels, are added to the training set, and the process repeats. A complementary technique in active learning is uncertainty sampling (see Section 3.1), where the instances about which the model is *least* confident are selected for querying.

然后用于对未标记数据进行分类。通常，最可靠的未标记实例，连同它们的预测标签，被添加到训练集中，然后重复这个过程。主动学习中的一个补充技术是不确定性采样(见3.1节)，其中选择模型最不自信的实例进行查询。

多个模型一致的预测可以用于训练，分歧大的挑出来查询

Similarly, co-training (Blum and Mitchell, 1998) and multi-view learning (de Sa, 1994) use ensemble methods for semi-supervised learning. Initially, separate models are trained with the labeled data (usually using separate, conditionally independent feature sets), which then classify the unlabeled data, and "teach" the other models with a few unlabeled examples (using predicted labels) about which they are most confident. This helps to reduce the size of the version space, i.e., the models must agree on the unlabeled data as well as the labeled data. Query-by-committee (see Section 3.2) is an active learning compliment here, as the committee represents different parts of the version space, and is used to query the unlabeled instances about which they do *not* agree.

同样，共同训练(Blum and Mitchell, 1998)和多视图学习(de Sa, 1994)使用集成方法进行半监督学习。最初，用标记的数据训练单独的模型(通常使用单独的、条件独立的特征集)，然后对未标记的数据进行分类，并用一些未标记的例子(使用预测的标签)“教导”其他模型，它们对哪些是最自信的。这有助于减少版本空间的大小，也就是说，模型必须在未标记数据和标记数据上达成一致。按委员会查询(参见3.2节)在这里是一种积极的学习恭维，因为委员会代表版本空间的不同部分，并用于查询他们不同意的未标记实例。

Through these illustrations, we see that active learning and semi-supervised learning attack the same problem from opposite directions. While semi-supervised methods exploit what the learner thinks it knows about the unlabeled data, active methods attempt to explore the unknown aspects[6]. It is therefore natural to think about combining the two. Some example formulations of semi-supervised active learning include McCallum and Nigam (1998), Muslea et al. (2000), Zhu et al. (2003), Zhou et al. (2004), Tür et al. (2005), Yu et al. (2006), and Tomanek and Hahn (2009).

通过这些例子，我们看到主动学习和半监督学习从相反的方向解决同一个问题。半监督方法利用学习者自认为对未标记数据的了解，而主动方法试图探索未知的方面。因此，考虑将两者结合起来是很自然的。半监督主动学习的一些例子公式包括McCallum和Nigam (1998)，Muslea等人(2000)，Zhu等人(2003)，Zhou等人(2004)，t<e:1>等人(2005)，Yu等人(2006)，以及Tomanek和哈恩（2009）。

7.2 Reinforcement Learning

In *reinforcement learning* (Sutton and Barto, 1998), the learner interacts with the world via "actions," and tries to find an optimal policy of behavior with respect to "rewards" it receives from the environment. For example, consider a machine that is learning how to play chess. In a supervised setting, one might provide the learner with board configurations from a database of chess games along with labels indicating which moves ultimately resulted in a win or loss. In a reinforcement setting, however, the machine actually plays the game against real or simulated opponents (Baxter et al., 2001). Each board configuration (state) allows for certain moves (actions), which result in rewards that are positive (e.g., cap-

在强化学习中(Sutton and Barto, 1998)，学习者通过“行动”与世界互动，并试图找到一个最优的行为策略，考虑到它从环境中获得的“奖励”。例如，考虑一台正在学习如何下棋的机器。在一个有监督的环境中，一个人可能会向学习者提供国际象棋棋局数据库中的棋盘配置，以及指示最终胜负的标签。然而，在强化设置中，机器实际上与真实或模拟对手进行游戏(Baxter et al.， 2001)。每个棋盘配置(状态)都允许特定的移动(行动)，这将导致积极的奖励(例如，上限)

[6] One might make the argument that active methods also "exploit" what is known rather than "exploring," by querying about what isn't known. This is a minor semantic issue.

turing the opponent's queen) or negative (e.g., having its own queen taken). The learner aims to improve as it plays more games.

给对手的皇后取图)或取负(例如，自己的皇后被取走)。学习者的目标是通过玩更多的游戏来提高自己。

The relationship with active learning is that, in order to perform well, the learner must be proactive. It is easy to converge on a policy of actions that have worked well in the past but are sub-optimal or inflexible. In order to improve, a reinforcement learner must take risks and try out actions for which it is uncertain about the outcome, just as an active learner requests labels for instances it is uncertain how to label. This is often called the "exploration-exploitation" trade-off in the reinforcement learning literature. Furthermore, Mihalkova and Mooney (2006) consider an explicitly active reinforcement learning approach which aims to reduce the number of actions required to find an optimal policy.

与主动学习的关系是，为了表现良好，学习者必须积极主动。人们很容易将注意力集中在过去效果良好、但不够理想或不够灵活的行动政策上。为了改进，强化学习器必须冒险并尝试不确定结果的行为，就像主动学习器要求标记不确定如何标记的实例一样。在强化学习文献中，这通常被称为“ 探索-利用 ”权衡。此外，Mihalkova和Mooney(2006)考虑了一种明确的主动强化学习方法，旨在减少找到最优策略所需的操作数量。

7.3 Submodular Optimization

在小集合里添加实体收益，要大于在大集合里添加同样的实体。条件是小集合是大集合的子集

Recently, there has been a growing interest in *submodular functions* (Nemhauser et al., 1978) in machine learning research. Submodularity is a property of set functions that intuitively formalizes the idea of "diminishing returns." That is, adding some instance x to the set $\mathcal{A}$ provides more gain in terms of the target function than adding x to a larger set $\mathcal{A}'$, where $\mathcal{A} \subseteq \mathcal{A}'$. Informally, since $\mathcal{A}'$ is a superset of $\mathcal{A}$ and already contains more information, adding x will not help as much. More formally, a set function F is submodular if it satisfies the property:

最近，人们对机器学习研究中的子模块函数(Nemhauser et al.， 1978)越来越感兴趣。子模块性是集合函数的一个性质，它直观地形式化了“ 收益递减 ”的概念。也就是说，将某个实例x添加到集合A中比将x添加到更大的集合A'中(其中AC A')提供了更多的目标函数增益。非正式地说，因为A'是A的超集，并且已经包含了更多的信息，所以添加x不会有太大的帮助。更正式地说，一个集合函数F是子模的，如果它满足下列性质:

$$F(\mathcal{A} \cup \{x\}) - F(\mathcal{A}) \geq F(\mathcal{A}' \cup \{x\}) - F(\mathcal{A}'),$$

or, equivalently:

$$F(\mathcal{A}) + F(\mathcal{B}) \geq F(\mathcal{A} \cup \mathcal{B}) + F(\mathcal{A} \cap \mathcal{B}),$$

for any two sets $\mathcal{A}$ and $\mathcal{B}$. The key advantage of submodularity is that, for monotonically non-decreasing submodular functions where $F(\emptyset) = 0$, a greedy algorithm for selecting N instances guarantees a performance of $(1 - 1/e) \times F(\mathcal{S}^*_N)$, where $F(\mathcal{S}^*_N)$ is the value of the optimal set of size N. In other words, using a greedy algorithm to optimize a submodular function gives us a lower-bound performance guarantee of around 63% of optimal; in practice these greedy solutions are often within 90% of optimal (Krause, 2008).

对于任意两个集合A和b，子模性的主要优点是，对于F(0) = 0的单调非递减子模函数，选择N个实例的贪心算法保证性能为(1 -1/e) × F(SN)，其中F(S)为大小为N的最优集合的值。换句话说，使用贪心算法优化子模函数可以保证性能的下限约为最优的63%;在实践中，这些贪婪的解决方案往往在最优的90%以内(Krause，2008)。

In learning settings where there is a fixed budget on gathering data, it is advantageous to formulate (or approximate) the objective function for data selection as a submodular function, because it guarantees near-optimal results with signif-

在收集数据有固定预算的学习环境中，将数据选择的目标函数表述(或近似)为子模块函数是有利的，因为它保证了具有显著性的接近最优结果

icantly less computational effort[7]. The relationship to active learning is simple: both aim to maximize some objective function while minimizing data acquisition costs (or remaining within a budget). Active learning strategies do not optimize to submodular functions in general, but Guestrin et al. (2005) show that maximizing mutual information among sensor locations using Gaussian processes (analogous to active learning by expected error reduction, see Section 3.4) can be approximated with a submodular function. Similarly, Hoi et al. (2006b) formulate the Fisher information ratio criterion (Section 3.5) for binary logistic regression as a submodular function, for use with batch-mode active learning (Section 6.1).

大大减少了计算工作量。”与主动学习的关系很简单:两者都旨在最大化某些目标函数，同时最小化数据获取成本(或保持在预算范围内)。主动学习策略通常不会优化到子模函数，但Guestrin等人(2005)表明，使用高斯过程(类似于通过期望误差减少的主动学习，参见第3.4节)最大化传感器位置之间的互信息可以用子模函数近似实现。同样，Hoi等人(2006b)将二元逻辑回归的Fisher信息比率准则(第3.5节)作为子模块函数，用于批处理模式主动学习(第6.1节)。

7.4 Equivalence Query Learning

An area closely related to active learning is learning with *equivalence queries* (Angluin, 1988). Similar to membership query learning (Section 2.1), here the learner is allowed to synthesize queries de novo. However, instead of generating an *instance* to be labeled by the oracle (or any other kind of learning constraint), the learner instead generates a *hypothesis* of the target concept class, and the oracle either confirms or denies that the hypothesis is correct. If it is incorrect, the oracle should provide a counter-example, i.e., an instance that would be labeled differently by the true concept and the query hypothesis.

There seem to be few practical applications of equivalence query learning, because an oracle often does not know (or cannot provide) an exact description of the concept class for most real-world problems. Otherwise, it would be sufficient to create an "expert system" by hand and machine learning is not required. However, it is an interesting intellectual exercise, and learning from combined membership and equivalence queries is in fact the basis of a popular inductive logic game called Zendo[8].

7.4等价查询学习与主动学习密切相关的一个领域是等效查询学习(Angluin，1988)。类似于成员查询学习(第2.1节)，这里允许学习者从头合成查询。然而，不是生成一个实例来由oracle(或任何其他类型的学习约束)标记，而是生成一个目标概念类的假设，并且oracle确认或否认该假设是正确的。如果它是不正确的，oracle应该提供一个反例，即，一个实例将被真实概念和查询假设不同地标记。等价查询学习的实际应用似乎很少，因为oracle通常不知道(或不能提供)大多数实际问题的概念类的确切描述。否则，手工创建一个“专家系统”就足够了，不需要机器学习。然而，这是一个有趣的智力练习，从组合成员和等价查询中学习实际上是一个流行的归纳逻辑游戏Zendo的基础。

7.5 Model Parroting and Compression

用于迁移的模型视作oracle

Different machine learning algorithms possess different properties. In some cases, it is desirable to induce a model using one type of model class, and then "transfer" that model's knowledge to a model of a different class with another set of properties. For example, artificial neural networks have been shown to achieve

7.5模型模仿和压缩不同的机器学习算法具有不同的性质。在某些情况下，需要使用一种类型的模型类来诱导模型，然后将该模型的知识“转移”到具有另一组属性的不同类的模型中。例如，人工神经网络已经被证明可以实现

[7] Many interesting set optimization problems are NP-hard, and can thus scale exponentially. So greedy approaches are usually more efficient.

[8] http://www.wunderland.com/icehouse/Zendo/

better generalization accuracy than decision trees for many applications. However, decision trees represent symbolic hypotheses of the learned concept, and are therefore much more comprehensible to humans, who can inspect the logical rules and understand what the model has learned. Craven and Shavlik (1996) proposed the TREPAN (Trees Parroting Networks) algorithm to extract highly accurate decision trees from trained artificial neural networks (or similarly opaque model classes, such as ensembles), providing comprehensible, symbolic interpretations. Several others (Buciluă et al., 2006; Liang et al., 2008) have adapted this idea to "compress" large, computationally expensive model classes (such as complex ensembles or structured-output models) into smaller, more efficient model classes (such as neural networks or simple linear classifiers).

These approaches can be thought of as active learning methods where the oracle is in fact another machine learning model (i.e., the one being parroted or compressed) rather than, say, a human annotator. In both cases, the "oracle model" can be trained using a small set of the available labeled data, and the "parrot model" is allowed to query the the oracle model for (i) the labels of any unlabeled data that is available, or (ii) synthesize new instances de novo. These two model parroting and compression approaches correspond to the pool-based and membership query scenarios for active learning, respectively.

在许多应用中，比决策树具有更好的泛化精度。然而，决策树代表了学习概念的象征性假设，因此对人类来说更容易理解，人类可以检查逻辑规则并理解模型已经学习了什么。Craven和Shavlik(1996)提出了TREPAN(树木鹦鹉网络)算法，从训练过的人工神经网络(或类似的不透明模型类，如ensembles)中提取高度精确的决策树，提供可理解的符号解释。其他几个人(buciluei et al.， 2006;Liang等人，2008)采用了这一思想，将大型、计算成本高的模型类(如复杂的集成或结构化输出模型)“压缩”为更小、更有效的模型类(如神经网络或简单的线性分类器)。这些方法可以被认为是主动学习方法，其中oracle实际上是另一个机器学习模型(即，被鹦鹉学舌或压缩的模型)，而不是人类注释器。在这两种情况下，“oracle模型”都可以使用一小部分可用的标记数据进行训练，而“鹦鹉模型”被允许查询oracle模型(i)可用的任何未标记数据的标签，或者(ii)从头合成新实例。这两种模型模仿和压缩方法分别对应于主动学习的基于池和成员查询场景。

8 Conclusion and Final Thoughts

Active learning is a growing area of research in machine learning, no doubt fueled by the reality that data is increasingly easy or inexpensive to obtain but difficult or costly to label for training. Over the past two decades, there has been much work in formulating and understanding the various ways in which queries are selected from the learner's perspective (Sections 2 and 3). This has generated a lot of evidence that the number of labeled examples necessary to train accurate models can be effectively reduced in a variety of applications (Section 4).

主动学习是机器学习中一个不断发展的研究领域，毫无疑问，这是由于数据越来越容易或廉价获取，但标记用于训练却越来越困难或昂贵。在过去的二十年里，从学习者的角度出发，在制定和理解选择查询的各种方式方面已经做了很多工作(第2节和第3节)。这已经产生了大量证据，表明在各种应用中，训练准确模型所需的标记示例数量可以有效地减少(第4节)。

Drawing on these foundations, the current surge of research seems to be aimed at applying active learning methods in practice, which has introduced many important problem variants and practical concerns (Sections 5 and 6). So this is an interesting time to be involved in machine learning and active learning in particular, as some basic questions have been answered but many more still remain. These issues span interdisciplinary topics from learning to statistics, cognitive science, and human-computer interaction to name a few. It is my hope that this survey is an effective summary for researchers (like you) who have an interest

在这些基础上，当前的研究浪潮似乎旨在将主动学习方法应用于实践，这引入了许多重要的问题变体和实际问题(第5节和第6节)。因此，这是一个参与机器学习和主动学习的有趣时期，因为一些基本问题已经得到回答，但更多的问题仍然存在。这些问题跨越了跨学科主题，从学习到统计学，认知科学，还有人机交互等等。我希望这个调查对有兴趣的研究人员(像你一样)来说是一个有效的总结

in active learning, helping to identify novel opportunities and solutions for this promising area of science and technology.

在主动学习中，帮助为这一有前途的科学技术领域找到新的机会和解决方案。

Acknowledgements

This survey began as a chapter in my PhD thesis. During that phase of my career, I am indebted to my advisor Mark Craven and committee members Jude Shavlik, Xiaojin "Jerry" Zhu, David Page, and Lewis Friedland, who offered valuable feedback and encouraged me to expand this into a general resource for the machine learning community. My own research and thinking on active learning has also been shaped by collaborations with several others, including Andrew McCallum, Gregory Druck, and Soumya Ray.

The insights and organization of ideas in the survey are not wholly my own, but draw from the conversations I've had with numerous researchers in the field. After putting out the first draft of this document, I received nearly a hundred emails with additions, corrections, and new perspectives, which have all been woven into the fabric of this revision; I thank everyone who took (and continues to take) the time to share your thoughts. In particular, I would like to thank (in alphabetical order): Jason Baldridge, Aron Culotta, Pinar Donmez, Russ Greiner, Carlos Guestrin, Robbie Haertel, Steve Hanneke, Ashish Kapoor, John Langford, Percy Liang, Prem Melville, Tom Mitchell, Clare Monteleoni, Ray Mooney, Foster Provost, Eric Ringger, Teddy Seidenfeld, Katrin Tomanek, and other colleagues who have discussed active learning with me, both online and in person.

References

N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1–9. Morgan Kaufmann, 1998.

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. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 25–32. ACM Press, 2009.

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

D. Angluin. Queries revisited. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 12–31. Springer-Verlag, 2001.

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.

M.F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 45–56. Springer, 2008.

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.

J. Baxter, A. Tridgell, and L. Weaver. Reinforcement learning and chess. In J. Furnkranz and M. Kubat, editors, *Machines that Learn to Play Games*, pages 91–116. Nova Science Publishers, 2001.

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance-weighted active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 49–56. ACM Press, 2009.

M. Bloodgood and V. Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 39–47. ACL Press, 2009.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 92–100. Morgan Kaufmann, 1998.

A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *Proceedings of the IEEE Symposium on the Foundations of Computer Science*, page 330. IEEE Press, 1996.

C. Bonwell and J. Eison. *Active Learning: Creating Excitement in the Classroom*. Jossey-Bass, 1991.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

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.

C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 535–541. ACM Press, 2006.

R. Burbidge, J.J. Rowland, and R.D. King. Active learning for regression based on query by committee. In *Proceedings of Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 209–218. Springer, 2007.

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.

X. Chai, L. Deng, Q. Yang, and C.X. Ling. Test-cost sentistive naive Bayes classification. In *Proceedings of the IEEE Conference on Data Mining (ICDM)*, pages 51–5. IEEE Press, 2004.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):237–304, 1995.

S.F. Chen and R. Rosenfeld. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50, 2000.

D. Cohn. Neural network exploration using optimal experiment design. In *Advances in Neural Information Processing Systems (NIPS)*, volume 6, pages 679–686. Morgan Kaufmann, 1994.

D. Cohn, L. Atlas, R. Ladner, M. El-Sharkawi, R. Marks II, M. Aggoune, and D. Park. Training connectionist networks with queries and selective sampling. In *Advances in Neural Information Processing Systems (NIPS)*. Morgan Kaufmann, 1990.

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

D. Cohn, Z. Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2006.

M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 8, pages 24–30. MIT Press, 1996.

N. Cressie. *Statistics for Spatial Data*. Wiley, 1991.

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.

I. Dagan and S. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 150–157. Morgan Kaufmann, 1995.

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 and D.J. Hsu. Hierarchical sampling for active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 208–215. ACM Press, 2008.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 249–263. Springer, 2005.

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.

V.R. de Sa. Learning classification with unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 6, pages 112–119. MIT Press, 1994.

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.

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.

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

S. Esmeir and S. Markovitch. Anytime induction of cost-sensitive trees. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 425–432. MIT Press, 2008.

V. Federov. *Theory of Optimal Experiments*. Academic Press, 1972.

P. Flaherty, M. Jordan, and A. Arkin. Robust design of biological experiments. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 363–370. MIT Press, 2006.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597, 1998.

C. Gasperin. Active learning for anaphora resolution. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 1–8. ACL Press, 2009.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

R. Gilad-Bachrach, A. Navot, and N. Tishby. Query by committee made real. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 443–450. MIT Press, 2006.

J. Goodman. Exponential priors for maximum entropy models. In *Proceedings of Human Language Technology and the North American Association for Computational Linguistics (HLT-NAACL)*, pages 305–312. ACL Press, 2004.

R. Greiner, A. Grove, and D. Roth. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 139:137–174, 2002.

N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering for image database categorization. In *Proceedings the ACM Workshop on Multimedia Information Retrieval (MIR)*, pages 9–16. ACM Press, 2005.

C. Guestrin, A. Krause, and A.P. Singh. Near-optimal sensor placements in gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 265–272. ACM Press, 2005.

Y. Guo and R. Greiner. Optimistic active learning using mutual information. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 823–829. AAAI Press, 2007.

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. A bound on the label complexity of agnostic active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 353–360. ACM Press, 2007.

A. Hauptmann, W. Lin, R. Yan, J. Yang, and M.Y. Chen. Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of the ACM Workshop on Multimedia Image Retrieval*, pages 385–394. ACM Press, 2006.

D. Haussler. Learning conjunctive concepts in structural domains. *Machine Learning*, 4(1):7–40, 1994.

T. Hofmann and J.M. Buhmann. Active data clustering. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, pages 528—534. Morgan Kaufmann, 1998.

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.

Y. Huang and T. Mitchell. Text clustering with extended user feedback. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 413–420. ACM Press, 2006.

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.

R. Hwa. Sample selection for statistical parsing. *Computational Linguistics*, 30 (3):73–77, 2004.

S. Ji and L. Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40:1474–1485, 2007.

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.

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.

R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan, and A. Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

C. Körner and S. Wrobel. Multi-class ensemble-based active learning. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 687–694. Springer, 2006.

A. Krause. *Optimizing Sensing: Theory and Applications*. PhD thesis, Carnegie Mellon University, 2008.

V. Krishnamurthy. Algorithms for optimal scheduling and management of hidden markov model sensors. *IEEE Transactions on Signal Processing*, 50(6):1382–1397, 2002.

S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 331–339. Morgan Kaufmann, 1995.

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, H. Daume, and D. Klein. Structure compilation: Trading structure for features. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 592–599. ACM Press, 2008.

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.

M. Lindenbaum, S. Markovitch, and D. Rusakov. Selective sampling for nearest neighbor classifiers. *Machine Learning*, 54(2):125–152, 2004.

C.X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 483–486. ACM Press, 2004.

Y. Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Computer Sciences*, 44:1936–1941, 2004.

R. Lomasky, C.E. Brodley, M. Aernecke, D. Walt, and M. Friedl. Active class selection. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 640–647. Springer, 2007.

Z. Lu and J. Bongard. Exploiting multiple classifier types with active learning. In *Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO)*, pages 1905–1906. ACM Press, 2009.

D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

G. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of the Association for Computational Linguistics (ACL)*. ACL Press, 2008.

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.

A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 359–367. Morgan Kaufmann, 1998.

P. Melville and R. Mooney. Diverse ensembles for active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 584–591. Morgan Kaufmann, 2004.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *Proceedings of the IEEE Conference on Data Mining (ICDM)*, pages 483–486. IEEE Press, 2004.

P. Melville, S.M. Yang, M. Saar-Tsechansky, and R. Mooney. Active learning for probability estimation using Jensen-Shannon divergence. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 268–279. Springer, 2005.

L. Mihalkova and R. Mooney. Using active relocation to aid reinforcement learning. In *Proccedings of the Florida Artificial Intelligence Research Society (FLAIRS)*, pages 580–585. AAAI Press, 2006.

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.

T. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.

T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

C. Monteleoni. *Learning with Online Constraints: Shifting Concepts and Active Learning*. PhD thesis, Massachusetts Institute of Technology, 2006.

R. Moskovitch, N. Nissim, D. Stopel, C. Feher, R. Englert, and Y. Elovici. Improving the detection of unknown computer worms activity using active learning. In *Proceedings of the German Conference on AI*, pages 489–493. Springer, 2007.

I. Muslea, S. Minton, and C.A. Knoblock. Selective sampling with redundant views. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 621–626. AAAI Press, 2000.

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.

H.T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 79–86. ACM Press, 2004.

F. Olsson. *Bootstrapping Named Entity Recognition by Means of Active Machine Learning*. PhD thesis, University of Gothenburg, 2008.

F. Olsson. A literature survey of active machine learning in the context of natural language processing. Technical Report T2009:06, Swedish Institute of Computer Science, 2009.

F. Olsson and K. Tomanek. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 138–146. ACL Press, 2009.

G. Paass and J. Kindermann. Bayesian query construction for neural network models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 7, pages 443–450. MIT Press, 1995.

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.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 441–448. Morgan Kaufmann, 2001.

M. Saar-Tsechansky, P. Melville, and F. Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009.

T. Scheffer, C. Decomain, and S. Wrobel. Active hidden Markov models for information extraction. In *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)*, pages 309–318. Springer-Verlag, 2001.

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

M.J. Schervish. *Theory of Statistics*. Springer, 1995.

B. Settles. *Curious Machines: Active Learning with Structured Instances*. PhD thesis, University of Wisconsin–Madison, 2008.

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.

H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the ACM Workshop on Computational Learning Theory*, pages 287–294, 1992.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

V.S. Sheng and C.X. Ling. Feature value acquisition in testing: A sequential batch test algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 809–816. ACM Press, 2006.

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. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Erlbaum, 2007.

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.

R. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

C.A. Thompson, M.E. Califf, and R.J. Mooney. Active learning for natural language parsing and information extraction. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 406–414. Morgan Kaufmann, 1999.

K. Tomanek and U. Hahn. Semi-supervised active learning for sequence labeling. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1039–1047. ACL Press, 2009.

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. *Active Learning: Theory and Applications*. PhD thesis, Stanford University, 2001.

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.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 999–1006. Morgan Kaufmann, 2000.

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.

L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984.

V.N. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

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.

A. Vlachos. A stopping criterion for active learning. *Computer Speech and Language*, 22(3):295–312, 2008.

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.

R. Yan, J. Yang, and A. Hauptmann. Automatically labeling video data using multi-class active learning. In *Proceedings of the International Conference on Computer Vision*, pages 516–523. IEEE Press, 2003.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 189–196. ACL Press, 1995.

H. Yu. SVM selective sampling for ranking with application to data retrieval. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 354–363. ACM Press, 2005.

K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1081–1087. ACM Press, 2006.

C. Zhang and T. Chen. An active learning framework for content based information retrieval. *IEEE Transactions on Multimedia*, 4(2):260–268, 2002.

T. Zhang and F.J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1191–1198. Morgan Kaufmann, 2000.

Z. Zheng and B. Padmanabhan. On active learning for data acquisition. In *Proceedings of the IEEE Conference on Data Mining (ICDM)*, pages 562–569. IEEE Press, 2002.

Z.H. Zhou, K.J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 425–435. Springer, 2004.

X. Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005a.

X. Zhu. Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin–Madison, 2005b.

X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the ICML Workshop on the Continuum from Labeled to Unlabeled Data*, pages 58–65, 2003.

Index