Active Inference and Learning for Classifying Streams

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

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

Abstract

In this position paper we introduce *Active Inference*, a paradigm for intelligently requesting human labels for inference and learning in situations with a finite budget for applying human resources for labeling cases. Many machine learning systems are applied to a stream of instances that can repeat, such as queries entered in a search engine or web pages for potential ad impressions. When a particular instance x can be subject to classification more than once, we have an additional complication to the budgeted learning setting. In such applications, frequently the distributions will be non-uniform; for instance, in the above applications the distributions $p(x)$ over examples are highly skewed and thus a few x's result in a large percentage of the actual cases for prediction. In such settings, it may be beneficial to allocate a human "labeling" budget selectively perform direct inference, requesting human labels on a selected subset of the instances to be provided to an end system in an effort to reduce misclassification cost on the x's with the highest expected utility. In estimating the utility of labeling a particular x, one must consider three factors: misclassification cost, the probability of encountering x, $p(x)$, and the value x and its associated label may bring for (active) learning. We will discuss the illustrative application of machine learning for safe advertising, where there is a limited budget for acquiring ground-truth labels for labeling web-pages.

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

1. Introduction

Active Inference is a general paradigm for applying a limited budget for requesting ground truth label information, which combines ideas from reject inference and active learning and is particularly useful in applications where cases to label are drawn with replacement from a non-uniform distribution. For example, when classifying search queries or web pages for advertising or web pages for filtering, one may see the same instances repeatedly.

Active inference is substantially different from traditional application of human resources in machine learning. Traditionally, ground truth information, often taking the form of feedback from expert labelers, is given to a machine learning system in an initialization process for model induction, or requested via active learning for further fine-tuning of models. This classification model will then estimate the labels for incoming instances; these estimated labels (or probability distributions over them) are provided to an end system that uses the estimates as the basis for a final decision. Active inference differs in that the system can request ground-truth labels directly for inference, by-passing the estimates from the model. Active inference is different as well from the (active) online setting, wherein a stream of instances is presented to a model and modeling procedure that feeds back predictions to some end system. Some subset of these instances are *subsequently* passed off to an expert for ground-truth labeling—information that is used by the modeling procedure for updating the model.

Active Inference differs from both the traditional batch learning and online learning settings by allowing information requests to be made with the explicit intent of performing direct inference. As we demonstrate next, this paradigm can offer substantial benefits in settings where instances, x, are drawn with replacement from some distribution, $p(x)$, particularly in cases where $p(x)$ is highly skewed, such that some instances appear much more frequently than others. Additionally, Active Inference is valuable in scenarios with a skewed

loss structure governing the costs of mistakes. Here the cost incurred by requesting an explicit "correct" label on an instance may substantially outweigh the expected risk taken by deferring classifications to an imperfect statistical model.

As a motivating example consider the problem of building classifiers for "safe" online advertising: helping advertisers to control the content adjacent to which their advertisements are placed. Certain categories of objectionable content such as hate speech and pornography are at odds with the carefully crafted corporate images associated with most brands. Given a stream of potential impressions, a safe advertising system is tasked with classifying each page as (not) objectionable, thereby allowing or preventing a brand ad to occur on a web page. Of course, the distribution on impressions is highly skewed: some urls appear extremely frequently in the ad stream, while others are effectively unique. Given the cost sensitivity to objectionable content, and to large numbers of falsely blocked good web pages, it is clear that not all urls should warrant equal effort; some urls may be sufficiently sensitive or frequent to have their own hard-set ground truth labels. We currently are developing and applying active inference methods to safe advertising in our work with AdSafe Media.[1]

The contribution of this position paper is to provide an introduction to Active Inference for machine learning systems tasked with classification on streams of instances drawn with replacement from some distribution, $p(x)$, with a limited budget for ground-truth label acquisition. We present (§ 2) a formal definition of Active Inference in a utility optimization framework for binary classification problems. We then present (§ 2.2) details and issues specific to stream-based Active Inference, where the decision engine is exposed to a stream of instances drawn with repetition from $p(x)$. We empirically show the benefits possible through Active Inference (§ 3), where we apply several label acquisition strategies to a simulation based on our motivating safe advertising application. A great deal of prior work has shown the benefits of active learning. We prove (§ 4) that the active inference strategy we introduced is a generalization of traditional uncertainty sampling for active learning. We close the paper with a discussion of the limitations of the introduction to active inference that we have presented in this position paper, serving to frame the directions for future work.

[1] http://www.adsafemedia.com

2. Active Inference on Data Streams

This section presents a proposed, basic formulation of active inference for binary classification. We first present the underlying fundamentals, which lead to a straightforward strategy for the case of pool-based active inference. We then discuss the complications that arise for stream-based active inference, and develop a proposed strategy for deciding the instances in the stream on which to expend our labeling budget.

2.1. Active inference fundamentals

Given an instance, x_i, drawn from some distribution, $p(x)$, with an associated label, y_i, drawn according to some $p(y_i|x_i)$, a standard classification system seeks to predict a posterior probability distribution over the class of x_i (for instance, as objectionable or not) using a predictive statistical model, $\hat{p}(y_i|x_i) = f(x_i)$. Based on this estimated posterior distribution, one can choose a particular classification $\hat{y}_i$, which will have an expected misclassification cost (or loss) L that one typically will want to minimize:

$$L(x_i, \hat{y}_i) = \sum_{y' \in \{0,1\}} \hat{p}(y'|x_i) C(\hat{y}_i, y')$$

Here $C(\hat{y}_i, y')$ is some function that yields the cost of predicting $\hat{y}$ when in fact the true label is y'.

Given a distribution, $p(x)$, from which the x's are drawn, a classification system typically seeks to minimize total expected loss:

$$\int_{x_i} L(x_i, \hat{y}_i) p(x_i)$$

Active Inference extends this traditional statistical classification setting by introducing strategies for direct inference at prediction time. If a given label, y_i, is known by or can be acquired by the classification system in advance, this label can be provided to the end system that uses the instance labels.[2] Acquiring such ground truth labels can come at a cost. For example, in the case of safe advertising, this cost takes the form of human annotation of web pages. While human labels can frequently be acquired at a very low cost, q_i, using micro-outsourcing systems (Sheng et al., 2008) such as Amazon's Mechanical Turk,[3] budget restrictions most likely allow only a small subset of incoming instances to be subject to examination.

To optimize the utilization of a restrictive budget, a

[2] Acquired labels can be provided possibly in combination with the statistically inferred labels as a means for combatting label noise. We will largely ignore label noise for this paper; it is an important complication.

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

system that performs active inference includes a strategy D for querying the oracle directly given a stream of instances. Such a strategy will decide when presented with an example x_i whether to acquire a label for the example or to use the statistical prediction. Strategy complexity can vary (e.g., examples being presented in a stream complicates matters), as we will discuss below. For the sake of simplicity for the moment, let us consider a strategy to be equivalent to the set of examples that will be labeled-if-seen. Let $\mathcal{D}$ be the corresponding power set. In this simple setting, a straightforward statement of a goal for active inference then would be to choose D by:

$$\arg\min_{D\in\mathcal{D}} \int_{x_i} (\mathbb{I}_D(x_i)(q_i + C(y_i, y_i)) + (1 - \mathbb{I}_D(x_i))L(x_i, \hat{y}_i))p(x_i)$$

such that $\sum_i \mathbb{I}_D(x_i)q_i \leq B$, where $\mathbb{I}_D(\cdot)$ is an indicator function for the strategy, equal to 1 when the strategy acquires the label for x_i (in our current oversimplified setting, $x_i \in D$) and 0 otherwise. A minor complication is that one would not know y_i to compute $C(y_i, y_i)$; however, often these costs are formulated to be zero, and otherwise $C(y_i, y_i)$ could be estimated similarly to $L(x_i, y_i)$. We will just assume it to be zero for the rest of this paper. Acquiring the label for instance x_i costs q_i. B is a budget governing the maximum number of label acquisitions the active inference system may make.

Given this (simplified) problem structure, the benefit *per occurrence* of x_i of acquiring y_i (once) is $\beta_i = C(\hat{y}_i, y_i) - q_i$. Of course we do not know y_i until after we have acquired it, so we need to estimate β_i as well:

$$\hat{\beta}_i = L(x_i, \hat{y}_i) - q_i$$

where $\hat{y}_i = \operatorname{argmin}_{y'_i} L(x_i, y'_i)$. A straightforward active inference strategy then is to seek $D \in \mathcal{D}$ that optimizes $\int_{x_i} p(x_i)\hat{\beta}_i$ while adhering to the budget, B. If the examples were presented in a pool, rather than a stream, this would be a straightforward optimization because $p(x_i)$ would reduce to the frequency of x_i in the pool.

2.2. Stream-based active inference

The first complication for stream-based active inference is that $p(x)$ may need to be estimated or updated on-the-fly during the inference process. On-line density estimation is an established field with several existing techniques. While appropriately choosing an estimate for $\hat{p}(x)$ is certainly critical to the performance of an Active Inference strategy, we leave a thorough evaluation of probability estimators and their associated influence on Active Inference for future work. Poisson or Gamma-Poisson may be sufficient for accurate modeling. Alternately, to include the possibility removal of instances from the pool of viable candidates (e.g. the removal of a webpage), a more detailed model is necessary, for instance, the Pareto/NBD framework of (Schmittlein et al., 1987), or the beta-geometric/NBD model in (Fader et al., 2005). Both of these techniques have been successfully employed in lifetime value calculations similar to those used here.

A second complication is that the (reduction in) cost associated with a particular x_i must be extrapolated into the future, and appropriately discounted. A third, related complication is that we will need a framework relating the budget to the stream (and the discounting). Do we have a fixed budget for the future (foreseeable or not)? Do we have a budget-per-unit time? This of course will be application dependent. For the development of the rest of the paper, we deal with these two related complications by assuming that we are given a budget for a given time period (or a budget-per-unit-time), and that we can ignore discounting: either because we really are most concerned with this immediate time period (a "square-wave" discount function), or because the discounting affects $p(x)$ uniformly, so weighting by $p(x)$ implicitly deals with the discounting. In our experience, having a budget for a particular time period is a usual application setting. For example, a business may budget so many dollars per month for human labeling of web pages. Next month there will be a new (possibly different) budget. So let's assume for the rest of this development that we know enough about the rate of seeing examples over the budget period that we can directly translate $\hat{p}(x_i)$ to $\hat{\phi}(x_i)$, the estimated frequency of seeing example x_i over the budget period.

A fourth complication is that in the stream setting we do not actually know the set of x_is that we will see over a particular time period, nor even the total set of (real) x_is that we might actually see. If x_i is a web page described by a bag-of-terms representation (for example), we certainly don't expect to see every possible x_i. Thus it is awkward, and may be ineffective, to treat D simply as a set of examples (as we could in the pool setting discussed briefly above). We would like to take the more general notion of D being a decision strategy that will incorporate $\hat{p}(x)$ (or $\hat{\phi}(x)$) and $\hat{p}(y|x)$ to produce a decision whenever an x_i presents itself: should we spend some of our budget to acquire its label? Let's now discuss this in more depth.

Recall that given the cost structure presented above, the expected benefit of acquiring y_i per occurrence of x_i is $\hat{\beta}_i = L(x_i, \hat{y}_i) - q_i$, where $\hat{y}_i = \underset{y'_i}{\text{argmin}}\ L(x_i, y'_i)$. The expected utility of labeling x_i is then $\mathcal{U}(x_i) = \hat{\beta}_i \hat{\phi}(x_i)$. Similarly to our on-line estimation of $\hat{p}(x)$, as we observe the stream and our models' predictions over the stream, we can estimate the distribution over $\mathcal{U}$. Let $\hat{\psi}(\mathcal{U})$ be our estimated probability (density) function over the different possible expected utilities for the various x_i.

Now we can formulate a proposed general label-acquisition strategy: label all x_i for which $\mathcal{U}(x_i) \geq \tau$. This would have a total expected benefit of

$$\Psi(\tau) = \int_{\tau}^{\infty} \hat{\psi}(\mathcal{U}) d\mathcal{U} \tag{1}$$

Then we can choose τ such that $\Psi(\tau) \leq B$, and aquire labels for any x_i for which $\mathcal{U}(x_i) \geq \tau$.[4] We refer to such a strategy as Expected Utility Maximizing Active Inference.

3. Experimental Validation

The benefits of such a strategy are illustrated by the following simple classification experiment, performed over a set of 35,000 web pages extracted from a stream of real ad impressions. Each url has been hand labeled as to the presence or absence of adult content. This dataset has a class skew of roughly 80 to 1. Predictive modeling was performed by logistic regression on a standard vector space representation of textual content present on web pages. While logistic regression is a popular technique for performing text classification, Active Inference is applicable regardless of the functional form of the base predictor used, so long as the base predictor can output probability estimates.

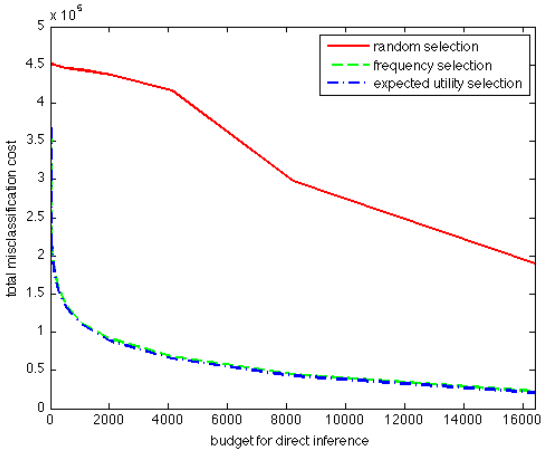
Over ten folds of cross validation, a power law distribution ($\alpha = 2$) was induced on the testing portion of each fold, to simulate the skewed distribution of pages in impression streams for display advertising (corresponding to browser visits to ad-supported web pages). At first blush, predictive models perform quite well at this task of classifying adult content (AUC = 0.97). However, consider cost-sensitive classification, where a misclassification incurs a cost of 1 in the case of labeling a non-adult instance as adult, and 10, 100, or even 1,000 in case where an adult instance is missed. (Assume a cost of 0 for correct classifications.) Given the large number of classifications to make, even a small error rate can lead to significant total cost. Moreover, because instances are drawn with repetition from the above power law distribution, a single instance may be misclassified many times, incurring a large total cost.

Consider three simple strategies for selecting instances for active inference based on a budget of B instances for labeling. For each strategy, instances are selected for gold-standard labeling at a cost. If encountered again in the stream, the instances are classified by direct inference using the already-purchased gold-standard label. Additionally, these labels could optionally be used to improve the training data available to the model, potentially reducing the future error rate (we will return to this below). For examples not seen previously, or for examples without explicit labels, the model is used to perform inference. In order to separate the quality of the active inference selection technique from the technique used for density estimation of $p(x)$, for this experiment, we assume $p(x)$ is estimated perfectly. We assume that the per query labeling cost $q_i = 1$ for all x_i, and that the total budget for labeling is B queries.

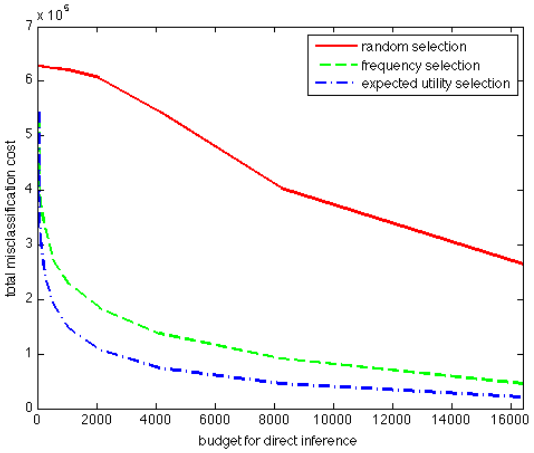
- **Random Sampling** B instances are sampled randomly and assigned correct labels.
- **Frequency Selection** The B most frequently occurring instances are selected for assignment of correct labels.
- **Expected Utility Maximization** The B instances yielding the highest expected label utility, as in Equation 1.

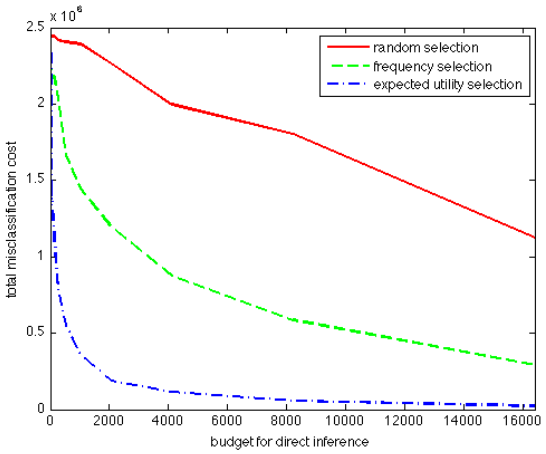
Figure 1 compares these three labeling strategies for varying total budgets. From these plots, we see that "active" inference can have a substantial impact on total misclassification cost. Further, we note the impact an intelligent Active Inference strategy can make. At a misclassification cost of 10 to 1, we see that including information on the distribution of instances, $p(x)$, can have a marked impact on total misclassification cost for a fixed budget; after labeling only 2,000 instances, the techniques incorporating this density information have a total misclassification cost almost $\frac{1}{5}$ of the cost given from random selection. As the costs of misclassification become more significant, incorporating expected label utility makes a notable difference. At 2,000 instances in the 1,000 to 1 cost setting, the expected utility technique has far less than half the total cost of even the frequency-based selection. (And note that this experiment does not incorporate any effect of differences in predictability based on $p(x)$, which may give frequency-based selection an unrealistic advantage—in applications such as those discussed above, one would expect the most frequent pages to be easier to predict.)

[4] We may want also to take into account the variance in $\hat{\psi}(\mathcal{U})$ in order to develop a strategy that will with confidence expend the budget as desired.



(a) 10 to 1





(c) 1,000 to 1

Figure 1. Total incurred costs for different instance selection strategies for different given budgets, B

4. Active Inference and Active Learning

What we have presented so far of course is only half the story. We are interested in applications where we also have to learn the model, and where we will want also to carefully choose the instances that we label for learning. Our decisions on spending our labeling budget should take into account not only the benefit of direct labeling for active inference, but also the benefit of labeling for improving the statistical models that will be used for inference. It may of course be the case that there are separate budgets for active inference and for labeling for learning. However, even in that case the active inference decisions will affect the set of instances from which to learn. Therefore, it seems reasonable to consider how to allocate a single budget so as to get the best overall performance, taking into account both (active) inference and (active) learning.

Our work in progress involves developing and evaluating strategies that try to "optimize" this overall process. A full development is beyond the scope of this short paper. However, let's develop one special case that provides some interesting insight.

Consider strategy presented above—developed particularly for active inference. We conjecture that this also may be an effective active learning strategy, and therefore that the combination of the inference and learning effects may provide a strong baseline against which more sophisticated strategies can be compared. The reason is that this strategy can be considered a form of generalized uncertainty sampling (Saar-Tsechansky & Provost, 2004; Saar-tsechansky & Provost, 2001).

Proposition: *The active inference strategy of selecting the instance(s) x_i with largest values for $\hat{\psi}(x_i)$ selects the same instance(s) as uncertainty sampling under conditions of uniform (estimated) instance frequency, uniform query cost, and uniform error cost.*

Proof: The proof proceeds simply by unwinding the derivation above. Consider the example $\bar{x}$ chosen by the active inference strategy, i.e., the instance with the largest $\hat{\psi}(\mathcal{U})$. If the estimated frequency distribution is uniform, then $\bar{x}$ is the example with the largest $\hat{\beta}_i$. If the acquisition cost is uniform, then this will be the example with the largest $L(x_i, \hat{y}_i)$. Now assume the error costs are uniform, and in particular w.l.o.g. the error cost is 1 if $\hat{y}_i$ is incorrect and 0 if it is correct. Then,

$$L(x_i, \hat{y}_i) = \sum_{y'} \hat{p}(y'|x_i)\mathbb{I}(y' \neq \hat{y})$$

where

$$\hat{y} = \operatorname*{argmin}_{y'} L(x_i, y')$$

and where $\mathbb{I}(\cdot)$ is an indicator function that is 1 if its argument is true and zero otherwise. Now, consider

the examples with $\hat{p}(y_i = 1|x_i) \geq 0.5$ (the derivation is symmetric for $\hat{p}(y_i = 1|x_i) \leq 0.5$). This implies that $\hat{y}_i = 1$, and therefore

$$L(x_i, \hat{y}_i) = \hat{p}(y_i = 0|x_i) \leq 0.5$$

Therefore, $\bar{x}$ will be the example with the largest $\hat{p}(y_i = 0|x_i)$, which will be the instance with $\hat{p}(y_i = 0|x_i)$ (and $\hat{p}(y_i = 1|x_i)$) closest to 0.5. This is exactly the criterion used by uncertainty sampling to rank instances for labeling, and so $\bar{x}$ will also be the instance chosen by uncertainty sampling. □

Under non-uniform distributions, this largest-$\hat{\psi}$ active inference strategy generalizes uncertainty sampling by preferring to label examples, ceteris paribus, if they would be more costly to get wrong, if labeling them is particularly cheap, and/or if they are particularly likely to "reappear." Thus, we could think of the active inference strategy as a cost-sensitive uncertainty sampling strategy, and therefore it has some relation to prior work on weighting uncertainty sampling (Saar-Tsechansky & Provost, 2004) and on cost-sensitive uncertainty sampling (Saar-Tsechansky & Provost, 2004). To our knowledge no prior work has weighted uncertainty sampling by $\hat{p}(x)$.

5. Related Work

There exists a body of work on active inference in networked data, represented for example by (Rattigan et al., 2007; Bilgic & Getoor, 2008). The problem setting of that work is different from that of this paper. Rather than facing a data stream with possibly repeated examples, their examples are interconnected in a network. Via relational statistical models and/or collective inference, labeling some examples in the network can have an impact on the predicted classifications of other nodes in the network. Thus, we face a problem of deciding which nodes to label to maximize performance. A closely related problem setting allows one to influence (at a cost) the actual classes of certain nodes in the network (e.g., by giving special offers), then in networks exhibiting network influence (Aral et al., 2009) we again face the problem of deciding on which nodes to expend our budget to maximize overall performance (Domingos & Richardson, 2001).

The work presented here focuses on data acquisition in high-skew settings (here we specifically emphasize the imbalanced distribution on $p(x)$). In a separate but related vein is the body of work investigating strategies for learning under substantial class skew. This work includes over-sampling the minority class or under-sampling the majority class (Chawla et al., 2002; Liu et al., 2009). A different branch of work investigates the application of non-uniform misclassification costs during training in order to give additional consideration to the class of interest (Domingos, 1999).

There has been some work on active learning on skewed data. Tomanek and Hahn 2009 investigate Query By Committee-based approaches to sampling labeled sentences for the task of named entity recognition. The goal of their selection strategy is to encourage class-balanced selections by incorporating class-specific costs. This work assumes that classifiers can often accurately infer which instances belong to the minority class, giving higher weight to instances thought to belong to the minority class and with a high degree of uncertainty. Our work differs from this by extending to extreme cases where initial performance is poor. Additionally, our techniques are more general, able to extend beyond the tasks faced in NLP.

Bloodgood and Shanker 2009 use a similar approach to (Tomanek & Hahn, 2009), incorporating class specific cost factors to encourage choosing from the minority class. Here the base rate is estimated on a small random sample. We note that in many realistic settings, random samples may not reveal any minority instances, thereby foiling this technique.

Zhu and Hovy 2007 investigate active learning in conjunction with over and under-sampling to alleviate the class imbalance problem. Here active learning is used to choose a set of instances for labeling, with sampling strategies used to improve the class distribution. Our work differs by seeking strategies for acquiring a good class distribution in the data, removing the necessity for performing sub-sampling.

Ertekin et al. 2007 focus on learning with highly imbalanced data sets. Given a large, imbalanced pool of labeled instances, the authors randomly sub-sample instances, choosing to keep only those that are closely positioned to the margin of a SVM classifier. The authors do not address the problem of seeking unlabeled instances in the wild. Furthermore, the margin-based active learning heuristic is very similar to uncertainty sampling, a strategy that we demonstrate to exhibit difficulty in the extremely skewed cases.

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

nificant imbalance in acquisition costs.

Online active learning is concerned with selecting instances for labeling from a stream. The labeled instances are incorporated into a classifier that is applied to the subsequent stream. Helmbold and Panizza 1997 first looked at the tradeoffs between the cost of errors and the costs of labels in online active learning. Subsequently there have been several proposed techniques for "label efficient" techniques including the b-sampling technique in (Bianchi et al., 2006). A similar sampling technique incorporating a logistic model of the confidence has been proposed by Sculley 2007. This work also proposed an approach not based on sampling, where labels are requested whenever the confidence is less than some threshold. These online active learning techniques have a different focus than the work presented here, as they are concerned with the iterative improvement of a base classifier exposed to a stream of instances, and do not incorporate labels for direct inference, nor is duplicity in the instance stream explicitly accounted for.

While none of the afore mentioned active learning techniques make explicit use of the distribution on instances, there is a set of so called "density sensitive" active learning techniques. These techniques are concerned with leveraging the diversity of the instance space, $p(x)$, where the assumption is each instance is seen only once. This exploration is generally done to explore the input space in an effort to overcome the cold-start problem faced by active learners. This problem has been examined by Zhu et al. 2008, work extended by Donmez and Carbonell 2008.

Nguyen and Smeulders 2004 present a framework for incorporating density information into an active learner. This is done through a local density-based label propagation model. This label-propagation can then be incorporated into a more traditional active learner, avoiding repeated labeling within clusters.

This branch of research seeks to find "clusters" of distinct content among the unlabeled instances. Because this family of techniques does not examine duplicity in the occurrences of instances, nor is there any facility for direct inference, density sensitive active learning only bares a passing similarity to the work presented here.

Incorporating a reject option into classification systems has been studied extensively. This work allows a classifier to "reject" those instances with a high expected label cost, or a high uncertainty. Our work can be thought of as a special case of this reject option, where instead of deferring a label, a explicit label is requested from an oracle, and potentially incorporated into the classifier's training set. Examples of research focused on classification with a reject option include (Fumera et al., 2003), who develop analysis specific to text categorization. Bartlett and Wegkamp 2008 present a cost-sensitive technique where a convex loss function similar to hinge loss is optimized. In general, this branch of research does not explicitly consider repeated draws from an underlaying distribution, $p(x)$, and is based on the assumption that it is often less costly to not label an instance than to label an instance incorrectly.

6. Conclusions, Limitations, and Future Work

The main contribution of this paper was to introduce the notions of active inference for data streams, and to present some preliminary results that hopefully will motivate future work. Active inference is a real problem faced in applications such as our running example of safe advertising.

We've demonstrated in this paper that active inference has the potential to reduce the cost of inference substantially under skewed example frequency and cost distributions. Furthermore, we showed that the intuitive active inference strategy we introduced turns out to be a generalization of traditional uncertainty sampling. Therefore, we conjecture that it will be a solid baseline against which to compare more sophisticated active inference strategies as they are developed.

Our introductory treatment so far oversimplifies the problem substantially, and its limitations provide a fallow field for future research:[5]:

- In realistic applications we do not know $p(x)$, the distribution from which the examples are draw with replacement. This distribution must be estimated on the fly, at the same time one is performing (active) inference and learning.
- Because we are estimating $p(x)$ on the fly, it may be the case that for a particular x we use the model to classify it for a while, and then eventually acquire its label. When are we certain enough to label, rather than wait "one more" time?
- Furthermore, many realistic settings have a dynamic $p(x)$: new instances appear not only because they may extremely rare for a fixed $p(x)$ and unseen up to now, but $p(x)$ also tends to change, bringing new instances into the system, and altering the frequencies with with known instances are encountered. In some cases, the dynamics of this distribution can be abrupt, with instances rising

[5]Some of these are mentioned in the text above.

rapidly in popularity: e.g., new popular web pages or new popular search terms (e.g., "volcano iceland"). Distribution estimation techniques need to take this into account.

- The budget needs to be managed over the stream, trading off several competing desires. Labeling pages early both maximizes the value of those particular labels and maximizes the value to model induction. Labeling later allows better estimation of p(x), and therefore may increase the value of the active inference. Furthermore, different budget frameworks are possible. For example, one may have a fixed budget of B up front or a budget per unit time that gets replenished.
- So far, this work has assumed that the labeling is done by an error-free oracle. However, in reality for the applications we are considering the labeling will be done by humans. Humans are error-prone, and the active inference frameworks and models should take noise in the labels into account explicitly.
- When there is noise in the labels, repeated labeling (Sheng et al., 2008) becomes a strategy that must be considered. This adds wonderful complexity to active inference. There no longer is a clear switch from model-based inference to human-based inference. Now we need to consider the fusion of different evidence, acquired at different costs. The model's estimation could be seen as just another labeling source; for certain examples it may even be more accurate than an average human labeler.
- We showed that our active inference strategy is in fact a generalization of uncertainty sampling. However, despite its remarkably consistent performance, there are many research papers showing improvements to uncertainty sampling. What is the best combined active inference and learning strategy, that manages a labeling budget to give the best utility in the long run?

Our on-going work attempts to address some of these challenges. All-in-all, active inference seems to be a topic that can support a large amount of future work in machine learning and beyond.

References

Aral, Sinan, Muchnik, Lev, and Sundararajan, Arun. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.

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

Bartlett, Peter L. and Wegkamp, Marten H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

Bianchi, Nicolò C., Gentile, Claudio, and Zaniboni, Luca. Worst-case analysis of selective sampling for linear classification. In *J. Mach. Learn. Res.*, volume 7, pp. 1205–1230, 2006.

Bilgic, Mustafa and Getoor, Lise. Effective label acquisition for collective classification. In *KDD '08*, 2008.

Bloodgood, Michael and Shanker, K. Vijay. Taking into account the differences between actively and passively acquired data: the case of active learning with support vector machines for imbalanced datasets. In *NAACL '09*, 2009.

Chawla, Nitesh V., Bowyer, Kevin W., and Kegelmeyer, Philip W. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16:321–357, 2002.

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

Domingos, Pedro and Richardson, Matthew. Mining the network value of customers. In *KDD*, 2001.

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

Ertekin, Seyda, Huang, Jian, Bottou, Leon, and Giles, Lee. Learning on the border: active learning in imbalanced data classification. In *CIKM '07*, New York, NY, USA, 2007.

Fader, Peter S., Hardie, Bruce G. S., and Lee, Ka Lok. "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 24(2):275–284, 2005.

Fumera, Giorgio, Pillai, Ignazio, and Roli, Fabio. Classification with reject option in text categorisation systems. In *In: Proc. 12th International Conference on Image Analysis and Processing. IEEE Computer Society*, pp. 582–587, 2003.

Helmbold, David and Panizza, Sandra. Some label efficient learning results. In *COLT '97: Proceedings of the tenth annual conference on Computational learning theory*. ACM, 1997.

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

Nguyen, Hieu T. and Smeulders, Arnold. Active learning using pre-clustering. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, 2004.

Rattigan, Matthew J., Maier, Marc, Jensen, David, Wu, Bin, Pei, Xin, Tan, Jianbin, and Wang, Yi. Exploiting network structure for active inference in collective classification. In *ICDM Workshops*, 2007.

Saar-tsechansky, Maytal and Provost, Foster. Active learning for class probability estimation and ranking. In *In Proc of the Seventeenth Int Joint Fonf on Artificial Intelligence (IJCAI-2001*, pp. 911–920, 2001.

Saar-Tsechansky, Maytal and Provost, Foster. Active sampling for class probability estimation and ranking. *Machine Learning*, 54 (2):153–178, 2004.

Schmittlein, David C., Morrison, Donald G., and Colombo, Richard. Counting your customers: who are they and what will they do next? *Manage. Sci.*, 33(1):1–24, 1987. ISSN 0025-1909.

Sculley, D. Online active learning methods for fast label-efficient spam filtering. In *Fourth Conf. on Email and AntiSpam*, 2007.

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

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

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

Zhu, Jingbo, Wang, Huizhen, Yao, Tianshun, and Tsou, Benjamin K. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING '08*, 2008.