# Sampling Strategies to Evaluate the Performance of Unknown Predictors

Hamed Valizadegan* Saeed Amizadeh † Milos Hauskrecht ‡

**Abstract**

The focus of this paper is on how to select a small sample of examples for labeling that can help us to evaluate many different classification models unknown at the time of sampling. We are particularly interested in studying the sampling strategies for problems in which the prevalence of the two classes is highly biased toward one of the classes. The evaluation measures of interest we want to estimate as accurately as possible are those obtained from the contingency table. We provide a careful theoretical analysis on sensitivity, specificity, and precision and show how sampling strategies should be adapted to the rate of skewness in data in order to effectively compute the three aforementioned evaluation measures.

## 1 Introduction

One of the important challenges of machine learning and data mining research is the evaluation of models or hypotheses with the help of data. The typical and the easiest approach to this problem is to draw a sample randomly from the underlying sample distribution. However, this approach may not be the best solution if each example in the sample comes at a cost that is associated, for example, with the example extraction or annotation of examples by a human expert. In such a case, one seeks a sampling solution that maximizes the benefits of the sample for the evaluation task, while minimizing its size.

As an example, consider the problem of assessment of the quality of a diagnostic model in Medicine, that decides whether a patient suffers from a certain disease. The typical statistics of interest when building such models are the sensitivity, the specificity and the precision of the model. Assuming the disease prevalence is low, the random selection of patient examples may lead to a large number of negative and a small number of positive examples, which may prevent us from accurately estimating the sensitivity and the precision of the model. To assure an accurate assessment of sensitivity, the random sample would have to be much larger which would make the evaluation process very costly.

*University of Pittsburgh. Email: hamed@cs.pitt.edu
†University of Pittsburgh. Email: saeed@cs.pitt.edu
‡University of Pittsburgh. Email: milos@cs.pitt.edu

The focus of our work is on the analysis of methods for cost-effective evaluation of classification models and their key statistics (e.g. sensitivity, specificity and precision) when the prevalence of the two classes is unequal and highly biased toward one of the classes. Our goal is to design strategies for choosing examples such that they can be used to evaluate accurately a large set of classification models or rules one may want to experiment with, and not just one model. The setting of our problem fits any problem for which we expect to build a benchmark set of examples, programs, images, documents, objects that need to be labeled for evaluating the performance of multiple designs and assess their performance. Here are a few examples:

- Consider the problem of evaluation of many possible diagnostic models in Medicine. Such models may originate from different sources; they can be proposed either by different experts, or automated algorithms, or some other model building processes. We want to get a small set of examples to evaluate as accurately and cost-efficiently as possible all these models and compute their key classification (diagnostic) statistics.

- Assume we are interested in developing an inexpensive test (assay) to accurately diagnose a disease from blood samples. The current state-of-the art is a very expensive but an accurate test. To assess how good the potential inexpensive test or tests are, we use a set of stored and labeled blood samples that can be used to run the new assay and evaluate how well it performs. We are interested in obtaining a small sample to be stored and labeled with an expensive test. The sample should be as small as possible and capable of assessing the quality of any test.

- Assume the goal of a company is to develop a system for classification of satellite (microscope) images. It considers and studies multiple designs. To test and assess the quality of the system a benchmark set of images must be selected and assesses by experts so that different designs can be readily evaluated. The question is how to select a small set of benchmark images sufficient for these evaluations.

Whilst our work bears some resemblance to the active

learning framework, we would like to stress that it is very different in its goals and hence the methodology does not transfer to our task. The key difference is that active learning relies on an underlying model (classification, distribution model) it aims to optimize (or learn) from data, and its sample selection strategy makes an attempt to learn the pre-selected model as efficiently and with as small sample size as possible. Hence, the primary output of the active learning is the model it learns. In our work, we do not pre-select and learn (optimize) the model, instead we want a sample that can help us to evaluate any classification model, regardless of its origin. Therefore, our primary output is the sample itself. Another difference is that we are interested in evaluating every possible model with a broader range of classification statistics (sensitivity, specificity, precision) while the active learning framework typically narrows its focus to only one statistics.

One might argue that active learning framework can be used to accurately label all the unlabeled examples by constructing a very good classifier. However, the active learning framework cannot guarantee a classifier that has controlled generalization ability when the labeling budget is very limited. The only guarantee (to a certain degree) is that the model constructed using active learning procedure is better than the one constructed using the passive selection of examples. Therefore, relying on the labels of examples obtained using a model can be very risky and lead to a poor evaluation of the future models (rules). Moreover, for the problems that we consider in this paper (e.g. diagnosing models in medicine), the construction of an accurate model is a challenging task and subject to future model construction. This continuous effort on model construction has been our main motivation to obtain a sample that allows us to evaluate the future models constructed by different procedures.

We start this paper by analyzing what types of examples we need to include in the evaluation set when the priors of the two classes changes. To get some insights to the types of analysis in this paper, let us consider two extreme cases: one that has classes of balanced sizes and the other one that is very unbalanced. Intuitively, when the two classes are balanced, a random selection of examples is likely to perform well in calculating all the interesting evaluation measures (i.e. precision, sensitivity, and specificity). However, a random selection of examples will lead to a poor evaluation of precision and sensitivity of the rules when the classes are unbalanced (suppose the positive class is the minority class). This is because the number of positive examples in the sample will be very limited and the overlap of these positive examples and the positive examples returned by the rule can be very small that leads to high uncertainty in evaluating these rules. The smaller the positive rate of the rule (the number of positive examples returned by the rule) is, the smaller this overlap and the higher the uncertainty will be. Therefore, intuitively, the rule of thumb is to sample more positive examples when the positive rate of the rule is small or the positive class is rare. However, the question is how qualitatively the positive rate of the rule and the unbalance ratio of classes on one hand and the number of positive examples on the other hand are related when we aim to obtain a good estimation of all aforementioned measures. Looking at it from a reverse perspective, the question we answer in this paper is what is our confidence in evaluating the rules of specific form for a specific problem (that has a particular unbalance ratio of classes) using a given sampling? Our analysis in this paper addresses this question. We show that when the prior of the classes changes, the prior in sampling should change with a rate proportional to the unbalance ratio of the classes. This finding is very different from recent work in the active learning area by [1] that advocates the approach in which the examples from the two classes are balanced.

The contributions of this work are:

- We derive bounds on the estimation error of computing precision, sensitivity, and specificity when a small sample size is used to compute these evaluation measures.
- Using these bounds, we show that the optimum sampling strategy is dependent on the skewness rate of the data and positive rate of the rule. One result of our analysis is that when the data is unbalanced, it is better to sample more from the minority class.
- The bounds also suggest that there is a tradeoff between accurate estimate of precision and sensitivity on one side and specificity on the other side. We show how different samplings from the two classes influence this tradeoff by studying the interaction between these bounds.
- We discuss how the guidelines induced by our analysis can be generalized to the case of evaluating rules by the area under the ROC curve (AUC).
- We discuss how sampling from minority class helps to increase the sample size by introducing blind labeling of examples to the majority class. Our discussion assumes a fixed and limited labeling budget.
- Since our analysis suggests we should sample from the minority class, an open question is how to accomplish this task. We propose practical ways to follow the guideline induced by the analysis. In particular, we suggests four different approaches one might utilize to sample from the minority class.
- We verify the efficacy of the proposed method on two UCI data sets by comparing the evaluation results of random sampling and sampling strategy suggested by the analysis.

This paper is organized as follows: We start with the related work in Section 2. Section 3 analyzes how the number of positive and negative examples in the labeled set influences the estimates of the three common evaluation measures: the sensitivity, the specificity and the precision. In Section 3.2 we derive bounds on the estimation errors for these measures. Then (in Section 3.3) we show that a good sampling procedure must trade-off the estimation accuracies of all these statistics and how this translates to controlling the rate of positive examples in the labeled set. Section 3.4 considers the case when some of the statistics and the quality of their estimates are more important than the estimates of other statistics and in Section 3.5 we describe how the current analysis can be adapted in order to obtain a good estimation of area under the ROC curve (AUC). In Section 4, we describe the practical approaches to implement the guidelines suggested by the analysis in Section 3. Section 5 studies how the methodology works in practice on two different data sets. We conclude the work in Section 6.

## 2 Related Work

The objective of this work is to select a set of unlabeled examples from a large data set such that when they are labeled by an expert they can be used to assess accurately multiple classification statistics for many (apriori unknown) classification models. To the best of our knowledge, there is no prior work on this problem. In the following we briefly review prior work in active learning, guided learning, and active risk estimation, three research topics most relevant to our problem.

**2.1 Active learning** The objective of active learning is to improve the process of learning a model while restricting the training sample size [14, 11]. The methods are motivated by the fact that example labeling is a costly and time-consuming task and that the number of examples we need to label to learn the model can be significantly decreased by choosing the most informative examples. To select the most informative examples, active learning is usually reduced to optimization of a certain criterion. Examples of such criteria include but are not restricted to: minimizing the size of version space for SVM [14], minimizing the variance of the estimated generalization error [3], minimizing the generalization error[10, 4] and minimization of empirical one-step look ahead classification risk [17]. The difference from our problem is that active learning methods optimize the model, while we want to find and optimize the sample itself.

One of our objectives is to analyze highly unbalanced data sets. The problem of active learning in presence of extremely unbalanced data set was recently considered by [1, 2, 13]. To address the problem the authors in [1] suggested to consider equal number of examples from two classes to help the learning classes. Our analyses in this paper show that examples in the minority class are more important. Hence this is a very different conclusion than the one proposed by [1] for the learning problem.

Another research work relevant to this paper is pool-based active learning [15] where the objective is to label all the examples in the pool as accurately as possible by constructing a model that classifies all unlabeled examples in the pool with a high confidence. This is unlike the regular active learning where the objective is to construct an accurate model able to classify unseen examples. Although pool-based active learning has some similarities to our work, our objective and tools are different; first we are looking for sampling guidelines effective for the purpose of evaluation, and second, we have a limited budget and cannot make a confident model to classify accurately all the examples in the pool. Notice that the finite-pool active learning is basically designed to simplify the task of reviewers screening a huge corpus of reports to find the relevant documents and it is assumed that enough budget (yet small when compared to label all examples) is available to label as many examples as needed to construct a good classifier [15].

**2.2 Guided Learning** The second research direction relevant to our problem is Guided Learning [1, 16]. The guided learning framework works by asking an expert (or an algorithm) to provide an example with certain characteristics [9]. It then constructs a model using those examples. Note that this is different from active learning: active learning presents samples to the expert to review and label, while guided learning only provides the expert with a guidance (or criteria to be applied) to select an example or a set of examples, and it is the duty of the expert (or a program) to find these examples. In the simplest form, guided learning is interested in finding the examples that belong to a specific class. When applying guided learning to highly imbalanced problems, the main open questions are how to identify the minority class(es) and how to find examples that belong to them.

Our sample selection framework is similar to guided learning in that it provides recommendations of what type (class) of examples to include in the data. The difference is that we do not attempt to learn (optimize) any specific model with a better sample of labeled examples, instead we want to find the sample itself and use it for evaluation purposes. To address the problem of how to select examples from the minority class, we can utilize the solutions introduced in the guided learning literature. Particularly, we rely either on the human expert feedback and his/her ability to identify the examples in the data set or a collection of surrogate classifiers capable of identifying with a better accuracy examples in the minority class that are then reviewed and assessed by a human.

**2.3 Active Risk Estimation** Active Risk Estimation recently introduced by Sawade et. al. [12] is probably the most relevant work to our study, where the authors study the problem of evaluating the risk of a given model accurately for a given evaluation measure at a given labeling budget. Unlike active risk estimation that selects examples to label in order to reduce the risk of evaluating one known model, our work is much more general in the sense that it does not assume there is only one apriori-known model to evaluate. The generality of our setting makes the task of sample selection much more difficult and does not allow one to have a detailed sampling strategies of what individual examples are beneficial to be selected. Instead, we can only find some hindsight or guidelines on what general sampling strategies are good.

## 3 Sampling for Evaluation

Suppose we have a large sample $S$ of unlabeled examples from two classes. We would like to choose a subset $L \subset S$ of representative examples from $S$ to be labeled by an expert, such that the labeled set would let us evaluate a set of (apriori unknown) classification rules or models (given to us later either by an expert or an algorithm ) as accurately as possible in terms of several evaluation measures. The evaluation measures of interest in our work are the precision, the sensitivity, and the specificity of the classification model, or more generally measures obtained from the contingency table.

A good evaluation set should estimate all aforementioned evaluation measures equally well and with similar accuracy guarantees. But how to select examples in set $L$ to assure this is the case? If classes in $S$ are balanced, we expect a random sample from $S$ will also preserve the balance in $L$, so random sampling is likely to work well. However, a balanced sample is harder to achieve if the classes in $S$ are unbalanced. Interestingly, if the two classes are extremely unbalanced, the accuracy guarantees can be (relatively) easily satisfied for statistics related to the majority class even without seeing labeled examples from this class. What seems like a contradiction can be easily explained as follows: assuming all unlabeled examples take on the majority class label, the overall error from mislabeling the minority class examples is small because their prevalence (prior) is small. In such a case we may prefer $L$ to be purely from the minority class or at least heavily biased towards this class.

In this section, we provide a careful analysis of the relation of the skewness rate (the ratio of positive class to the whole sample size), the size of the labeled data set, and the quality of different evaluation measure estimates.

**3.1 Preliminary** Let $S$ denotes the set of all examples in the data set. We define the *rule* $R : S \mapsto \{+, -\}$ as a labeling function which maps the members of $S$ to either the positive or the negative class. The set of all possible rules is denoted by $\mathcal{R}$. Let $O \in \mathcal{R}$ be a special rule which returns the *true* labelings of the examples in $S$ ("$O$" stands for the Oracle). Throughout the paper, $L \subseteq S$ and $U \subseteq S$ refer to the set of examples in $S$ for which the true labeling are, respectively, known and unknown; thus, $L \cup U = S$, $L \cap U = \phi$. Furthermore, and without loss of generality, we assume that the positive class is the minority class and we denote by $B \in U$ the set of examples that are blindly labeled with the majority class (negative class). For $A \subseteq S$, $R \in \mathcal{R}$ and $c \in \{+, -\}$, we define:

- $|A|$ as the cardinality of set $A$,
- $A_R^c \equiv \{x \in A \mid R(x) = c\}$,
- $\epsilon_R^c(A) \equiv \frac{|A_R^c|}{|A|}$

For instance, $S_O^+$ means the set of all true positive examples in $S$ while $\epsilon_O^+(S)$ is the *rate* of true positives in $S$. It is trivial to see that for any given $A \subseteq S$ and $R \in \mathcal{R}$, we have $A = A_R^+ \cup A_R^-$ and $\epsilon_R^-(A) = 1 - \epsilon_R^+(A)$.

For oracle $O$ and the rule $R \in \mathcal{R}$ ($R \neq O$), the quantities $\epsilon_O^+(S)$ and $\epsilon_R^+(S)$ are of special interest in this work. The former is the rate of true positives in the data set while the latter is called the positive rate of rule $R$. Both of these quantities are dictated by the domain and do not change due to any specific sampling strategy. Notice that $\epsilon_R^+(S)$ is unknown at the time of sampling.

In the following analyses , we first assume the partitioning of S into L and U is known. We then use these results to determine what partitioning yields better estimates of evaluation measures. Table 1 shows the elements of contingency table for a rule $R$ and oracle $O$.

Table 1: Contingency table for rule $R$ and oracle $O$.

| | $S_O^+$ | $S_O^-$ |
|---|---|---|
| $S_R^+$ | $\lvert S_O^+ \cap S_R^+\rvert$ | $\lvert S_O^- \cap S_R^+\rvert$ |
| $S_R^-$ | $\lvert S_O^+ \cap S_R^-\rvert$ | $\lvert S_O^- \cap S_R^-\rvert$ |

Using this contingency table, precision, sensitivity and specificity of rule $R$ over set $S = U \cup L$ are defined as follows:

$$Pr(R) = \frac{|S_O^+ \cap S_R^+|}{|S_R^+|} = \frac{|U_O^+ \cap U_R^+| + |L_O^+ \cap L_R^+|}{|S_R^+|} \tag{3.1}$$

$$Se(R) = \frac{|S_O^+ \cap S_R^+|}{|S_O^+|} = \frac{|U_O^+ \cap U_R^+| + |L_O^+ \cap L_R^+|}{|S_O^+|} \tag{3.2}$$

$$Sp(R) = \frac{|S_O^- \cap S_R^-|}{|S_O^-|} = \frac{|U_O^- \cap U_R^-| + |L_O^- \cap L_R^-|}{|S_O^-|} \tag{3.3}$$

where *Pr*, *Se*, and *Sp* stand for the precision, the sensitivity and the specificity, respectively. We would like to identify characteristics of set $L$ or $U$ (one determines the other) that lead to a good estimation of the desired evaluation measures over the whole set $S$.

**3.2 Estimation Bounds** The exact evaluation of precision, sensitivity, and specificity on $S$ is dependent on both $U$ and $L$; for example, the precision is calculated from $|U_O^+ \cap U_R^+|$ and $|L_O^+ \cap L_R^+|$. For a given rule $R$, the dependency on set $L$ is known and we would like to choose $U$ such that the uncertainty of estimating these evaluation measures for different rules is small. Although the labels in set $U$ are not known, we know that $|U_O^+ \cap U_R^+| \sim B(n,p)$ follows a binomial distribution with parameters $n = |U_O^+|$ and $p = \epsilon_O^+(U)$. Using the properties of binomial distribution, we can compute the expected value $E(|U_O^+ \cap U_R^+|) = |U_O^+|\epsilon_O^+(U)$ and variance $\sigma^2(|U_O^+ \cap U_R^+|) = |U_R^+|\epsilon_O^+(U)(1-\epsilon_O^+(U))$. Given that the labels of examples in $L$ are known, the expected value and variance of precision for different labelings on $U$ for a specific but unknown rule $R$ of size $|S_R^+|$ is defined as follows:

$$E(Pr) = \frac{E(|U_O^+ \cap U_R^+|)}{|S_R^+|} + \frac{|L_O^+ \cap L_R^+|}{|S_R^+|} = \frac{|U_R^+|\epsilon_O^+(U)}{|S_R^+|} + \frac{|L_O^+ \cap L_R^+|}{|S_R^+|} \tag{3.4}$$

(3.5)

$$\sigma^2(Pr) = \frac{\sigma^2(|U_O^+ \cap U_R^+|)}{|S_R^+|^2} = \frac{|U_R^+|\epsilon_O^+(U)(1-\epsilon_O^+(U))}{|S_R^+|^2}$$

Given these statistics, we can obtain the concentration bounds for the values of precision and study the variation of precision when the selection of samples in $U$ changes. Using Bernstein bound[1][5], we can bound the deviation of the expected value of precision for unknown labeling of $U$ from the true precision as follows:

$$P\left(|Pr - E(Pr)| > \xi_{Pr}\right) \leq 2\exp\left(-\frac{\xi_{Pr}^2}{4\sigma^2(Pr)}\right) = 2\exp\left(-\frac{\xi_{Pr}^2|S_R^+|^2}{4|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)}\right) \tag{3.6}$$

where $\xi_{pr}$ is a positive value. The above equation simply explains that the value of precision is concentrated around its mean with a density dependent on the properties of oracle $O$ (the uncertainty of labeling on $U$) and the rule $R$ (e.g. the positive examples returned by the rule). We can rewrite the above bound in the probably approximately correct (PAC) form as given in the following proposition.

PROPOSITION 3.1. *For any $\delta \in [0,1]$, the difference between the true precision and the expected value of precision defined by Equation 3.4 satisfies:*

$$|Pr - E(Pr)| \leq 2\frac{\sqrt{|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)\log(\frac{2}{\delta})}}{|S_R^+|} \tag{3.7}$$

*with probability at least $1-\delta$.*

[1]Notice that the bound we use here is the Bernstein bound for small values of $\xi_{Pr}$ relative to $\sigma^2$. Check [5] for details.

*Proof.* The proof is standard and as follows. By defining $\delta = 2\exp\left(-\frac{\xi_{Pr}^2|S_R^+|^2}{4|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)}\right)$, solving it for $\xi_{Pr}$, and replacing it in Equation 3.6, we get:

$$P\left(|Pr - E(Pr)| > 2\frac{\sqrt{|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)\log(\frac{2}{\delta})}}{|S_R^+|}\right) \leq \delta$$

or

$$P\left(|Pr - E(Pr)| \leq 2\frac{\sqrt{|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)\log(\frac{2}{\delta})}}{|S_R^+|}\right) > 1-\delta$$

which is equivalent to the claim of the proposition.

**Remark I:** Notice that in order to get a better estimation of precision, we need to have small values on the right hand side of Equation 3.7. The right hand side of Equation 3.7 is inversely related to $|S_R^+|$ that implies we have a better approximation of precision for the rules with large $|S_R^+|$. Moreover, the numerator shows how the selection of a particular labeled set $L$ affects this uncertainty: the direct relation of numerator to $|U_R^+|$ and $\epsilon_O^+(U)(1-\epsilon_O^+(U))$ implies that smaller values of $|U_R^+|$ and $\epsilon_O^+(U)(1-\epsilon_O^+(U))$ are preferable. To get smaller value for $\epsilon_O^+(U)(1-\epsilon_O^+(U))$, $U$ should become very unbalanced because smaller value for $\epsilon_O^+(U)$ or $1-\epsilon_O^+(U)$ leads to smaller value in $\epsilon_O^+(U)(1-\epsilon_O^+(U))$. Since the uncertainty is also inversely related to the value of $|U_R^+|$ (look at the numerator on the right hand side), more positive examples in $L$ are required to reduce the right hand side of Equation 3.7 and obtain a tighter bound. As a summary, by labeling more positive examples, we reduce the number of positive examples left in $U$ and obtain smaller uncertainty when estimating the precision.

Similarly, we can compute the mean and the variance for sensitivity and specificity as follows:

$$E(Se) = \frac{|U_R^+|\epsilon_O^+(U)}{|S_O^+|} + \frac{|L_O^+ \cap L_R^+|}{|S_O^+|} \tag{3.8}$$

$$\sigma^2(Se) = \frac{|U_R^+|\epsilon_O^+(U)(1-\epsilon_O^+(U))}{|S_O^+|^2} \tag{3.9}$$

$$E(Sp) = \frac{|U_R^-|(1-\epsilon_O^+(U))}{|S_O^-|} + \frac{|L_O^+ \cap L_R^-|}{|S_O^-|} \tag{3.10}$$

$$\sigma^2(Pr) = \frac{|U_R^-|\epsilon_O^+(U)(1-\epsilon_O^+(U))}{|S_O^-|^2} \tag{3.11}$$

Consequently, we can obtain the bounds for sensitivity and specificity:

(3.12)

$$P\left(|Se - E(Se)| > \xi_{Se}\right) \leq 2\exp\left(-\frac{\xi_{Se}^2|S_O^+|^2}{4|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)}\right)$$

(3.13)

$$P\left(|Sp - E(Sp)| > \xi_{Sp}\right) \leq 2\exp\left(-\frac{\xi_{Sp}^2|S_O^-|^2}{4|U_R^-|\epsilon_O^+(U)\epsilon_O^-(U)}\right)$$

The following propositions show how $L$ affects the accuracy of sensitivity and specificity estimates.

PROPOSITION 3.2. *For any $\delta \in [0,1]$, the difference between the true sensitivity and the expected value of sensitivity defined by Equation 3.8 satisfies:*

$$|Se - E(Se)| \leq 2\frac{\sqrt{|U_R^+|\epsilon_O^+(U)\epsilon_O^-(U)\log(\frac{2}{\delta})}}{|S_O^+|} \tag{3.14}$$

*with probability at least $1-\delta$.*

PROPOSITION 3.3. *For any $\delta \in [0,1]$, the difference between the true specificity and the expected value of specificity defined by Equation 3.10 satisfies:*

$$|Sp - E(Sp)| \leq 2\frac{\sqrt{|U_R^-|\epsilon_O^+(U)\epsilon_O^-(U)\log(\frac{2}{\delta})}}{|S_O^-|} \tag{3.15}$$

*with probability at least $1-\delta$.*

The proofs of Propositions 3.2 and 3.3 are similar to the proof of Proposition 3.1.

**Remark II:** The interpretation of the bound for sensitivity is similar to the bound on the precision; to improve the sensitivity estimate we need more positive examples in L. The interpretation of the specificity bound is different: in order to minimize the right hand side of the bound for specificity given in Equation 3.15, more negative examples need to be collected in $L$. This is because smaller values of $|U_O^-|$ and $\epsilon_O^+(U)(1-\epsilon_O^+(U)$ ask for sampling more negative examples in $L$. Thus the goals for improving the estimates of the precision and sensitivity on one side and specificity on the other are opposite. In general, decreasing one increases the other side.

PROPOSITION 3.4. *When the budget size $|L|$ is very small compared to the sample size $|S|$ (i.e. $|L| \ll |S|$), we have the following approximations for almost all rules*[2]*:*

$$|U_R^+| \approx \frac{|U|\epsilon_O^+(U)|S_R^+|}{|S|\epsilon_O^+(S)} \tag{3.16}$$

$$|U_R^-| \approx \frac{|U||S_R^-|(1-\epsilon_O^+(U)))}{|S|(1-\epsilon_O^+(S))} \tag{3.17}$$

*Proof.* First notice that when $|L| \ll |S|$, we have $\epsilon_O^+(U) \approx \epsilon_O^+(S)$. We also have $\epsilon_R^+(U) \approx \epsilon_R^+(S)$ if $\epsilon_O^+(S) \ll \frac{|L|}{|S|}$ and $\epsilon_R^+(S) \ll \frac{|L|}{|S|}$. Therefore:

$$\begin{aligned} |U_R^+| &= |U|\epsilon_R^+(U) \approx |U|\epsilon_R^+(S) = \frac{|U|}{|S|}|S_R^+| \\ &= \frac{|U|\epsilon_O^+(S)}{|S|\epsilon_O^+(S)}|S_R^+| \approx \frac{|U|\epsilon_O^+(U)}{|S|\epsilon_O^+(S)}|S_R^+| \end{aligned} \tag{3.18}$$

[2]Notice that we can explain this in probably approximately correct (PAC) approach. However we preferred not to make this proposition overcomplicated for the practical purposes.

The proof for the second approximation is similar and more intuitive considering that the negative class is majority.

Replacing Equations 3.16 and 3.17 in Equations 3.7, 3.14, and 3.15, we have the following bounds when we use $\epsilon_R^+(S) = \frac{|S_R^+|}{|S|}$:

$$|Pr - E(Pr)| \leq 2\sqrt{\frac{|U|\epsilon_O^+(U)^2\epsilon_O^-(U)\log(\frac{2}{\delta})}{|S|^2\epsilon_R^+(S)\epsilon_O^+(S)}} \tag{3.19}$$

$$|Se - E(Se)| \leq 2\sqrt{\frac{|U|\epsilon_R^+(S)\epsilon_O^+(U)^2\epsilon_O^-(U)\log(\frac{2}{\delta})}{|S|^2\epsilon_O^+(S)^3}} \tag{3.20}$$

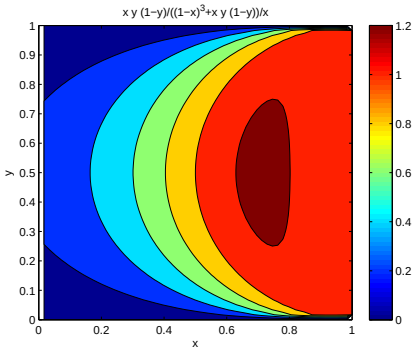
$$|Sp - E(Sp)| \leq 2\sqrt{\frac{|U|\epsilon_R^-(S)\epsilon_O^+(U)\epsilon_O^-(U)^2\log(\frac{2}{\delta})}{|S|^2\epsilon_O^-(S)^3}} \tag{3.21}$$

The above bounds depend on two types of parameters. One group includes $\epsilon_O^+(S)$ and $\epsilon_R^+(S)$ which are the characteristics of the underlying data set and the rule and are not dependent on a particular approach of sampling. We have no control on this group of parameters. $\epsilon_O^+(U)$ is the second type of parameter that depends on the sampling strategy (how we partition $S$ into $L$ and $U$). Thus, any optimization of these bounds involves finding the optimal value of $\epsilon_O^+(U)$.

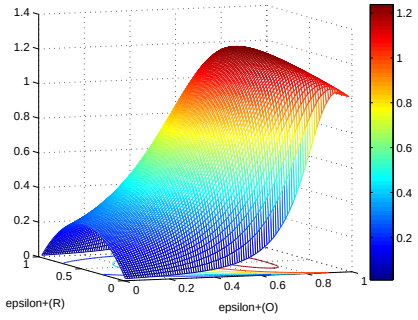
**3.3 Sampling Strategy** As mentioned in the previous subsection, the accuracy of estimates of the three evaluation measures depends on $\epsilon_O^+(U)$, as a variable that is dependent on our sampling strategy. As explained by Remark I and Remark II, we need small values of $\epsilon_O^+(U)$ to have a good estimation of precision and sensitivity and large values of $\epsilon_O^+(U)$ to have a good estimation on specificity. Therefore, getting a good estimation for precision and sensitivity is in contrary to getting a good estimation for specificity. We would like to quantitatively measure how the estimation accuracy of different evaluation measures changes with different number of positive examples in $L$ or $U$ (or equivalently $\epsilon_O^+(L)$ or $\epsilon_O^+(U)$)[3]. To understand the interaction between the estimation accuracy of the three evaluation measures, we study how the changes in the value of $\epsilon_O^+(U)$ trades off between the estimation accuracy of precision, sensitivity, and specificity.

The bounds provided in Equations 3.19, 3.20, and 3.21 depend on several rule and problem specific parameters (i.e. $\epsilon_R^+(S)$, $|S|$, $\epsilon_O^+(S)$, $L$ or $U$, and $\delta$) that makes the task of visualizing the values of these bounds difficult. To address this problem, we study only the values of $\epsilon_O^+(U)$ at which similar upper bounds for different evaluation measures are obtained. By the knowledge of these boundary values, we

[3]Notice that $\epsilon_O^+(L)$ and $\epsilon_O^+(U)$ are closely related; knowing one gives the other one.




(a) Top View




(b) Side View (3D)

Figure 1: $\epsilon_O^+(U)/\epsilon_O^+(S)$ for precision vs. specificity as a function of $\epsilon_O^+(S)$ (the $x$ axis) and $\epsilon_R^+(S)$ (the $y$ axis).

can study at what values we trade the accuracy estimation of one evaluation measures with another. Since we cannot study the trade-off values of $\epsilon_O^+(U)$ for all three evaluation measures at once, we study the change in the value of $\epsilon_O^+(U)$ that trades off precision vs. sensitivity and sensitivity vs. specificity separately, in the following subsections.

**3.3.1 Precision vs. Specificity** To obtain the boundary values of $\epsilon_0^+(U)$ that trades off precision vs. specificity, we need to have similar right hand side for the bounds of these evaluation measures. By making the right hand side of bounds in Equation 3.19 and 3.21 equal, we obtain:

$$\frac{\epsilon_R^+(S)\epsilon_O^+(S)}{\epsilon_O^+(U)} = \frac{(1-\epsilon_O^+(S))^3}{(1-\epsilon_R^+(S))(1-\epsilon_O^+(U))} \Rightarrow$$
$$\frac{(1-\epsilon_O^+(U)}{\epsilon_O^+(U)} = \frac{(1-\epsilon_O^+(S))^3}{(1-\epsilon_R^+(S))\epsilon_R^+(S)\epsilon_O^+(S)} \Rightarrow$$

(3.22)
$$\epsilon_O^+(U) = \frac{(1-\epsilon_R^+(S))\epsilon_R^+(S)\epsilon_O^+(S)}{(1-\epsilon_O^+(S))^3 + (1-\epsilon_R^+(S))\epsilon_R^+(S)\epsilon_O^+(S)}$$

The value of $\epsilon_O^+(U)$ determines the rate of positive examples in set $U$ (and equivalently in set $L$) at which the same estimation accuracy are obtained for precision and specificity. We can also find the exact required number of positive examples with the knowledge of $\epsilon_O^+(S)$, $\epsilon_R^+(S)$, $\epsilon_O^+(U)$, and $|L|$ as follows:

$$|L_O^+| = \min\left(|L|, \left[|S|\epsilon_O^+(S) - |U|\epsilon_O^+(U)\right]_+\right) \tag{3.23}$$

where $[a]_+$ returns $a$ if $a > 0$; otherwise it returns zero. Although the above equation provides the required number of positive examples, we need a more qualitative criteria, independent of the budget size $|L|$, to study the sampling strategy for different values of $\epsilon_O^+(S)$ and $\epsilon_R^+(S)$. It is easy to see that $\epsilon_O^+(U) < \epsilon_O^+(S)$ is an indicator to get more positive examples in $L$, and $\epsilon_O^+(U) > \epsilon_O^+(S)$ is an indicator to get more negative examples in $L$. Therefore, we define quantity $\alpha$ as a variable determining which class of examples is preferable in $L$.

$$\alpha = \frac{\epsilon_O^+(U)}{\epsilon_O^+(S)} \tag{3.24}$$

The magnitude of $\alpha$ is inversely related to the required number of positive examples in $L$; i.e. a small value for $\alpha$ asks for more positive examples while a large value for $\alpha$ asks for more negative examples in $L$. Figure 1 shows the value $\alpha$ as a function of $\epsilon_O^+(S)$ and $\epsilon_R^+(S)$, suggested by Equation 3.22. This figure suggests the following results: 1) We expect to reduce the rate of positive examples in set $U$ compared to that in set $S$ for small values of $\epsilon_O^+(S)$. This is equivalent to say that more positive examples in $L$ are required. 2) We expect to get more positive examples in $L$ for rules with very small or large rate $\epsilon_R^+(S)$.

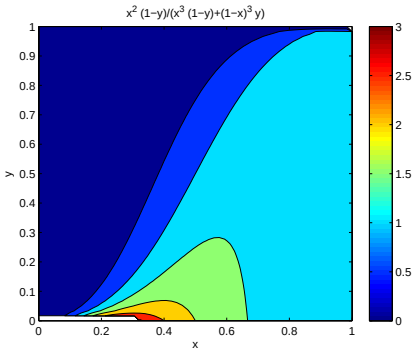
**3.3.2 Sensitivity vs. Specificity** Similarly, the boundary values to trade off sensitivity vs. specificity can be obtained:

$$\frac{\epsilon_O^+(S)^3}{\epsilon_R^+(S)\epsilon_O^+(U)} = \frac{(1-\epsilon_O^+(S))^3}{(1-\epsilon_R^+(S))(1-\epsilon_O^+(U)} \Rightarrow$$
$$\frac{(1-\epsilon_O^+(U)}{\epsilon_O^+(U)} = \frac{(1-\epsilon_O^+(S))^3\epsilon_R^+(S)}{\epsilon_O^+(S)^3(1-\epsilon_R^+(S))} \Rightarrow$$

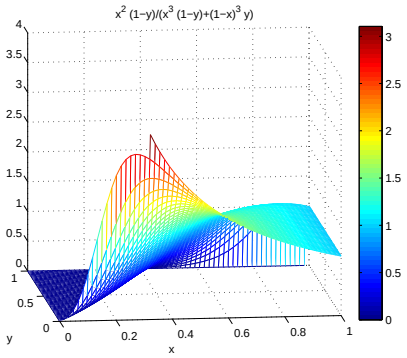
(3.25)
$$\epsilon_O^+(U) = \frac{\epsilon_O^+(S)^3(1-\epsilon_R^+(S))}{\epsilon_O^+(S)^3(1-\epsilon_R^+(S)) + (1-\epsilon_O^+(S))^3\epsilon_R^+(S)}$$

Figure 2 shows the value $\alpha$ as a function of $\epsilon_O^+(S)$ and $\epsilon_R^+(S)$ for sensitivity vs. specificity, suggested by Equation 3.25. Notice that in order to guarantee good bounds on both sensitivity and specificity, we need more positive examples in $L$ if $\epsilon_O^+(S)$ is small compared to $\epsilon_R^+(S)$. Moreover, compared to the discussion in the previous subsection, here the small value for $\epsilon_O^+(S)$ is relative and the concept covers a wider domain of situations.

**3.3.3 Special case of $\epsilon_R^+(S) \approx \epsilon_O^+(S)$** In this section, we give a comparison of precision, sensitivity, and specificity for




(a) Top View




(b) Side View (3D)

Figure 2: $\alpha$ for sensitivity vs. specificity as a function of $\epsilon_O^+(S)$ (the $x$ axis) and $\epsilon_R^+(S)$ (the $y$ axis).

rules that have the specific form of $\epsilon_R^+(S) \approx \epsilon_O^+(S)$ i.e. rules that have the same rate in returning positive examples as oracle. This set of rules are particularly interesting because $\epsilon_R^+(S)$ does not provide any information about the quality of the rule $R$. To see this, notice that if $\epsilon_R^+(S) \gg \epsilon_O^+(S)$, we can make the general statement that $R$ has a lot of false positive alarms and a small number of false negative alarms in average. On the other hand, if $\epsilon_R^+(S) \ll \epsilon_O^+(S)$, there should be a lot of false negative alarms and in average a small number of false positive alarms. So the case $\epsilon_R^+(S) \approx \epsilon_O^+(S)$ provides minimums information in terms of estimating the three evaluation measures. Replacing $\epsilon_R^+(S) \approx \epsilon_O^+(S)$ in Equations 3.19, 3.20, and 3.21, we have:

$$|Pr - E(Pr)| \leq 2\sqrt{\frac{|U|\epsilon_O^+(U)^2\epsilon_O^-(U)\log(\frac{2}{\delta})}{|S|^2\epsilon_O^+(S)^2}} \tag{3.26}$$

$$|Se - E(Se)| \leq 2\sqrt{\frac{|U|\epsilon_O^+(U)^2\epsilon_O^-(U)\log(\frac{2}{\delta})}{|S|^2\epsilon_O^+(S)^2}} \tag{3.27}$$

$$|Sp - E(Sp)| \leq 2\sqrt{\frac{|U|\epsilon_O^+(U)\epsilon_O^-(U)^2\log(\frac{2}{\delta})}{|S|^2\epsilon_O^-(S)^2}} \tag{3.28}$$

Interestingly, the first two bounds become equivalent and we only need to compare precision (sensitivity) vs. specificity. Similar to the previous section, by making the right hand side of these two bounds equal, we get the following boundary value for $\epsilon_O^+(U)$.

$$\epsilon_O^+(U) = \frac{\epsilon_O^+(S)^2}{\epsilon_O^+(S)^2 + (1-\epsilon_O^+(S))^2} \tag{3.29}$$

Figure 3 shows $\alpha$ as a function of $\epsilon_O^+(S)$, suggested by the above equation. As can be seen, for small values of $\epsilon_O^+(S)$ we would like to reduce the rate of positive examples in $U$ compared to $S$, i.e. $\epsilon_O^+(S) \geq \epsilon_O^+(U)$, and for large values of $\epsilon_O^+(S)$ we need $\epsilon_O^+(S) \leq \epsilon_O^+(U)$. In other words, we need to get more positive in $L$ for small values of $\epsilon_O^+(S)$ and more negative examples in $L$ for large values of $\epsilon_O^+(S)$. This suggests sampling from the minority class.

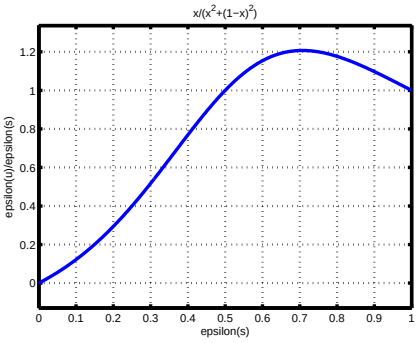



Figure 3: $\alpha$ for precision (sensitivity) vs. specificity for good rules

**3.4 Cost Analysis of Different Evaluation Measures** If a cost is associated with the estimation of each evaluation measure, one can utilize such costs directly in the analysis, in a similar approach to what we performed in the previous subsections. To see this, suppose $C_{pr}$, $C_{se}$, and $C_{sp}$ are the costs respectively associated with the estimation of precision, sensitivity and specificity. In order to consider these costs in the analysis of precision vs. specificity, for example, we need to have:

$$\frac{\epsilon_R^+(S)\epsilon_O^+(S)}{\epsilon_O^+(U)} = \frac{C_{sp}}{C_{pr}} \times \frac{(1-\epsilon_O^+(S))^3}{(1-\epsilon_R^+(S))(1-\epsilon_O^+(U)}$$

It is easy to see that the costs will be multiplied as a constant coefficient to the previous boundary values for $\epsilon_O^+(U)$.

**3.5 Sampling to Evaluate Area under the ROC curve (AUC)** A receiver operating characteristic (ROC) [7] is the plot of the sensitivity vs. 1-specificity for a binary classifier when the discrimination threshold of the classifier is varied. The area under the ROC curve or AUC is one standard measure of evaluating classification models particularly when the

problem is skewed. One might be interested in finding sampling strategies that results a sample of examples for a good approximation of AUC. In the setting of this paper, where the sample is being used to evaluate unknown rules, obtaining a sample to accurately estimate AUC of unknown rules is a tricky problem, yet very related to approximating precision and specificity. To see this, notice that estimating AUC of a given classifier can be summarized in estimating precision and specificity of all the rules obtained from that classifier with varying threshold. Hence, an accurate estimate of AUC for multiple unknown classifiers can be obtained by a sampling strategy that produces an accurate estimation of precision and specificity of all the rules obtained from such classifiers. In other words, the discussion in the previous section is valid for sampling to evaluate unknown rules using AUC.

## 4 Sampling in Practice

The previous discussion provides a guideline to the sampling strategy based on the parameters of the problem such as the nature of $\epsilon_O^+(S)$, the total number of examples $|S|$, the budget size $|L|$, and some preliminary knowledge about the characteristics of the rules we are interested to evaluate (e.g. $\epsilon_R^+(S)$ is small). Once this guideline for sampling is available (based on the analysis), the next interesting question is how to follow the strategy dictated by the guideline. For example, if the guideline asks for more positive examples, how should we retrieve positive examples? We call this problem *finite-pool sampling*. Notice that this problem is different from 1) the problem of regular active learning [11] and guided learning [16] where the objective is to construct a good predictive model; 2) the problem of finite-pool active learning [15] where the objective is to label all the examples in the pool. Unlike finite-pool active learning, the budget in finite-pool sampling is limited and we cannot afford to label as many examples as required to construct a classifier with a good confidence on labeling all the examples in the pool [15]. Nonetheless, the same procedures utilized for finite-pool active learning [15] and guided learning [16] to follow their suggested guidelines can be used for finite-pool sampling.

Here, we summarize four different schemes one can use to implement the task of finite-pool sampling. We note that while some of these schemes do not guarantee the examples from the target class are always selected, they at least try to follow the analysis (i.e. they make every effort to obtain more examples from the minority class).

- The simplest approach is to ask an expert to provide an example from the class the sampling guideline dictates. This example is then included in set $L$ with the corresponding label. The limitation of this approach is that all the burden of finding the example among all possible examples in the data set $S$ is on the expert. This is similar to the strategy utilized in biomedical citation screening [15], where the reviewers manually screen the reports to retrieve the most relevant studies to a research question. Notice that this approach is also similar to guided learning approach (discussed in the related work) where the reviewers are provided with certain characteristics of the data and then the reviewers find such examples and label them [1, 16].

- The second approach still requires expert's input, but eliminates expert's responsibility to search for individual examples. The idea is to start from a small initial set of labeled (positive and negative) examples and use them to learn a classification model. The model is then used to label all remaining unlabeled examples. One of the examples the model classifies to the minority class is presented to the expert to provide the true label. The example with its true label is then included in $L$. As more labeled examples are collected, the model is gradually improved. Active learning approaches can be also utilized to help this process, like in [15].

- The third approach is similar to the second approach and differs in how the model is initialized. The idea here is to use expert's input to define conditions that are more predictive of the minority class than the baseline population. The examples are then drawn either from examples that satisfy these conditions, or alternatively, from the model trained on the labeling induced by these conditions. The model can be further refined if examples with true labels become available, as was handled in the second approach.

- Finally, yet another approach may rely on the implicit domain information. For example, in getting labeled examples for information retrieval or recommendation systems, the possible relevant items could be detected implicitly by monitoring the user behavior; e.g. the items that get clicked or checked out are relevant. This is the click-through feedback widely utilized in information retrieval [8].

**4.1 Blind Labeling** In the previous subsections, we showed that by sampling from the minority class, we reduce the estimation uncertainty for the evaluation measures. An intuitive justification of sampling from minority class is that obtaining examples from the majority class is cheap if the prevalence of the two classes is highly biased toward one of the classes. We say it is cheap because if we randomly sample examples and assign them the label of majority class, we make a small error as we will see in this section. Without loss of generality, let us assume that $|S_O^+| \leq |S_O^-|$; i.e. the negative class is the majority one. If the rate of positive examples is very small (i.e. $\epsilon_O^+(S) = \frac{|S_O^+|}{|S|} \ll \frac{1}{2}$), we propose

to blindly assign negative label to a subset $B$ of examples from set $U$ ($B \in S$). By blind labeling, we not only increase the number of labeled examples in $L$ with a small error but also we keep almost the same rate of positive examples in set $L$ as original set $S$. This is particularly important because if the rate of positive examples in $L$ is very different than that in set $S$, we might over/under-estimate the values of the evaluation measures. To perform the blind labeling, we propose to bring $\frac{|L_O^+|-|L|\epsilon_O^+(S)}{\epsilon_O^+(S)}$ random examples from $S$ to $L$ and blindly label them as the majority class (negative). Let us call the new sampled set $\hat{L}$ that contains examples in $L$ and the extra blindly labeled negative examples. $\hat{L}$ has the same rate of examples from two classes as $S$ because

$$\frac{|\hat{L}_O^+|}{|\hat{L}|} = \frac{|L_O^+|}{|L| + \frac{|L_O^+|-|L|\epsilon_O^+(S)}{\epsilon_O^+(S)}} = \epsilon_O^+(S) \tag{4.30}$$

The error caused by the blind sampling will be small if the rate of examples from the minority class is small enough. To see this, notice that if we sample $|B|$ examples from $U$ for blind labeling, the average number of positive examples in $B$ will be $|B|\epsilon_O^+(U)$, a small number if $|L|$ and $\epsilon_O^+(S)$ are small.

One practical problem with this approach is that it needs the knowledge of $\epsilon_O^+(S)$. In many domains, there is a domain knowledge about approximate value of $\epsilon_O^+(S)$. In case that there is not such knowledge, one approach to obtain an approximation is to get an initial set of randomly selected labeled examples and estimate $\epsilon_O^+(S)$ using this set. This set can further be used to learn an initial model to select positive examples. The following proposition shows that the approximation can be reasonably good for very small or large values of $\epsilon_O^+(S)$.

PROPOSITION 4.1. *For any $\delta \in [0,1]$, with probability at least $1-\delta$, we have the following bound on the difference between $\epsilon_O^+$ and the approximation made using a random sample $M$:*

$$|\epsilon_O^+(M) - \epsilon_O^+(S)| \leq 2\sqrt{\frac{\epsilon_O^+(S)\epsilon_O^-(S)\log(\frac{2}{\delta})}{|M|}} \tag{4.31}$$

*Proof.* Notice that the mean and variance of $\epsilon_O^+(M)$ is:

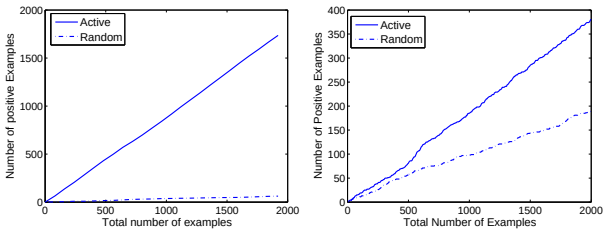
$$E(\epsilon_O^+(M)) = E(\frac{|M_O^+|}{|M|}) = \frac{E(|M_O^+|)}{|M|} = \frac{|M|\epsilon_O^+(S)}{|M|} = \epsilon_O^+(S)$$

$$\delta^2(\epsilon_O^+(M)) = \frac{\delta^2(|M_O^+|)}{|M|^2} = \frac{|M|\epsilon_O^+(S)\epsilon_O^-(S)}{|M|^2} = \frac{\epsilon_O^+(S)\epsilon_O^-(S)}{|M|}$$

Using Bernstein inequality, we have

$$P\left(|\epsilon_O^+(M) - \epsilon_O^+(S)|| > \xi\right) \leq 2\exp\left(-\frac{\xi^2|M|}{4\epsilon_O^+(S)\epsilon_O^-(S)}\right) \tag{4.32}$$

Similar to the proof of Proposition 3.1, we can conclude the result in this proposition.




(a) CovType (b) Connect-4

Figure 4: The retrieved number of positive examples over the trials.

## 5 Experiments

In this section, we verify the efficacy of the guidelines proposed by the analysis in this paper on covtype and connect-4 data sets from UCI data repository [6], as described here. Covtype has about $580,000$ records, each represented by $54$ attributes and one of the 7 classes. We preprocessed this data and constructed a binary problem by considering classes $4$ and $7$ as positives and other classes as negatives[4]. This results in $23257$ positive and $557755$ negative examples, a positive rate of $0.04$. Connect-4 has $67557$ records, with each record represented by $126$ features and one of the three classes (-1,0,1). We construct a binary problem from this data by considering class $0$ as the positive class, which results a positive rate of $0.1$.

To retrieve positive examples in the active mode, we utilize $100$ randomly selected examples and construct an initial decision tree classifier. Using the recommendation of this decision tree, we randomly choose $5$ positive examples[5], and add them to the labeled set. We then construct a new classifier and repeat the same procedure until we collect a total number of $2000$ examples. Note that not all $5$ selected examples in each round are positive. Figure 4 shows the retrieved number of positive examples over the trials for both the active and random sampling for two data sets.

As the baseline, we randomly sample from the data set until we obtain $2000$ examples. To compare the results, we create rules with different rate of positive examples, i.e. $|S_R^+| \in \{10000, 20000, 30000, 40000, 50000\}$. For each $S_R^+$, we create $100$ random rules and report the average performance and the confidence interval of both active sampling and random sampling. We report the absolute value of the difference between the true value of each evaluation measure (the gold standard computed using the whole data set) and the value computed on the subsets created by active/random sampling.

Figures 5 and 6 shows the result for rules with varying

[4]Notice that the method is not sensitive to the choice of the classes as long as the result is an unbalanced data set.

[5]The procedure is robust to the different setting of this parameter.

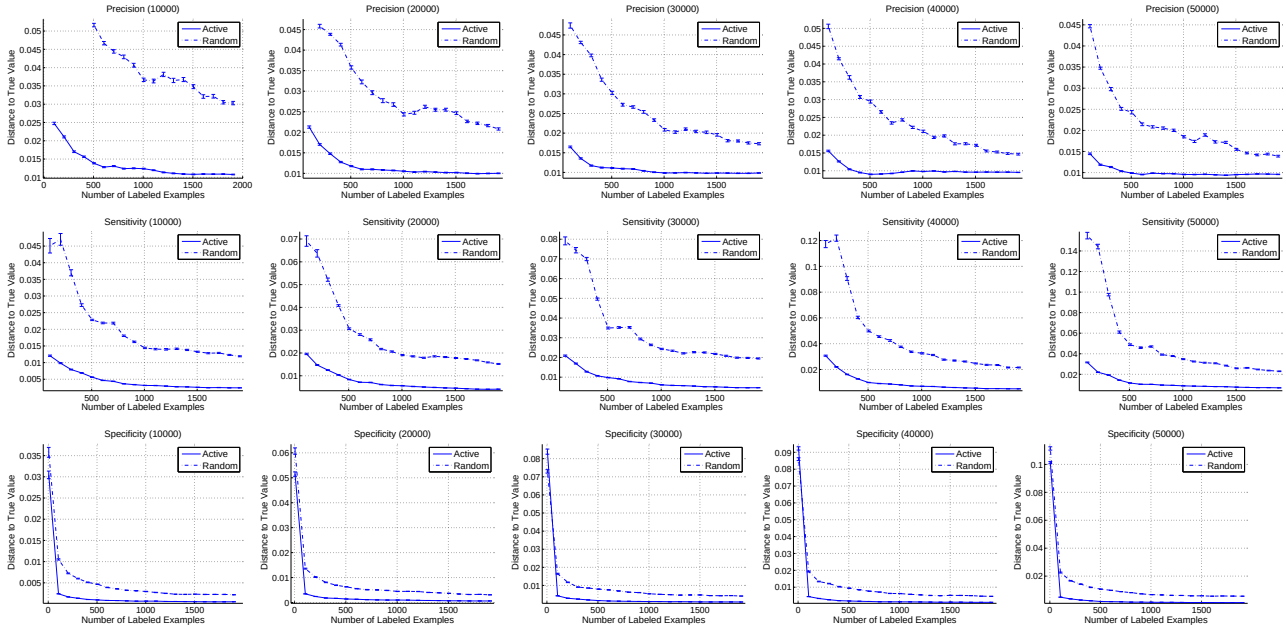



Figure 5: The difference between the true value of evaluation measure (computed over $S$) and the estimated ones on CovType data set. The dot line is the estimation using random sampling and the solid line is the estimation using active sampling. Column $i$ in the figure shows the average of 100 random generation of rules of size $i * 10000$.

value of $\epsilon^+_R(S)$ on two data sets. Column $i^{th}$ in these features is the average result in estimating three evaluation measures on 100 rules with positive rate of $\epsilon^+_R(S) = i \times 10000$. As can be seen, the active sampling approach results in a better approximation of the evaluation measures. The results confirms that with only a small labeled set of examples, we obtain much better estimation accuracy than random sampling. Consider the computation of precision on CovType data set for rules with $\epsilon^+_R(S) = 10000$, as an example. With only about 80 labeled examples, we can obtain an accuracy of 0.025 using active sampling while to obtain the same estimation accuracy using random sampling, more than 2000 examples are required. Overall in all the figures for CovType, a reduction rate of 95% in the number of labeled examples is achieved when compared to the random sampling of examples. The reduction rate for connect-4 data set is smaller than that for the CovType data sets. This is because CovType data set is very unbalanced in which random sampling performs very poor as shown by the analysis. Also notice that the smaller is the $\epsilon^+_R(S)$, the more effective is sampling from minority class, as predicted by the analysis. As we go from unbalanced data set and unbalanced rules (rules with small $\epsilon^+_O(S)$) to balanced data set and balanced rules, the advantage of sampling from one class disappears and it becomes more beneficial to perform random sampling.

## 6 Conclusion

We provided a theoretical analysis of the sampling strategy for evaluating classification rules of unknown nature. In particular, we showed that for unbalanced data sets, it is much more beneficial to sample from the rare class. We discuss how the proposed framework can be generalized to the case where a cost associated with estimating each evaluation measure. We also verified and confirmed our theoretical analyses experimentally using two data sets from the UCI data repository.

## 7 Acknowledgement

This research work was supported by the grants R01LM010019 and R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## References

[1] J. Attenberg and F. J. Provost. Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *Proceedings of the 16th ACM SIGKDD 2010*, pages 423–432. ACM, 2010.

[2] M. Bloodgood and V. Shanker. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In *HLT-NAACL (Short Papers)*, pages

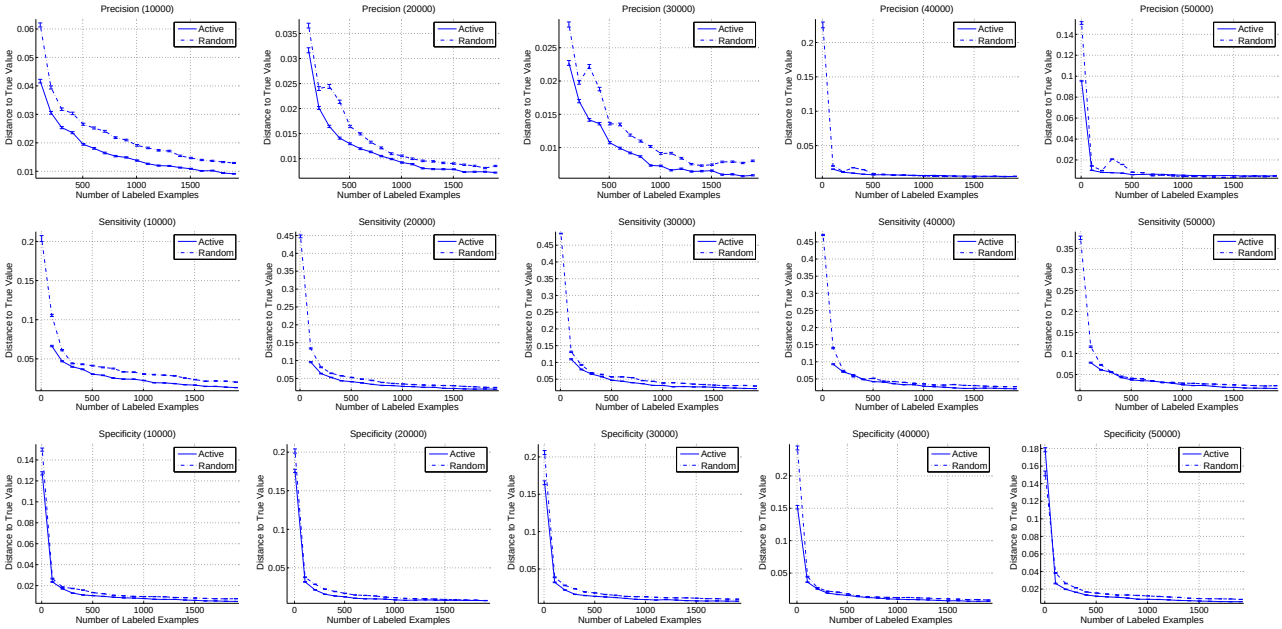



Figure 6: The difference between the true value of evaluation measure (computed over $S$) and the estimated ones on Connect-4 data set. The dot line is the estimation using random sampling and the solid line is the estimation using active sampling. Column $i$ in the figure shows the average of 100 random generation of rules of size $i * 10000$.


137–140. The Association for Computational Linguistics, 2009.

[3] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research (JAIR)*, 4:129–145, 1996.

[4] Pinar Donmez, Jaime G. Carbonell, and Paul N. Bennett. Dual strategy active learning. In *Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4701, pages 116–127. Springer, 2007.

[5] Devdatt Dubhashi and Sandeep Sen. *Concentration Of Measure For Randomized Algorithms: Techniques And Analysis*. 2005.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982.

[8] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

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

[10] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML'01*, 2001.

[11] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153–178, 2004.

[12] Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *International Conference on Machine Learning*, pages 951–958, 2010.

[13] K. Tomanek and U. Hahn. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), September 1-4, 2009, Redondo Beach, California, USA*, pages 105–112. ACM, 2009.

[14] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proc. ICML*, pages 999–1006. Morgan Kaufmann, San Francisco, CA, 2000.

[15] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD*, KDD '10, pages 173–182, New York, NY, USA, 2010. ACM.

[16] G. M. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

[17] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003.