Submitted to *Management Science*
manuscript MS-00665-2006

Active Feature-Value Acquisition

Maytal Saar-Tsechansky
McCombs School of Business, The Univeristy of Texas at Austin
maytal@mail.utexas.edu

Prem Melville
IBM Research
pmelvil@us.ibm.com

Foster Provost
Stern School of Business, New York University
fprovost@stern.nyu.edu

Most induction algorithms for building predictive models take as input training data in the form of feature vectors. Acquiring the values of features may be costly, and simply acquiring all values may be wasteful, or prohibitively expensive. Active feature-value acquisition (AFA) selects features incrementally in an attempt to improve the predictive model most cost-effectively. This paper presents a framework for AFA based on estimating information value. While straightforward in principle, estimations and approximations must be made to apply the framework in practice. We present an acquisition policy, Sampled Expected Utility (SEU), that employs particular estimations to enable effective ranking of potential acquisitions in settings where relatively little information is available about the underlying domain. We then present experimental results showing that, as compared to the policy of using representative sampling for feature acquisition, SEU reduces the cost of producing a model of a desired accuracy and exhibits consistent performance across domains. We also extend the framework to a more general modeling setting in which feature values as well as class labels are missing and are costly to acquire.

Key words: Information acquistion, predicitve modeling

1. Introduction

"...the shift from relying on existing information collected for other purposes to using information collected specifically for research purposes is analogous to primitive man's shifting from food collecting to agriculture..." (Siegel and Fouraker 1960)

Predictive models play a key role in numerous business intelligence tasks. Models are induced from historical data to predict customer behavior or to detect adversarial acts such as fraud. A critical factor affecting the knowledge captured by such a model is the *quality* of the information from which the model is induced—the "training data." In the context of predictive modeling, the quality of information pertains to the training sample's composition, the accuracy of the values, and the number of unknown values.

For many predictive modeling tasks, potentially pertinent information is not immediately available, but can be acquired at a cost. Traditionally, *information acquisition* and *inductive modeling* are addressed independently; data are collected irrespective of the modeling objectives. However, information acquisition and predictive modeling in fact are mutually dependent: newly acquired information affects the model induced from the data, and the knowledge captured by the model can help determine what new information would be most useful to acquire (Simon and Lea 1974). We would like to take advantage of this relationship, and develop feature-value acquisition *policies* for predictive model induction—procedures for evaluating and selecting feature-value acquisitions which will be used for model induction. Mookerjee and Mannino (1997) address a similar problem where costly feature-values of a *test* instance for which *inference* is requested are unknown and are acquired sequentially given an existing knowledge base. Here we study a complementary problem in which values must be acquired for *induction*.

The availability of generic, effective and computationally feasible information-acquisition policies for model induction can affect business practices by transforming existing information-acquisition practices and by changing the manner by which firms interact with consumers. As an example, consider the generation of personalized recommendations to customers. Often, a recommender system's underlying predictive model employs customers' ratings of prior purchases as predictors of a customer's preference for a product she has not yet purchased. The availability of many ratings from a large number of customers is critical for successful induction of an accurate model of consumer preferences. However, without costly incentives, most consumers rarely provide this valuable feedback. To improve the model's predictive accuracy, it is infeasible to acquire feedback from all consumers about all products, even those they have already purchased. A better acquisition policy would determine which ratings from which customers would be most cost-effective to acquire via costly incentives, in order to obtain the desired modeling objective for the least cost (Huang 2007). Similar scenarios emerge in other modeling tasks where missing feature-values can be acquired at a cost. These include modeling of medical treatment effectiveness and diagnostics from medical databases, where patients' information, such as details on prior hospitalizations and prior medical tests, are notoriously incomplete. Intelligent information-acquisition policies can also dramatically change already established information-acquisition models: presently, firms acquire bundles of psychographic, consumption, and lifestyle data periodically from third-party suppliers, such as Axiom, to support business intelligence modeling for tasks such as risk scoring, customer retention, and personalized marketing. As with other information goods, a firm should consider how to bundle and price information on consumers. Effective acquisition policies will enable firms to identify and to acquire different types of information for different consumers at potentially different prices, to enhance modeling cost-effectively. These capabilities may also enable small firms to reap the benefits from business intelligence modeling, allowing them to enrich their potentially limited data by selectively acquiring useful information.

Given training data with missing feature-values, an arbitrary classification-model induction algorithm, a set of prospective feature-value acquisitions, and the cost of acquiring each specific feature-value (the cost of features may vary feature-to-feature and instance-to-instance), the general AFA problem is to acquire feature-values so as to obtain a desired performance level for minimum cost. In this paper we consider performance to be some function of the model's generalization accuracy.[1] However, because we do not know a priori the population under consideration, the generalization performance cannot be known exactly, and we must estimate it from a sample. Thus, AFA policies cannot be provably optimal, and so we employ a heuristic measure of the performance from feature-value acquisition.

Even given a heuristic measure of performance, in principle, identifying the feature-value set that yields the desired performance objective at a minimum cost requires considering all possible sets of prospective acquisitions. Unfortunately, this is not feasible to compute for most interesting problems. Moreover, given a finite sample, both statistical learning theory and practical experience tell us that more search through possible models often leads to worse performance, due to problems of multiple comparisons (Vapnik 1998, Jensen and Cohen 2000). We will revisit this issue in Section 5. We therefore revise the objective of AFA for this paper as follows. Given a performance measure, we aim to identify the individual feature-value to acquire next, in order to achieve the greatest improvement in the performance measure per unit cost, which implies a greedy, myopic acquisition policy.

The primary contributions of this paper are a general framework for addressing AFA and a specific method for solving AFA problems based on appropriate heuristics. To our knowledge, no prior work[2] addresses general AFA. We propose an acquisition policy that produces acquisition schedules iteratively based on estimates of the expected utility from different potential acquisitions. In principle this is straightforward, but the AFA setting renders utility estimation particularly challenging: estimations often must be made based on little available information, and ought to be sensitive so as to capture the benefit to induction of individual feature-values. Further, because the space of possible acquisitions can be immense, estimating the value of each potential acquisition may be computationally infeasible. We develop and study empirically the impact of measures for capturing the value to induction of single feature-values in the presence of scarce data, and propose different mechanisms to reduce the complexity of the estimation.

This expected-utility approach has several important advantages. The framework is general and can be applied to derive an acquisition schedule for any induction technique. This is important because due to the inherent bias of different modeling techniques and professional regulations in some industries, no single technique is applied across all problems. Another important advantage of the expected-utility approach is that it can be applied to improve any utility function derived from the model's predictive performance, such as estimated generalization accuracy, expected profit in a particular setting, or the expected cost of model

[1] The model's expected accuracy over the population under consideration.

[2] With the exception of a short paper on our preliminary studies (Melville et al. 2005).

error. Finally, the approach can utilize information about the varying cost of information to derive an acquisition schedule, not assuming that the cost of acquiring an unknown value is fixed for all features and/or for all instances. Experimental results demonstrate that the resulting method provides significantly better models for a given cost than those obtained with other acquisition policies. Since the method utilizes acquisition cost information, it is particularly advantageous in challenging tasks for which there is significant variance across potential acquisitions with respect to their informativeness and their cost.

Finally, another contribution of this paper is an extension of the policy to a more general acquisition problem. For some modeling tasks *class labels* (i.e., dependent variables' values) are missing as well as feature values, and either or both may be acquired at cost. We show that because our framework estimates the value to induction of different acquisitions it allows the dependent variable to be treated as yet another feature, and thus the AFA framework and method can be extended directly to address this new problem. In practice, the method interleaves the acquisition of class labels and feature values, based on the marginal expected value from each acquisition; we show it to be superior both to uniform acquisitions and to policies that consider the acquisition only of feature-values or only of class labels.

2. Active Feature-value Acquisition

Assume a classifier induction problem where each instance is represented with n independent variables plus a discrete, dependent "class variable." The available data set of m instances can be represented by the "incomplete" matrix F, where $F_{i,j}$ corresponds to the value of the j-th feature of the i-th instance, which may be missing. Missing elements in the matrix F represent missing feature-values that can be acquired at a cost. In general, the cost of different feature-values may vary, depending on the nature of the particular feature or of the instance for which the information is missing.

Algorithm 1 General Active Feature-value Acquisition Framework

```
Given: $F$: initial (incomplete) instance-feature matrix; $Y = \{y_i, i = 1, ..., m\}$: class labels for all instances; $T$: training set = $< F, Y >$; $L$: classifier induction algorithm; $\beta$: size of query batch; $C$: cost matrix for all instance-feature pairs;
Initialize set of possible queries $Q$ to $\{q_{i,j} : i = 1, ..., m; j = 1, ..., n;$ such that $F_{i,j}$ is missing$\}$
Repeat until stopping criterion is met
    Induce a classifier, $M = L(T)$
    $\forall q_{i,j} \in Q$ compute $score(q_{i,j}, c_{i,j}, L, T)$
    Select the subset, S, of $\beta$ feature value with the highest $score$s
    $\forall q_{i,j} \in S$
        Acquire values for $F_{i,j}$
    Remove $S$ from $Q$
End Repeat
Return $M = L(T)$
```

We present an iterative, sequential acquisition framework, where at each acquisition phase, alternative acquisitions are evaluated in order to acquire the value of $F_{i,j}$ at the cost $C_{i,j}$ that provides the largest improvement per unit cost in the performance objective. The iterative framework for AFA is presented in Algorithm 1. The framework is independent of the classification modeling technique; it is given a learner, L, which includes a model induction algorithm and a missing value treatment to allow for induction from the incomplete matrix F.[3] At each phase a "score" is estimated and assigned to each potential acquisition, reflecting the estimated added value per unit cost of the acquisition. The acquisition with the highest score is selected and the corresponding feature-value is acquired; a particular approach for assigning scores to potential acquisitions will be described in detail in Section 2. Once a value is acquired, the training data and the information acquisition cost are appropriately updated and this process is repeated until some stopping criterion is met, e.g. a desirable model accuracy has been obtained. Often, many values must be acquired in order to obtain a desired performance level. In order to reduce the computational burden or based on domain constraints, at each iteration an AFA policy may acquire a “batch” of $\beta \geq 1$ values. As before, computing the set that will result in the greatest improvement in the heuristic measure is computationally complex. In this paper, we select the values with the highest individual scores; the sensitivity to this choice is examined in Section 3.3.

We now present a method for Active Feature-value Acquisition based on computing the value of the information that may be acquired. The central component of the computation also presents the main difficulties with its implementation: the computation of the value of information prior to acquisition, when only partial knowledge about the acquired information is available. We discuss three difficulties with the computation, and present approximation techniques to address these difficulties. Together they comprise the proposed AFA method: *Sampled Expected Utility* (SEU).

We estimate the value of a potential acquisition by its expected marginal contribution to predictive performance. Because the true value of the missing feature is unknown prior to its acquisition, it is necessary to estimate the potential impact of an acquisition for different possible acquisition outcomes. The acquisition with the highest information value will be the one that results in the maximum utility in expectation, given a model, a model induction algorithm, and a particular utility function. For the latter, the objective may be to maximize the model’s generalization accuracy, or to maximize future profit, or to minimize the costs incurred due to incorrect predictions, etc. A utility score captures the expected improvement from each potential acquisition. Assuming feature j has K distinct possible values $v_1, ..., v_K$, the expected utility of the acquisition $q_{i,j}$, or "query" for short, is given by:

$$E(q_{i,j}) = \sum_{k=1}^{K} \mathcal{U}(F_{i,j} = v_k) P(F_{i,j} = v_k) \tag{1}$$

[3] Induction algorithms either include an internal mechanism for incorporating instances with missing feature-values (Quinlan 1993) or require that missing values be imputed first. Henceforth, we assume that the induction algorithm includes or is coupled with some treatment for instances with missing values.

where $P(F_{i,j} = v_k)$ is the probability that $F_{i,j}$ has the value v_k, and $\mathcal{U}(F_{i,j} = v_k)$ is the utility of knowing (via acquisition) that the feature-value $F_{i,j}$ is v_k. The utility $\mathcal{U}(\cdot)$ is the marginal improvement in performance per unit of acquisition cost:

$$\mathcal{U}(F_{i,j} = v_k) = \frac{\mathcal{A}(F, F_{i,j} = v_k)}{C_{i,j}} \tag{2}$$

where $\mathcal{A}(F, F_{i,j} = v_k)$ is the change in value to induction from augmenting F with $F_{i,j} = v_k$; and $C_{i,j}$ is the cost of acquiring $F_{i,j}$. This *Expected Utility* policy therefore corresponds to selecting the query that will result in the estimated largest increase in performance per unit cost in expectation. If all feature costs are equal, this corresponds to selecting the query that would result in the classifier with the highest expected performance. Otherwise, *Expected Utility* allows several low-yield, high-margin acquisitions to be selected instead of one higher-yield acquisition with less expected improvement per unit cost.

In principle this approach would allow the estimation of the value of each possible acquisition, and then the selection of acquisitions by ranking them by their information-value estimates. However, there are significant hurdles to its practical implementation. We introduce the challenges next, and then address each in turn in the following three subsections.

Challenge 1. Estimating contribution to induction. As outlined in Eq. 2, for each query ($\forall q_{ij} \in Q$) computing expected utility requires the estimation of the value, $\mathcal{A}(\cdot)$, to induction from the acquisition. Here we assume classification accuracy to be the performance metric of interest; as discussed, the framework applies to other goals such as minimizing misclassification cost or maximizing profit. As we will see, in order to estimate the expected improvement in classification performance, it is necessary to detect expected changes in the modeling technique's average class probability estimation that are conducive to improved classification accuracy. It turns out that the obvious measure, classification accuracy itself, is not sensitive enough to such changes.

Challenge 2. Estimating value distributions. For estimating the expected contribution of different acquisitions, a prerequisite is to estimate the conditional distribution $P(F_{i,j} = v_k)$ for each missing value of $F_{i,j}$, as needed in Eq. 1. We must identify an estimation mechanism appropriate for the AFA setting: many feature-values may be missing thereby rendering some modeling mechanisms more effective than others. For example, some mechanisms require that missing predictors or their distributions be estimated to produce a prediction. This adds yet another layer of estimation which may undermine the model's prediction, sometimes significantly (Saar-Tsechansky and Provost 2007).

Challenge 3. Reducing the consideration set. Even if all unknown values were estimated accurately, selecting the best from *all* potential acquisitions would require estimating the utility of, in the worst case, mn queries. This would be very expensive computationally and is infeasible for most interesting problems.

2.1. Estimating an acquisition's contribution to performance

Let us consider the measure to be used for estimating the value to induction from an acquisition, $\mathcal{A}(F, F_{i,j} = v_k)$. Let us assume for this discussion that acquiring new information aims to improve the model's classification accuracy, for a binary classification problem (thus, the decision threshold for maximum a posteriori classification is 0.5). Assuming that estimations will be computed by averaging over a set of hold-out examples, this suggests a simple criterion for identifying an effective utility measure A—prefer acquisitions that improve estimated accuracy:

Criterion 1: *For a given hold-out instance, $A(f_1) \succ A(f_2)$ if $f_1 > \theta$ and $f_2 < \theta$, where f_i refers to the model's estimated probability that the given instance belongs to the true class, θ* is the decision boundary, *and $a \succ b$ denotes that a is better than b.*

An obvious measure for this contribution is the model's classification accuracy itself (i.e., the estimated generalization accuracy) over the augmented sample F. However, as we show below, classification accuracy does not capture fine-grained changes in the models and therefore we would like a more sensitive measure for evaluating the benefits from prospective acquisitions.

To understand why, it is necessary to examine the dynamics of the modeling setting. Specifically, the training data—and therefore the models induced—continuously change as new information is acquired. Rather than examine the classification performance of a particular model induced from one version of the data, it is useful to examine how new acquisitions affect the *distribution* of estimations induced from different likely variations of the training data. Friedman's analysis of the relationship between training data and classification error (Friedman 1997) examines how changes in an induction technique's average estimation of the *probability* of the true class affect the likelihood of classification error. For the sake of discussion, assume binary classification and let $f(y|x)$ and $\hat{f}(y|x)$ denote the actual probability and the model's estimated probability that an instance belongs to class y, respectively, where x is the input vector of observable attributes. Following Friedman's analysis, the probability that the predicted class $\hat{y}$ is estimated (erroneously) not to be the most likely class y can be approximated with a standard normal distribution by:

$$P(\hat{y} \neq y) = \tilde{\Phi}\left[sign(f - 1/2)\frac{E\hat{f} - 1/2}{\sqrt{\operatorname{var}\hat{f}}}\right] \tag{3}$$

where $\tilde{\Phi}$ is the upper tail area of the standard normal distribution, $\operatorname{var}\hat{f}$ denotes the estimation variance resulting from variations in the training sample and where $E\hat{f}$ denotes the mean of the probability estimation $\hat{f}(y|x)$ generated by models induced from different variations of the training sample. Henceforth we refer to $E\hat{f}$ and to $\operatorname{var}\hat{f}$ as the average probability estimation and estimation variance, respectively. When the average probability estimation leads to incorrect class prediction and given a certain estimation variance, the likelihood of incorrect classification decreases as the average probability estimation of the true class increases. Equation 3 also reveals that the likelihood of correct classification can improve when the average probability estimation *already* leads to a correct prediction. As shown in Equation 3, given a certain estimation variance, $\operatorname{var} \hat{f}$, if the true class probability f and the average probability estimation $E\hat{f}$ lead to the same (correct) classification, then further from the decision boundary the average probability estimation $E\hat{f}$ is, the higher is the likelihood of a reduction in classification error. This is because it is less probable for the estimation variance to cause an erroneous classification. This analysis suggests two criteria for an effective measure A of the utility from an acquisition. First, a more general criterion than criterion 1: A should favor more extreme (correct) estimates of class membership probability.

Criterion 1′: *For a given hold-out instance, $A(f_1) \succ A(f_2)$ iff $f_1 > f_2$, where f_i refers to the model's estimated probability that the given instance belongs to the true class*. This is a specialization of criterion 1; criterion 1 always will hold if criterion 1′ does (but not vice versa).

Second, for a correctly classified example, for a fixed-size change in the estimate of class membership probability, A should favor changes to estimates nearer to the decision boundary θ (in the analysis above this boundary is assumed to be 0.5):

Criterion 2: *For a given hold-out instance, $A(f_1) - A(f_1 + \Delta) \succ A(f_2) - A(f_2 + \Delta)$, $\forall \Delta : 0 < \Delta \leqslant \min(1 - f_1, 1 - f_2)$, and $\theta < f_1 < f_2$.*

Based on these two criteria, we can assess the adequacy of different possibilities for the utility measure A, including intuitive alternatives such as estimated accuracy (error rate), or the estimate $\hat{f}$ of the probability of class membership itself. Specifically, classification accuracy is not an adequate AFA utility measure because it does not satisfy either criterion (1′) or (2) completely. This is illustrated for binary classification by the dotted line in Figure 1, which shows classification error (1-accuracy) as a function of the model's estimated probability of the *true* class, assuming maximum a posteriori classification.

Let us consider an alternative AFA utility measure, *Log Gain* (LG) (also known as cross-entropy). For a model induced from a training set F, let $\hat{f}_F(y|x)$ be the probability estimated by the model that instance x belongs to class y, and $\delta(A)$ is an indicator function such that $\delta = 1$ if A is the correct class and $\delta = 0$ otherwise. Let $LG(x) = \sum\limits_{y} -\delta(y) \log_2 \hat{f}_F(y|x)$; Log Gain is "better" as its value decreases. Consider an evaluation data set of t instances; let the value to induction from an acquisition resulting in a training set F be given by the sum of Log Gains over these t instances: $\mathcal{A}(F) = \sum\limits_{e=1}^{t} LG(x_e)$. Hence for each value V_k that feature $F_{i,j}$ can take we would induce a model from the augmented data set and compute this sum of Log Gains. As illustrated in Figure 1 Log Gain satisfies both criteria. It captures important changes in the model's estimation following an acquisition, which will allow the AFA policy to focus on acquisitions that decrease the likelihood of classification error: using Log Gain will result in higher scores for acquisitions

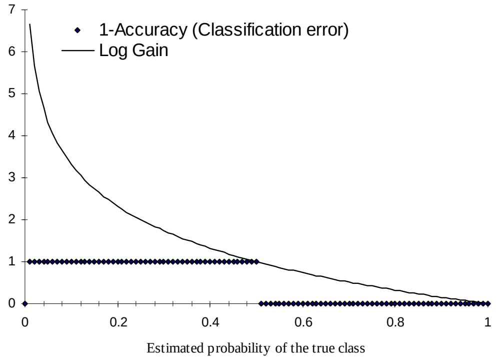


Figure 1 **Log Gain and classification error vs. the probability of the true class**

that increase the average probability estimation of the true class when, on average, the model's class prediction is incorrect. It will also promote acquisitions that lead to more extreme probability estimations when the model's class prediction is accurate—reducing the risk of erroneous classification as new information is added to the training sample. In principle, other measures that promote these two objectives should benefit the AFA policy as well.

2.2. Estimating feature-value distribution

Let us now address the second term in equation 1. We need to estimate the conditional probability distribution of a missing feature value given the known values. For each feature j, the probability $P(F_{i,j} = v_k)$ in Eq. 1 will be inferred from a model $M_{i,j}$, based on the other information available about instance i (what is known about the other features and the class).

Unfortunately, because of the AFA setting, the instances to which the feature-distribution-estimation model $M_{i,j}$ would be applied may have many missing values. Considering an arbitrary predictive model, predictors whose values are required for inference (rather than for induction) may not be available. The loss in predictive accuracy stemming from the need to estimate missing predictors' values or their distributions can be avoided if the model incorporates only predictors whose values are known for this instance (Saar-Tsechansky and Provost 2007). However, for AFA this would entail considering a tremendous number of combinations, as at any point in an acquisition schedule different instances may include widely different sets of known feature-values.

For the main results of this paper, in order to estimate the feature-value distribution model $M_{i,j}$ we employ only one predictor: the class variable. This is because for our setting, the class is guaranteed to be known at inference time. In principle, one can employ any set of known predictors to estimate a missing feature-value distribution. In Section 3 we validate empirically the benefits of relying on known predictors exclusively. Conditioning on the class variable outperforms a simpler model that does not condition the missing feature distribution at all, because the latter does not take into account any instance-specific information in the estimation. A straightforward application of more complex modeling that captures the interactions among predictors, but that requires the estimation of unknown predictors for inference, does not improve AFA performance.

2.3. Reducing the consideration set

Estimating the expectation $\hat{E}(\cdot)$ for each query, $q_{i,j}$, requires training one classifier for each possible value of $F_{i,j}$. Therefore, exhaustively evaluating all possible queries is infeasible for most interesting problems. One way to make this exploration tractable is by applying a computationally fast approach to identify a subset of all the possible queries which will subsequently be considered for acquisition. In particular, let the exploration parameter α ($1 \leq \alpha \leq \frac{mn}{\beta}$) control the size of the sample to be considered for acquisition. (Recall that β is the batch size, m the number of examples, and n the number of features.) To acquire a batch of β queries, first a sub-sample of $\alpha\beta$ queries is selected from the available pool of prospective acquisitions; then the expected utility of each query in this sub-sample is evaluated using equation 1. The value of α can be set depending on the amount of time the user is willing to spend on this process and the effectiveness of the selection scheme. We consider two computationally fast approaches to identify a subset of queries.

The first approach, *Uniform Sampling (US)*, identifies a representative subset of missing feature-values via a uniform random sample of queries. However, when the consideration set is drawn uniformly at random, particularly informative acquisitions may be left out of the consideration set. An alternative approach is to limit the consideration set to a subset of queries that are more likely to be informative for model induction than a query drawn at random. In particular, we propose selecting the consideration set of queries from particularly informative *instances*. This invokes the subproblem: what then constitutes an *informative* instance for model induction? We conjecture that acquired feature-values are more likely to have an impact on classification accuracy when the acquired values belong to a *misclassified* example and, as such, embed predictive patterns that are not consistent with the current model. Next, correctly classified instances are more informative if their class prediction is uncertain. The use of uncertainty for active data acquisition originated in work on optimum experimental design (Federov 1972) and has been extensively applied in the active learning literature (Cohn et al. 1994, Saar-Tsechansky and Provost 2004). For a probabilistic model, a lack of discriminative patterns results in uncertain predictions where the model assigns similar likelihoods for class membership of different classes. Formally, for an instance x, let $P_y(x)$ be the estimated probability that x belongs to class y as predicted by the model. Then the uncertainty score is given by $P_{y_1}(x) - P_{y_2}(x)$, where $P_{y_1}(x)$ and $P_{y_2}(x)$ are the first-highest and second-highest predicted class probability estimates respectively. Motivated by this reasoning, *Error Sampling* (ES) ranks informative instances higher if they are misclassified by the current model. Next *Error Sampling* ranks instances in increasing order of the uncertainty score.

We call the approaches in which *Uniform Sampling* and *Error Sampling* are used to reduce the set of missing values considered for acquisition, *Sampled Expected Utility-ES* (SEU-ES) and *Sampled Expected Utility-US* (SEU-US), respectively.

3. Experimental Evaluation

We now present a comprehensive set of experiments which demonstrate the efficacy of AFA and the benefits of the measures we described in Section 2.

3.1. Objectives and Methodology

We begin with the key empirical question of whether feature-values can be acquired cost-effectively with AFA as compared to a default policy in which a representative set of feature-value acquisitions are drawn uniformly at random. Next, we present extensive empirical results which carefully examine the benefits of the measures we propose for AFA as compared to alternatives. These evaluations provide empirical support to the arguments we present in Section 2 regarding measures that are likely to be particularly effective for AFA. We then examine the upper-bound performance that could be obtained with an omniscient "oracle" and how close SEU is to this performance. In addition, we examine which of the two measures that SEU employs is closest to an omniscient measure for the corresponding quantity—these results suggest how improvements in each of SEU's estimations can contribute to SEU's overall performance so as to approach the performance of the oracle. Finally, we perform sensitivity analyses exploring how different settings affect SEU's performance. These include SEU performance with different feature-value cost structures and parameters that determine the size of the initial sample provided to SEU, the number of examples it considers for acquisition, and the number of examples acquired in each acquisition phase.

The Primary Results. To address the first question we compare, as a function of acquisition cost, the classification performance obtained by the policies SEU-ES, SEU-US, and a policy (Uniform) that selects acquisitions uniformly at random. This study also aims to examine the merits of the two approaches, uniform sampling (US) and error sampling (ES), which we propose for reducing the set of prospective acquisitions considered by SEU.

We then demonstrate empirically the properties of the utility measure we propose for AFA in Section 2. The ability of Sampled Expected Utility to rank potential acquisitions accurately will be affected by the accuracy of its estimates of the quantities in Eq. 1, viz., the value to induction from a prospective acquisition of a value $F_{i,j} = v_k$, ($\mathcal{A}(F, F_{i,j} = v_k)$), and the estimated distribution of values for each unknown feature ($P(F_{i,j} = v_k)$). Specifically, in Section 2.1 we showed analytically that *Log Gain* effectively captures important changes in the estimated class probabilities following an acquisition, which affect the likelihood of erroneous classification. As such, Log Gain improves SEU′s ability to identify acquisitions that are particularly likely to reduce classification error. By contrast, using classification accuracy as the utility measure does not capture changes in probability estimation, except when the estimated class membership also changes. To demonstrate this effect on SEU performance we compare SEU with a modified policy that employs classification accuracy on the training set to estimate the value of prospective acquisitions. We refer to this policy as *SEU-Accuracy*. We also examine our choice for estimating the probability distribution over the values that a prospective acquisition may produce. As we discuss in Section 2.2, prior research has concluded that employing only predictors whose values are known during inference improves prediction significantly. Based on these findings in this study we conditioned the estimation on the (known) class label. To validate the merits of this approach we compare it to two alternative methods that reflect two extremes with respect to reliance on predictors and their availability at inference time. The first is a simple approach which computes the *unconditional frequency* of feature-values, based simply on their frequency in the training data. We refer to this approach as *SEU-Frequency*. In the second approach, we use tree induction, employing all other features and the class label as predictors to estimate the probability distribution of a prospective acquisition. This approach, *SEU-DT* aims to capitalize on the interactions between predictors for inference. However, as discussed in Section 2.2, inference is likely to suffer if predictors whose values are unknown must be estimated at inference time. The conditional distribution approach we employ in SEU lies between these two extremes—rather than relying on unknown values or estimating a simple unconditional distribution, SEU conditions the estimation on the known label.

The Oracle Policies. To derive an upper bound for SEU's performance, we employ an omniscient policy (the *Oracle*) that knows the true values of missing features to determine the feature-value acquisition that will lead to the greatest improvement in generalization performance. In addition, we assume the Oracle has access to the held-out test data so as to compute the actual improvement in Log Gain following an acquisition. As with SEU, to render the evaluation feasible, rather than evaluate all possible acquisitions the Oracle selects the best acquisition among a sample of $\alpha\beta$ prospective acquisitions. Both policies select prospective acquisitions from the same set of prospective acquisitions. We also present experiments which decompose the advantages conferred by the Oracle over the imperfect estimations performed by SEU. Specifically, we decompose the relative advantages into the Oracle's knowledge of the true model's performance measured over the held-out test data as compared to SEU's estimation over the training set, and the Oracle's knowledge of the true values of prospective acquisitions as compared to SEU's *estimation* of the *expected* benefits from prospective acquisitions over all possible values that a missing feature may have. Recall that in SEU we estimate the distribution of a missing value to compute the benefits from its acquisition *in expectation*. If the actual values of prospective acquisitions were known one could compute the benefit to model induction from acquiring the corresponding value directly rather than estimate these benefits in expectation. To evaluate the upper-bound performance that can be obtained by SEU if the actual values of missing features were known, we constructed a new policy, the *Feature Oracle*, that has access to the true values of prospective acquisitions for the purpose of evaluating acquisitions. Both SEU and the Feature Oracle estimate the benefit to induction over the training set and evaluate the same set of prospective acquisitions in each acquisition phase. To evaluate the benefits from assessing the model's accuracy directly over the test data we compare SEU's performance to that of the *Performance Oracle*—a policy in which the improvement in Log Gain is computed on the test data. We fix all other components of the polices so that the Performance Oracle and SEU employ the same measure to estimate feature-value distributions and evaluate the same consideration set of prospective acquisitions at each acquisition phase. We compare SEU to the *Performance Oracle* and *Feature Oracle* to estimate how improvements in each of SEU's estimations can contribute to SEU's overall performance so as to approximate the upper-bound performance.

Sensitivity Analyses. Finally, we explore the performance of SEU under different settings. The first of these evaluations considers the robustness of SEU's performance under different feature-value acquisition costs. SEU also incorporates several parameters such as the size of the sample of feature-values whose contributions to learning are evaluated by SEU, the number of feature-values acquired in each acquisition phase, and the size of the initial sample provided to SEU for evaluating the contributions of prospective acquisitions. We explore in turn how each of these design parameters affects SEU's performance (Langley 2000, Hevner et al. 2004).

Experimental Setup. The empirical evaluations are performed over a set of data sets from a variety of domains. Four data sets,[4] *expedia*, *etoys*, *priceline*, and *qvc* contain information about web users and their visits to large retail web sites. The target (dependent) variable indicates whether a user made a purchase during a visit. The predictors describe customers' surfing behaviors at the site as well as at other sites over time. We induce models to estimate whether a purchase will occur during a given session and employ the acquisition policies to estimate which unknown feature-values are most cost-effective to acquire. These data sets contain both continuous and categorical features; therefore for simplicity when estimating value distributions we converted all the continuous features to categorical features using the discretization method of Fayyad and Irani (1993). The remaining data sets are available from the UC Irvine repository (Blake and Merz 1998) and pertain to a variety of domains.

[4] From the related study by Zheng and Padmanabhan (2006).

The performance of each acquisition policy is evaluated over 10 independent runs of 10-fold cross-validation as follows. For each cross-validation run, the 10-fold partition was selected at random. In each fold of the cross-validation, all policies were provided with the same subset of initial feature-values, drawn uniformly at random from the training portion. All the remaining feature-values in the training data constitute the initial pool of potential acquisitions. At each acquisition phase, each policy acquires the values of a set of queries from the pool of prospective acquisitions; then, a new model is induced and its classification accuracy is measured on the test data. This process is repeated until a desired number of feature-values has been acquired. To reduce computation costs in the experiments, we acquire queries in fixed-size batches at each iteration. For problems where learning requires more training information, we acquired a larger number of feature-values at each phase. For each data set, we selected the initial random sample size to be such that the induced model performed at least better than assigning all instances to the majority class. We later explore the policy's performance for smaller numbers of initial feature-values and for different batch sizes. The test data set contains complete instances to allow us to estimate the true generalization accuracy of the constructed model. We set the exploration parameter α to 10. For model induction we used J48 classification-tree induction, which is the Weka (Witten and Frank 1999) implementation of C4.5 (Quinlan 1993). Integral to this induction algorithm is a missing value treatment, enabling induction from the incomplete data set. In addition, Laplace smoothing was used with J48 to improve class probability estimates.

We compare the performance of any two policies, A and B, by computing the percentage reduction in classification error rate obtained by A over B at each acquisition phase and report the average reduction over all acquisition phases. We refer to this average as the *average percentage error reduction* (Saar-Tsechansky and Provost 2004). The reduction in error obtained with policy A over the error of policy B is considered to be significant if the errors produced by policy A are lower than the corresponding errors (i.e., at the same acquisition phase) produced by policy B, according to a paired t-test ($p<0.05$) across all the acquisition phases. The learning curves that we present below and the average percentage error reduction we report reflect average performance of each policy over the 10 runs of 10-fold cross-validation.

3.2. Results

The Primary Results. Table 1 presents the average error reductions obtained by different SEU policies with respect to the uniform sampling policy, which acquires a representative set of feature-values drawn uniformly at random. The number of acquisitions, β, acquired at each acquisition phase and the size of the initial sample are also presented in Table 1. In this and subsequent tables each significant value ($p<0.05$) is marked with an asterisk (*).

Let us first examine whether SEU effectively decreases classification error as compared to a uniform sampling policy. In Table 1, the fourth and fifth columns present the average error reductions obtained

			Consideration set alternatives		Alternative for $A(F, F_{i,j}=v_k)$	Alternatives for $P(F_{i,j}=v_k)$	
Data Set	β	Initial Sample	SEU-US	SEU-ES	SEU-Accuracy	SEU-Frequency	SEU-DT
	(batch size)	(no. of instances)					
audiology	100	147	14.61^{*}	19.72^{*}	$7.81^{*\ddagger}$	$8.38^{*\ddagger}$	$11.31^{*\ddagger}$
car	50	1033	10.93^{*}	11.11^{*}	$4.25^{*\ddagger}$	$8.14^{*\ddagger}$	9.42^{*}
eToys	100	125	19.11^{*}	49.18^{*}	$10.17^{*\ddagger}$	$17.99^{*\ddagger}$	$16.14^{*\ddagger}$
expedia	100	350	10.04^{*}	16.61^{*}	$6.38^{*\ddagger}$	$5.92^{*\ddagger}$	11.03^{*}
lymph	20	38	7.20^{*}	3.05^{*}	$5.56^{*\ddagger}$	6.30^{*}	6.70^{*}
priceline	100	75	10.69^{*}	$1.24^{\dagger}$	$4.46^{*\ddagger}$	9.82^{*}	$9.17^{*\ddagger}$
qvc	100	225	3.76^{*}	14.82^{*}	$-0.20^{\ddagger}$	2.83^{*}	$1.30^{*\ddagger}$
vote	10	59	18.30^{*}	8.33^{*}	$6.23^{*\ddagger}$	$12.83^{*\ddagger}$	$15.32^{*\ddagger}$
Average			11.83	15.50	5.58	9.02	10.04

*Policy is better than uniform, $p<0.05$ ($^{\dagger}p<0.06$) ‡SEU-US is better than the alternative policy, $p<0.05$

Table 1 Error reductions of SEU variants as compared to a uniform random acquisition policy.

by SEU-US and SEU-ES with respect to uniform query sampling. Figure 2 presents the performance of the three policies on four data sets that exhibit the different patterns of performance we observe. For all data sets *Sampled Expected Utility* builds more accurate models than uniform query sampling. The differences in performance on all data sets, except for SEU-ES on *priceline*, are statistically significant.[5] These results demonstrate that the expected utility framework and the specific methods we employ to estimate the expected improvement in performance are indeed effective for AFA: *Sampled Expected Utility* selects queries that on average are more informative for induction than queries selected uniformly at random.

To underscore the advantage of using SEU, one can observe the cost benefit of using SEU to build a model exhibiting a desired performance level as compared to using a uniform acquisition policy. For example, on the *etoys* data set, uniform query sampling had to acquire approximately 1800 feature-values in order to obtain an accuracy of 94%. SEU-ES had to acquire fewer than 400 feature-values to achieve the same accuracy. When data-acquisition costs are considerable, this could translate to substantial savings in the cost of building accurate models.

The results also indicate that the method employed to select the queries to be considered for acquisition can have a significant impact on the outcome. Both SEU-US and SEU-ES acquire useful feature-values that significantly improve the model's performance. For some data sets, such as *etoys* and *audiology*, SEU-ES selects significantly more informative acquisitions than SEU-US, suggesting that *Error Sampling* identifies a superior subset of acquisitions to be considered for acquisition than those drawn on average via uniform query sampling. In other data sets, e.g. *priceline*, SEU-US is preferable. Recall that *Error Sampling* selects

[5] Note that for SEU-ES on *priceline*, the improvement at least is significant at the 0.06 level.

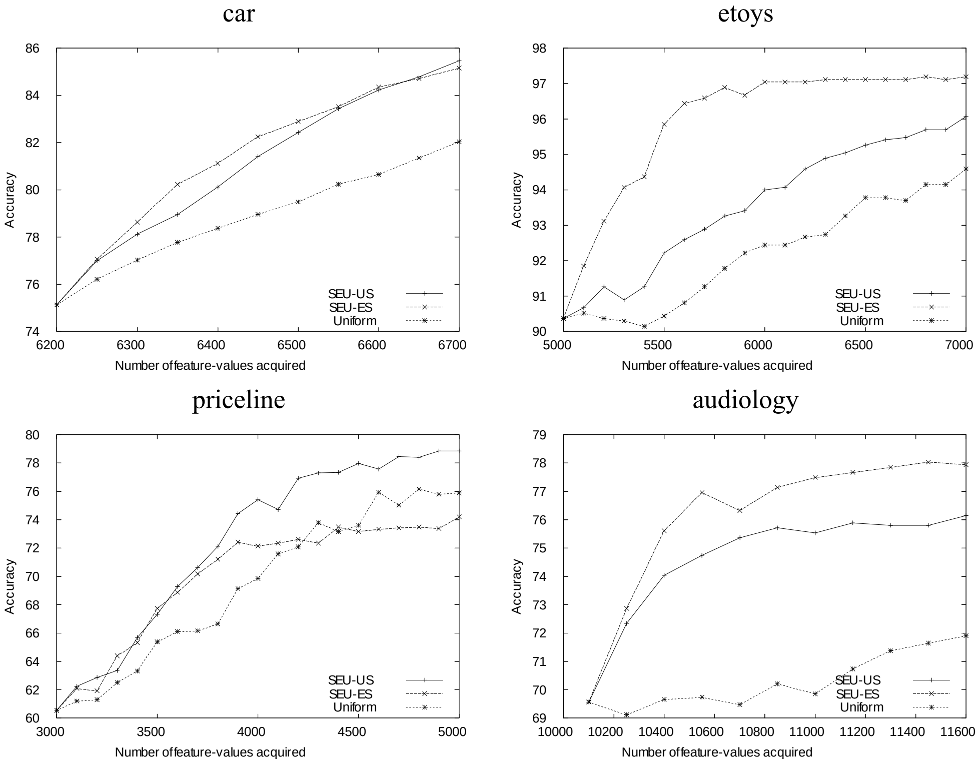
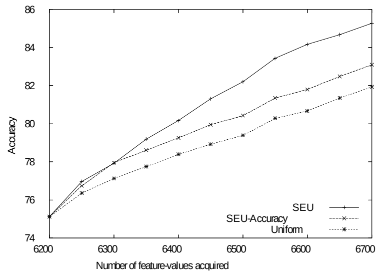


Figure 2 **Four characteristic patterns of the improvement of classification accuracy as a function of the number of feature-values acquired, assuming uniform feature costs.**

entire instances, and thus all the missing feature values for these instances simultaneously become candidates for acquisition. Fleshing out a smaller set of examples may be more or less preferable in different domains. Also, if examples are incorrectly classified simply because they are outliers, *Error Sampling* will stumble because it will prefer all the unknown features for these examples. Moreover, if only a few features are relevant, selecting all the features will only dilute the candidate set.

The average error reduction obtained with SEU-US over all acquisition phases ranges between 3.76% and 19.11%. SEU-ES often results in even more substantial savings, but its performance is more varied than that of SEU-US. The average error reduction obtained by SEU-ES ranges between 1.24% and 49.18%. Because it forms the consideration set based on the entire instance, Error-Sampling may sometimes fail to select instances with a highly informative feature-value if the entire instance seems less informative as compared to another instance.

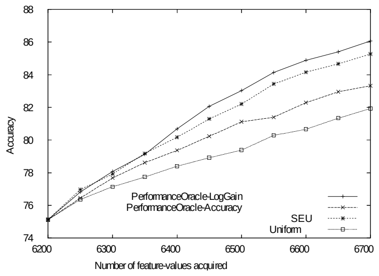
In sum, both SEU policies provide considerable advantage over uniform query sampling. SEU-ES usually is the better of the two, and sometimes can provide very substantial savings. SEU-US is more consistent and thus would be a more conservative choice. In the next section we expand on these results, focusing on the more conservative SEU-US policy. For the remainder of this paper, unless specified otherwise, SEU will refer to SEU-US.



(a) car

Data Set	Oracle-LogGain vs. Oracle-Accuracy
audiology	2.10*
car	8.34*
etoys	0.70
expedia	3.45*
lymph	6.83*
priceline	2.45*
qvc	2.94*
vote	6.52*

$^*p < 0.05$ (b)

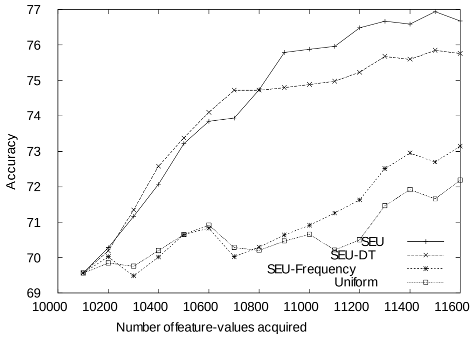


(c) car

Figure 3 Classification accuracy vs. Log Gain for estimating the value of an acquisition. The three comparisons are described in the text.

We now compare SEU's performance to its performance using alternative utility measures, providing empirical support for the desired properties outlined in Section 2. First we examine empirically whether Log Gain is a better measure of prospective utility than is classification accuracy. The sixth column of Table 1 shows the average error reduction obtained by SEU-Accuracy over uniform sampling. We mark with an asterix (*) the data sets where SEU-Accuracy is significantly better than uniform acquisition, and with a double-dagger (‡) the data sets for which SEU is significantly better than SEU-Accuracy (all of them). Figure 3(a) shows the performance of each policy as well as of uniform sampling for the *car* data set. The improvements obtained by SEU (with Log Gain) can be substantially higher than those obtained by SEU-Accuracy, up to more than 12% average error reduction. These results demonstrate that by capturing changes in the probability estimation, Log Gain indeed is able to select significantly more informative feature-values to acquire, leading to better models on average.

A possible reason for the inferior performance obtained by SEU-Accuracy may be the difficulty of precisely estimating classification accuracy using only the training data. While in practice only the training data are available to the SEU policy, it is useful to establish whether Log Gain is more informative even when an oracle computes the value of prospective acquisitions directly on the test data, or whether classification accuracy is preferable when it is estimated with sufficient precision, in spite of its step-like form. To address this question, we compared SEU's performance to versions of the policy where Log Gain and classification accuracy are measured on the *held-out test data*, rather than on the training data. We refer to these policies as *Performance Oracle-LogGain* and *Performance Oracle-Accuracy*, respectively. Figure 3(b) presents the error reduction obtained with Performance Oracle-LogGain as compared to Performance Oracle-Accuracy. For the Car data set, Figure 3(c) presents the performance obtained by the oracles, SEU, and the uniform acquisition policy. The results confirm the advantage from detecting changes in the model's



audiology

Figure 4 Three different methods for estimating a feature's value distribution, as compared to uniform acquisition.

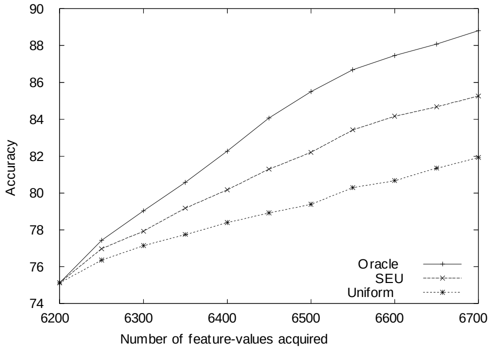
probability estimation via Log Gain over classification accuracy—Log Gain is able to identify more informative acquisitions even when the impact of an acquisition is evaluated with absolute precision over the test data.

Now, let us turn to the value distribution estimation employed by SEU. Table 1 presents average reductions in error when using conditional distributions (SEU, fourth column) as compared to using unconditional frequency estimation (SEU-Frequency, seventh column) and tree induction (SEU-DT, eighth column). For the *audio* data set, Figure 4 shows the performance of SEU with each estimation approach. These results suggest that unconditional distributions provide poorer estimates of the feature-value distribution as compared to the conditional distributions; the average improvement over the uniform policy over all data sets is 9.02% as compared to 11.83% obtained by SEU. For individual data sets SEU's relative advantage with respect to SEU-Frequency reached up to 6.23%. We also find that SEU is often better or comparable to SEU-DT which can capture more complex patterns, but relies on predictors that may be unknown at inference time. Errors in estimating missing predictors or their distributions contribute to prediction error cumulatively, and furthermore, because the induction technique implicitly assumes all predictors will be available during inference, it is less likely to capture alternative predictive patterns involving feature-values that will be available during inference than when it relies exclusively on known predictors (Saar-Tsechansky and Provost 2007).

The Oracle Policies. Let us now examine the upper-bound performance that can be obtained with SEU. Figure 5(a) shows the average error reduction obtained by the Oracle as compared to SEU. Figure 5(b) shows the performance obtained by the Oracle, SEU, and the uniform policies for the car data set. The Oracle obtains between 0.63% and 14.9% average error reduction and its benefit is statistically significant in most cases. We also measured the error reduction obtained by each, the Oracle and SEU, as compared to uniform sampling in order to discover what proportion of the improvement obtained by the Oracle is

Data Set	Oracle vs. SEU	SEU Error Reduction as % of Oracle's
audiology	9.47*	49.21
car	13.11*	53.17
eToys	3.11*	40.51
expedia	4.09*	42.89
lymph	14.90*	36.38
priceline	0.63	56.54
qvc	2.42*	47.31
vote	0.66	62.43

*p<0.05 (a)



(b) car

Figure 5 Error reduction of the omniscient Oracle as compared to SEU. SEU achieves about half the total error reduction over uniform acquisition.

obtained by SEU. In Figure 5(a) we denote this measure as "SEU Error Reduction as % of Oracle's." SEU consistently achieves about half the "optimal" error reduction.

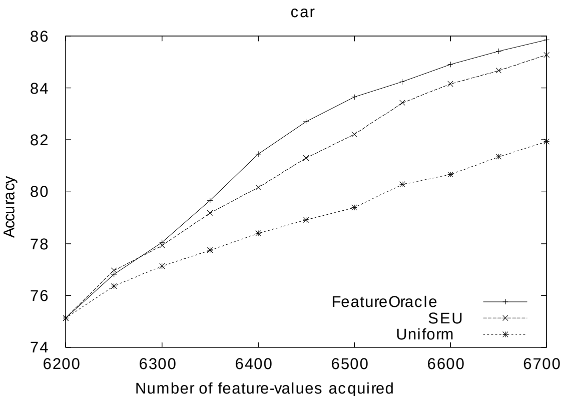
We can decompose the advantages conferred by each of the Oracle's perfect measures of the quantities in equation 1 over the corresponding imperfect estimations performed by SEU. Figure 6(a) presents the average error reduction obtained with the Feature Oracle as compared to the SEU policy. For the car data set Figure 6(b) shows the performance of the Feature Oracle, SEU and uniform sampling. As shown, acquisitions made by the SEU policy often result in models that perform comparably to those induced with acquisitions made by the Feature Oracle. The Feature Oracle performs statistically significantly better in only three data sets; in these cases the Feature Oracle's acquisitions lead to models that are between 2.19% and 4% more accurate than those produced by SEU, on average. Thus, for the purpose of choosing acquisitions based on computed expected utility, SEU estimates the distribution of missing values fairly well; however, in principle, there is some room for improvements so as to reach the Feature Oracle's performance. As we show in Section 3, models that are expected to perform well in this setting are those that can produce an estimation of the distribution without the need to impute or estimate the distribution missing predictors. Approaches that are more computationally intensive (and generally more accurate) than the one we employ may improve the performance further (Saar-Tsechansky and Provost 2007).

To evaluate the benefits from assessing the model's accuracy directly over the test data we compare SEU's performance to that of the Performance Oracle. Figure 7(a) presents the average percent error reduction obtained with the Performance Oracle as compared to SEU. With the exception of a single data set, the Performance Oracle's ability to rank potential acquisitions by their impact on the test data results in models that are statistically significantly more accurate than those induced with SEU. The Performance Oracle produced models that are 2.17% to 6.29% more accurate than those obtained with SEU.

Our decomposition of the Oracle's performance suggests that its access to the held-out test data confers

Data Set	Feature Oracle vs. SEU
audiology	-1.93
car	3.87*
eToys	-2.60
expedia	4.00*
lymph	2.19*
priceline	-2.24
qvc	-2.08
vote	0.64

*p<0.05



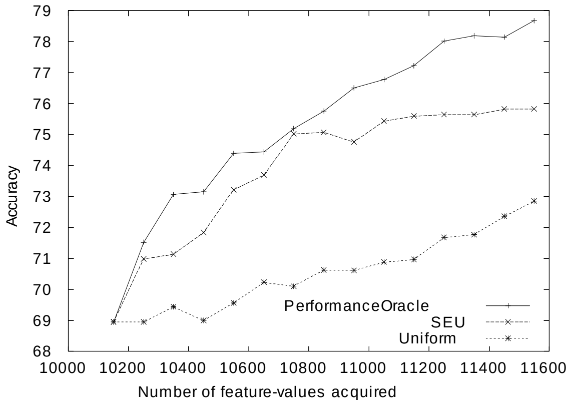
(b)

Figure 6 Oracle with complete knowledge of feature values (only) as compared to SEU.

Data Set	Performance Oracle vs. SEU
audiology	5.63*
car	2.74*
etoys	6.29*
expedia	3.56*
lymph	2.17*
priceline	2.64*
qvc	-0.18
vote	5.05*

*p<0.05

(a)



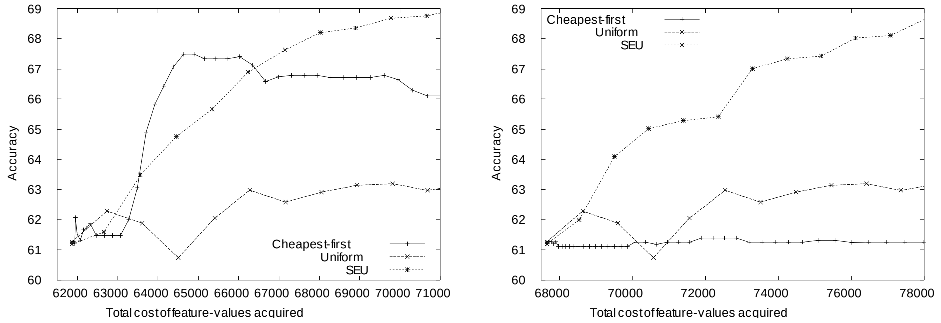
(b) audiology

Figure 7 Oracle with complete knowledge of test-set performance (only) as compared to SEU.

the most significant advantage to its performance over SEU. Unfortunately, in practice, this information cannot be available to an acquisition policy. SEU's estimation of the feature-value distribution already results in a comparable performance to that of the Feature Oracle for most data sets. However, comparing the decomposed results to the results of the (complete) Oracle, we see that there seems to be an interaction that leads to the overall error reduction being larger than the sum of the individual reductions.

3.3. Sensitivity Analyses

We now explore SEU's performance under different experimental settings. In the first set of experiments we evaluate SEU's performance when attributes vary largely both in the information they provide about the class and in their costs. To make the problem setting challenging, we constructed synthetic data in the following way. For the lymph data set, which has 18 features, we added an equal number of binary features with randomly selected values so as to provide no information on the class. In addition, for each feature we associate a cost drawn uniformly at random from 1 to 100. We evaluated the policies' performances for 5



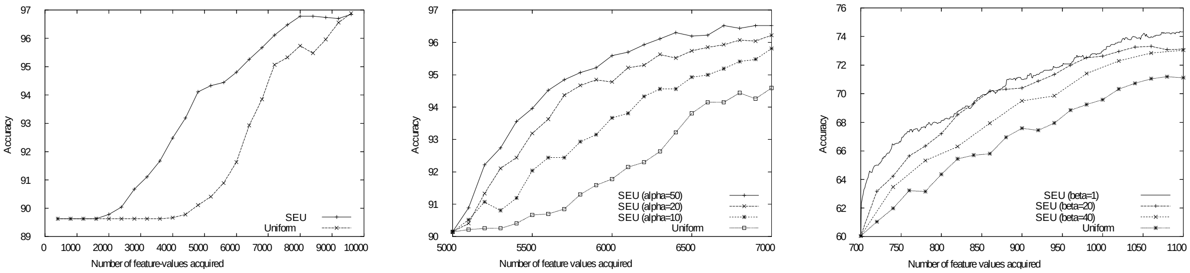
(a) Acquisition-cost structure 1 (b) Acquisition-cost structure 2

Figure 8 Comparing acquisition policies under different acquisition-cost structures.

different assignments of feature costs.

Since uniform sampling does not take feature costs into account, we also compare SEU with a baseline strategy that does. This approach, Cheapest-first, selects feature-values for acquisition in order of increasing cost. The results for all randomly assigned cost structures show that for the same cost, SEU consistently builds more accurate models than the uniform policy. Figure 8 presents results for two representative cost structures. SEU's superiority is more substantial than that observed with uniform costs for the original data sets. This is because SEU's ability to capture the value of acquisitions per unit cost is more critical when there are features of varying information value and cost. In contrast, the performance for Cheapest-first is quite varied for different cost assignments. When there are highly informative features that are inexpensive, Cheapest-first of course performs quite well (e.g., Figure 8(a)), since its underlying assumption holds. In such cases, SEU would not perform as well because it imperfectly estimates the expected improvement from each acquisition. On the other hand, when many inexpensive features are also uninformative (probably a more realistic scenario), Cheapest-first performs worse than the uniform policy (Figure 8b). SEU, however, estimates the trade-off between cost and expected improvement in performance, and although the estimation clearly is imperfect, it consistently selects better queries than random acquisitions for all cost structures.

We now examine in turn how each of SEU's parameters affects its performance. SEU requires some training data to estimate the expected contribution to induction of prospective acquisitions. In the experiments so far, we evaluated the performance of SEU once the model induced from the initial sample performs comparably to a majority classifier. Here, we also explore how SEU performs when it is provided with a smaller initial sample up until it exhausts the pool of potential acquisitions. As before, we initialized the training set to a random set of feature-values, and the number of values is equal to 10 times the number of features for the corresponding data set. A representative pattern that we observed is presented in Figure 9(a) for the eToys data set. As shown, because of the small amount of training data both policies require additional data to produce predictive patterns and improve performance. However, the acquisitions made



(a) Exhaustive acquisition of values (b) SEU with different α values (c) SEU with different β values

Figure 9 Changes to SEU's parameters have the expected effects on performance.

by SEU are substantially more informative and thus SEU acquires fewer features to achieve a given level of performance. Thus, even when only a very small number of values are available initially, it is preferable to employ SEU. In addition, as is typical with information-acquisition policies, once both policies exhaust the pool of acquisitions the performances of the models they produce converge.

To alleviate computational costs, SEU evaluates only a subset of prospective acquisitions at each acquisition phase. The size of this consideration set, as determined by the parameter α, is likely to affect SEU's performance. For the eToys data set, Figure 9(b) presents SEU performance for different consideration-set sizes. The results are quite intuitive—the larger the consideration set size, the more likely it is that more informative feature values are identified and acquired, improving the model's performance. As shown, even for a small α value of 10, SEU acquires more informative acquisitions on average than those selected by uniform sampling. If there are no computational constraints, however, a larger consideration set clearly is preferable.

The SEU policy evaluates the expected contribution of each individual feature-value acquisition. However, in practice and in the experiments conducted in this study, more than a single feature-value may be acquired simultaneously in each acquisition phase. While we find that SEU is effective in this setting, it is important to explore how SEU with batch acquisitions performs compared to when a single value is acquired at each phase. For the Lymph data set Figure 9(c) shows SEU performance when a single value is acquired at each phase, as compared to when multiple feature-values are acquired simultaneously. Here as well the results are quite intuitive. Because SEU estimates the contribution of single values, it performs best when it acquires one value at a time; similarly, when batch acquisitions are performed, SEU performs better the smaller the batch size is. When there are no significant computational constraints, the acquisition of an individual value at each iteration is preferable.

4. Active Information Acquisition

Now we demonstrate further the value of the expected utility framework by exploring its extension to a more general task. In principle, other sorts of information besides feature-values also may be acquired, at a cost, to benefit induction. We refer to the task of simultaneously evaluating the acquisition of any information pertinent to induction as Active Information Acquisition (AIA). We consider the case of AIA for which both unknown feature-values and class labels can be acquired. The motivation for this problem combines the motivations for traditional active learning (Cohn et al. 1994) and for active feature-value acquisition. Consider, for example, data used to model customers' responses to an offer. Different feature-values may be missing for different customers, and the responses to offers for particular customers can be acquired at a cost. The latter costs may stem from contacting a customer or from the opportunity cost arising from offering a sub-optimal offer to a potential buyer. Given these acquisition costs it would be beneficial for an AIA policy to suggest what would be the most effective acquisitions.

The expected utility framework extends directly to handle this problem. The advantage of our approach is that it evaluates all acquisition types by the same measure—i.e., the marginal expected contribution to the predictive performance per unit cost. In Algorithm 1, missing classes can be included as potential queries in step 2. As a technical point, in this setting we cannot use class-conditional distributions to estimate the feature-value distributions of missing features (or classes), since we do not have class labels for all instances. Instead, we will use an instance of the base learner (tree induction in this case) to estimate the value distribution of the feature under consideration, as done in SEU-DT in Section 3. We refer to this policy as Sampled Expected Utility-AIA (SEU-AIA).

4.1. Determining the Consideration Set for AIA

To make the expected utility approach tractable, here too we reduce the set of candidate queries. However, in the AIA setting, drawing the consideration set uniformly at random would be biased against missing class labels—there typically are fewer class labels than feature-values and a uniform sample would tend to reflect this. Such a bias could be detrimental to induction because a class label tends to be much more informative than a single feature value. To reduce the set of queries evaluated by AIA, we employ a computationally inexpensive heuristic which aims to capture the relative information value of a prospective acquisition per unit cost, before it is explicitly computed by AIA. Specifically, we compute a weight for each prospective acquisition that is proportional to an estimation of the information conveyed by this value for model induction (described next), normalized by the value's cost.

More specifically, for candidate-set reduction, we evaluate the contribution to induction of all feature-values of a given a feature F_i by a cost-normalized variant of the information gain $IG(F_i, L)$(Quinlan 1993) of the feature F_i for class variable L. The information gain is given by $IG(F_i, L) = H(L) - H(L|F_i) = H(L) - \sum_j p(F_i = v_j) H(L|F_i = v_j)$, where $H(Z)$ denotes Shannon's information entropy of variable Z, and feature F_i can have one of j values $v_1, ...v_j$. The information gain captures the reduction in the entropy of the class variable once the value of feature F_i is known. Thus features that carry more information for determining the class will have higher information gain and also are more valuable to induction. The score assigned to a feature-value is the corresponding feature's information gain normalized by the feature-value's cost. Thus feature-values with high information gain and lower costs will be assigned higher weights. These weights then guide the sampling of the consideration set.

In supervised learning, instances whose class labels are missing are not used for induction, thus when labels are acquired the values of the known feature values of the respective instance also become available for induction. To capture the value to induction when a class label is acquired, we compute the sum of information gains for all the known features of the respective instance. Formally, let M_k denote the set of all known feature-values for instance k, then the weight assigned to the value from acquiring the label of instance k is give by $\sum_{F_i \in M_k} IG(F_i, L)$.

The consideration set is composed of the prospective acquisitions with the highest weights. This selection scheme is tractable because it does not require intensive computations for each missing value, and estimates the same information gain for all queries of a given feature. Once the consideration set is determined, SEU is applied to estimate the expected value of each individual query in the set.

4.2. SEU-AIA Performance

To assess the performance of SEU-AIA for this general information acquisition task, we remove class labels and feature-values from the training data uniformly at random. We compare the performance of SEU-AIA to acquiring missing values uniformly at random. For this set of experiments we assume that all features and class labels have the same cost; next, we will present experiments with non-uniform costs. The purpose of this comparison is to verify that SEU-AIA effectively estimates the expected contribution of missing values of both types, so as to rank them accurately, and to produce better predictive models for a given cost.

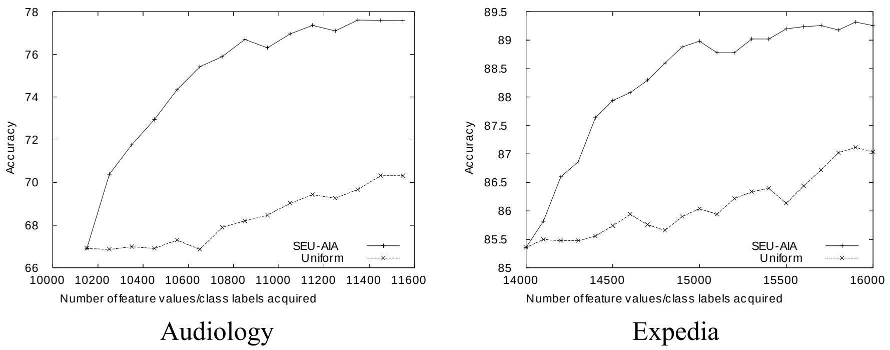
Table 10(a) presents a summary of results comparing SEU-AIA with uniform sampling. On all data sets, acquiring information using SEU-AIA results in significantly better models than using uniform acquisition. Figure 10(b) shows the results for audiology and expedia, demonstrating the substantial impact. SEU-AIA consistently acquires informative values for modeling, which result in models superior to those obtained by uniform acquisition. SEU-AIA evaluates and compares the different types of information effectively and provides a significant lift in predictive performance. To our knowledge, this is the first demonstration of an effective policy for this general information acquisition problem.

To gain further insight into SEU-AIA's choice of acquisitions we compare experimentally the performance of SEU-AIA with that of an active learning (AL) policy, which employs SEU-AIA, but only considers class labels for acquisitions, and to SEU which only evaluates feature acquisitions. We perform these evaluations under different cost structures in which either class labels or feature-values are significantly

Data Set	SEU-AIA vs. Uniform
	Average
audiology	21.29*
car	0.52*
eToys	39.71*
expedia	15.97*
lymph	16.42*
priceline	28.54*
qvc	23.47*
vote	20.05*

$^*p<0.05$

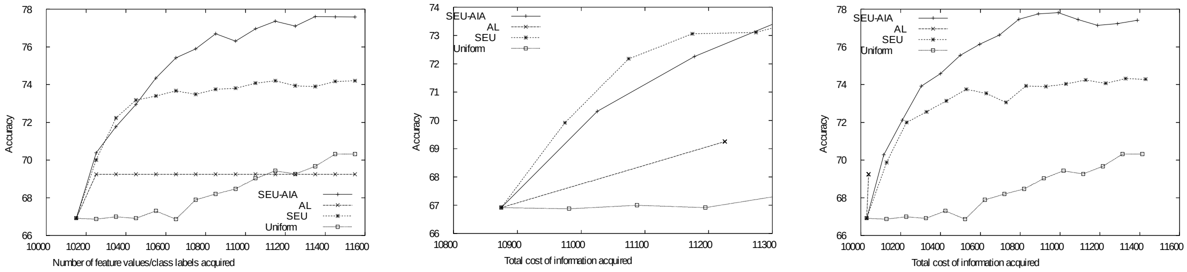
(a)



Audiology Expedia

(b)

Figure 10 Active Information Acquisition vs. Uniform Random Sampling when both features and class labels must be acquired at a cost. (a) Average error reductions, and (b) & (c) accuracy reductions.



(a) Feature cost = label cost (b) Label cost = 6 x feature cost (c) Feature cost = 6 x label cost

Figure 11 Different costs structures favor different basic acquisition strategies. SEU-AIA shows robust performance, adjusting acquisitions based on costs.

more expensive to acquire. These comparisons allow us to assess whether SEU effectively manages acquisitions of class labels and of feature-values so as to produce models that are comparable or superior to those produced with either active learning or AFA alone.

Figure 11 shows classification accuracy as a function of acquisition cost for SEU-AIA, Active Learning (AL), SEU, and uniform acquisition for three different cost structures. Figure 11(a) shows results for a cost structure in which class labels and feature values are equally expensive, whereas Figures 11(b) and (c) present results for cost structures in which feature-values or class labels are significantly more expensive, respectively. Note that as there are fewer class labels than feature-values, the curves describing AL performance may appear truncated, particularly if label costs are significantly lower than feature costs or vice versa.

Our results suggest that while policies that consider the acquisition of only one type of information can perform well for cost structures in which the values they acquire are informative and inexpensive, their performance is inconsistent and they perform poorly with other cost structures. By contrast, SEU-AIA effectively handles the trade-off between the informativeness of different types of values and the cost of acquiring them, providing consistent, good performance across all cost structures. For example, for the cost structure shown in Figure 11(a), no one type of information is consistently more cost-effective than the other and it therefore is beneficial to alternate between acquisitions of class labels and of feature-values. As shown in Figure 11(a), SEU-AIA performs better than each of the other policies, which consider only the acquisition of class labels or of feature-values. This confirms that acquisitions of both feature-values and class labels is more cost-effective. For the cost structures in which feature-values are significantly more expensive than class labels or vise versa, either AL or SEU, respectively, perform best. This is because the extreme cost structure leads to one type of acquisitions being consistently more cost-effective than the other. While SEU-AIA's estimation of the value of prospective acquisitions is imperfect, its performance in these settings approximates that of the better of AL or SEU, demonstrating that SEU-AIA accurately estimates the usefulness of acquisitions that are more cost-effective, regardless of their types. By contrast, the policies which consider the acquisition of only one type of information, namely AL and SEU, perform poorly in one of the settings. For example, in the cost settings shown in Figure 11(c) AL performs very well. However, AL performs poorly for the cost structure in Figure 11(b) because it does not consider the highly cost-effective feature-values that can be acquired. Because it is not known a priori how policies that acquire only feature-values or only class labels will perform under different cost structures and given the varied performance of these policies, SEU-AIA is attractive, allowing one consistently to identify different types of cost-effective acquisitions.

5. Limitations and Future Work

Despite the effectiveness of the expected utility framework and the SEU policy, there are limitations that provide avenues for future work.

- We discussed in Section 1 why any policy for AFA must be heuristic, as generalization accuracy itself cannot be computed exactly. However, even for a heuristic measure of performance, the question of whether more complex optimization over all possible set acquisitions will improve AFA remains an open and interesting question. We note that an optimization procedure to identify the best of all possible acquisition sets can be thought of as a multiple comparisons procedure (Jensen and Cohen 2000), using a finite set of training data to estimate generalization performance of alternative models. Such settings may lead to pathologies such as overfitting and oversearch—indeed, there has been growing evidence that search for optimal "combinations" of factors for predictive modeling often is inferior to greedy search.

- In some settings it may be possible to acquire different sets of feature values for a single price. For example, access to different information sources, such as archived patient records of different care providers, may each be costly, but once a record is accessed a set of values can be acquired for a single price. Thus the problem becomes: which set of values would be most cost-effective to acquire next? In principle, our framework can be applied directly. However, estimating the value distribution of all possible assignments for a given set and computing the value to induction from each assignment would be very inefficient in practice. One possibility is to assume statistical independence of values in a given set, alleviating the joint value-distribution estimation. Subsequently, one may employ Monte Carlo estimation to draw value assignments for each prospective set to approximate the expected contribution to model induction.

- For estimating the value distribution of a prospective acquisition it is recommended to use only feature-values that will be known at inference time (Saar-Tsechansky and Provost 2007). In principle, one can model this distribution using any set of known predictors. As we note earlier, this approach can be inefficient with most modeling techniques, because it may require inducing a different model to capture the interactions among each unique set of known predictors. One alternative is to employ the naïve Bayes assumption of conditional independence, which allows inference without the need to estimate the values or the distributions of missing predictors. In addition, it is trivial to marginalize any missing variables by not including them during inference.

The models induced may be improved by using a missing value treatment that takes into consideration the potential bias in the missingness pattern created by a selective (non-uniform) acquisition policy. In the empirical evaluations in this paper we employ the treatment integral to C4.5. It would be informative for future work to explore the performance with other missing value treatments and modeling techniques.

- The Expected Utility framework allows one to incorporate performance objectives other than accuracy, such as the (economic) benefit from model use.

- If/once enough values are available, feature selection may be applied to complement AFA. Prospective value acquisitions pertaining to features which are excluded by feature selection may not be considered for acquisition. Otherwise, AFA itself would have to learn that these values provide no utility.

6. Related Work

The problem of sequential information acquisition has been addressed in a variety of settings, with various types of information being acquired to satisfy a variety of objectives. To the best of our knowledge, the policies we present are the first approaches designed for the problem of incrementally acquiring feature-values for inducing a general classifier when costs are specified for individual entries of F.

An early and highly influential stream of research pertains to the classic multi-armed bandit problem introduced by Robbins (1952). McCardle (1985) applies similar reasoning. The objective in the multi-armed bandit problem and the technology adoption decision problem is the estimation of single parameter or a set of independent parameters, e.g., the probability of success. Other researchers also have addressed information acquisition for decision making. Moore and Whinston (1986, 1987) develop a theoretical decision-making model for a decision-maker who can acquire information at a cost in order to reduce the uncertainty associated with a given decision. Feature-value acquisition for case-retrieval and inference algorithms was treated by Mookerjee and Mannino (1997) and Mannino and Mookerjee (1999), who study sequential feature-value acquisition for test instances. They address a setting where all feature-values of a test instance are missing but can be acquired sequentially during inference to minimize the overall acquisition costs. Mookerjee and Mannino also demonstrate that joining concept formation and retrieval strategy results in significantly lower acquisition cost during inference as compared to when the two phases are addressed independently. Differently from this prior work, we develop policies for acquiring information to improve predictive model *induction*.

The notion of information acquisition designed for predictive model induction has been addressed by several prior lines of work. Three decades ago, authors identified the significance of the interdependence between induction as a search in the space of all possible concepts/models and the selection of training data used to direct the search. Simon and Lea (1974) describe conceptually how induction involves simultaneous search of two spaces—the results of searching the model space can affect how training data will be sampled. Techniques from Optimal Experimental Design (Federov 1972) and from Active Learning (Cohn et al. 1994, Freund et al. 1997) assume *class labels* are unknown. Thus in active learning *complete* instances are acquired to enhance learning. A related problem to active learning was addressed by Zheng and Padmanabhan (2002, 2006), where instances with incomplete feature-values are not used for induction, and similar to active learning, a policy is proposed to identify useful *instances* for which to acquire complete information. Melville et al. (2004) address the same problem, but assume that incomplete instances can be used for induction, using some missing-value treatment. The approach we develop here for feature-value acquisition is inspired by the method proposed by Roy and McCallum (2001) for active learning. Roy and McCallum examine the expected improvement from acquiring class labels for naïve Bayes classifier induction. Our approach for active information acquisition presented in section 4, we generalize the AFA problem to include the class variable as another feature to acquire, thus subsuming both AFA and traditional active learning.

Also closely related, Lizotte et al. (2003) study *budgeted learning* where an amount to be spent towards feature-value acquisition is specified a priori. There are two main differences between AFA and the budgeted learning problem and policies proposed by Lizotte et al. First, the fundamental goals are different. AFA seeks acquisitions that will give the best model for any intermediate investment in information acquisition; thus, the order of acquisitions is critical. Budgeted learning, on the other hand, seeks the best model under the assumption that the entire budget will be spent on acquisitions; thus, the order of acquisitions is largely unimportant (beyond the critical question of acquire or not). The AFA setting is appealing because it allows reaching a performance goal on a fraction of the budget. The policies of Lizotte et al. (2003) also assume the induction of a naïve Bayes classifier with its conditional independence assumption, and equal costs for acquiring the values of a feature regardless of the instance. These assumptions combine to enable queries of the form, "Acquire the value of feature j for *any* instance in class k." In the AFA setting, the applicability to a general classifier (not just naïve Bayes) as well as the potentially variable cost for entries in the matrix F requires the ability to assign different values to the acquisition of a given feature for different instances. We experiment with and discuss the implications of different cost structures in Section 3.3. The main implication of the conditional independence assumption employed in Lizotte et al. (2003) is that it substantially limits the number of queries to be considered by the policy. In principle, the formulation of the policies proposed in Lizotte et al. (2003) can be extended to consider unique benefits from each feature-value acquisition (as in (Veeramachaneni et al. 2006)); however the computational complexity of their implementation may be prohibitive, particularly that of Single Feature Lookahead with a large lookahead. In Sections 2.3 and 4 we address means to reduce the consideration set of queries for the policies we propose. Williams et al. (2005) also address a similar problem; however differently from the work above they construct an acquisition policy designed specifically for a logistic regression model, assume that the modeling performance and acquisition costs must be given by the same units, and assume that the training data also define the complete set of instances for which prediction will be required.

Some work on *cost-sensitive* learning (Turney 2000), such as CS-ID3 (Tan and Schlimmer 1990), also attempts to minimize the cost of acquiring features during training; however, it only considers acquiring information for the current training instance. The LAC* algorithm (Greiner et al. 2002) acquires a random sample of complete instances in repeated *exploration* phases that are intermixed with *exploitation* phases, using the current classifier to classify instances economically.

Work on feature (variable/model) selection (John et al. 1994) assumes known feature-values and selects a subset of features to use for model induction. During feature selection the known feature-values are used to estimate the relative contribution from including *all* the values for a given feature. In principle, feature selection procedures could be complementary to an AFA policy, as discussed in Section 5.

7. Conclusions

We presented a general approach to active feature-value acquisition that acquires feature-values based on the estimated expected improvement in model performance per unit cost. We also proposed a specific measure for the utility of a prospective acquisition that captures the benefit from the acquisition given the dynamics of modeling with incrementally changing training data, and a method for estimating the distributions of possible query results in the presence of missing values. We showed how this computationally intensive policy can be made faster and remain effective by constraining the search to a sample of potential feature-value acquisitions. The resulting technique, Sampled Expected Utility, consistently yields better models per unit acquisition cost, when compared to uniform query sampling. The technique's advantage is particularly apparent when feature-values have varying information values and incur different costs.

We also studied SEU's component measures as compared to alternatives and to omniscient oracles, which know exactly the quantities being estimated. These studies reveal that SEU's measure of prospective feature-value distributions is largely effective, and that the greatest improvement in performance by the oracles is obtained when they know the exact impact of different values that may be acquired. A sensitivity analysis of the method's parameters produced intuitive results—SEU performs better when the consideration set of prospective acquisitions is larger and when the number of values acquired at each phase is smaller.

Finally, we showed how the SEU framework for feature-value acquisition can be effectively applied to address the more general information-acquisition problem in which missing (training) class labels and feature-values both may be acquired at a cost. SEU is able to alternate between acquisitions of class labels and of feature-values based on their relative contributions to learning per unit cost. In this general setting SEU produces better models for a given cost as compared to a uniform policy as well as compared to policies that acquire only feature-values or only class labels.

As we discussed in the introduction, the availability of a generic, effective, and computationally efficient policy for information acquisition offers opportunities to change the manner by which firms, which rely on consumer feedback to generate valuable business intelligence, interact with consumers to enhance data-driven intelligence cost-effectively. It also presents opportunities to transform business practices where information is acquired regularly in bundles: intelligent acquisition policies will allow firms to selectively identify only the most cost-effective values to acquire from potentially different sources to improve induction. Furthermore, such policies are likely to render information acquisition from third-party providers feasible for small businesses with potentially limited budgets.

The 1960 Siegel and Fournaker quote with which we started the paper is as relevant as ever. The meaning of "research purposes" has changed dramatically in half a century, especially with the million-fold improvement in computing power per unit cost. In this paper we have shown new ways to bring this tremendous computing power to bear to evaluate the value of information to improve decision making.

Acknowledgments

We thank the anonymous reviewers for their substantive comments. We are especially grateful to the associate editor for his/her excellent suggestions. We thank Duy Vu for his generous help in enhancing the code. We also thank Priyank Mishra, Hasrat Godil, Prateek Gupta, and Saurabh Baji for their help in code development.

References

Blake, C. L., C. J. Merz. 1998. UCI repository of machine learning databases. Http://www.ics.uci.edu/~mlearn/MLRepository.html.

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

Fayyad, U. M., K. B. Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. 1022–1027.

Federov, V. 1972. *Theory of optimal experiments*. Academic Press.

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

Friedman, J. H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**(1) 55–77.

Greiner, R., A. Grove, D. Roth. 2002. Learning cost-sensitive active classifiers. *Artificial Intelligence* **139**(2) 137–174.

Hevner, A., S. March, J. Park, S. Ram. 2004. Design science in information systems research. *MIS Quarterly* **28**(1) 75–105.

Huang, Zan. 2007. Selectively acquiring ratings for product recommendation. *Proceedings of the Ninth International Conference on Electronic Commerce (ICEC '07)*. 379–388.

Jensen, David D., Paul R. Cohen. 2000. Multiple comparisons in induction algorithms. *Machine Learning* **38**(3) 309–338.

John, George H., Ron Kohavi, Karl Pfleger. 1994. Irrelevant features and the subset selection problem. *Proc. of 11th Intl. Conf. on Machine Learning (ICML-94)*. 121–129.

Langley, P. 2000. Crafting papers on machine learning. *Proceedings of 17th International Conference on Machine Learning (ICML-2000)*. 1207–1212.

Lizotte, D., O. Madani, R. Greiner. 2003. Budgeted learning of naive-Bayes classifiers. *Proceedings of 19th Conference on Uncertainty in Artificial Intelligence (UAI-2003)*.

Mannino, M. V., V. S. Mookerjee. 1999. Optimizing expert systems: Heuristics for efficiently generating low cost information acquisition strategies. *Informs Journal on Computing* **11**(3) 278–291.

McCardle, K. 1985. Information acquisition and the adoption of new technology. *Management Science* **31**(11) 1372–1389.

Melville, P., M. Saar-Tsechansky, F. Provost, R. Mooney. 2005. An expected utility approach to active feature-value acquisition. *Proceedings of the IEEE International Conference on Data Mining*. 745–748.

Melville, Prem, Maytal Saar-Tsechansky, Foster Provost, Raymond Mooney. 2004. Active feature-value acquisition for classifier induction. *Proc. of 3rd IEEE Intl. Conf. on Data Mining (ICDM-04)*.

Mookerjee, V. S., M. V. Mannino. 1997. Sequential decision models for expert system optimization. *IEEE Transactions on Knowledge and Data Engineering* **9**(5) 675 – 687.

Moore, J.C., A. B. Whinston. 1986. A model of decision-making with sequential information-acquisition (part 1). *Decision Support Systems* **2**(4) 285–307.

Moore, J.C., A. B. Whinston. 1987. A model of decision-making with sequential information-acquisition (part 2). *Decision Support Systems* **3**(1) 47–72.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo,CA.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society* **55** 527–535.

Roy, N., A. McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*. 441–448.

Saar-Tsechansky, M., F. Provost. 2004. Active sampling for class probability estimation and ranking. *Machine Learning* **54** 153–178.

Saar-Tsechansky, M., F. Provost. 2007. Handling missing values when applying classification models. *Journal of Machine Learning Research* **8**.

Siegel, S., L.E. Fouraker. 1960. *Bargaining and Group Decision Making*. McGraw-Hill.

Simon, H. A., G. Lea. 1974. *Knowledge and Cognition*, chap. Problem solving and rule induction: A unified view. Potomac, MD: Erlbaum.

Tan, M., J. C. Schlimmer. 1990. Two case studies in cost-sensitive concept acquisition. *Proc. of 8th Natl. Conf. on Artificial Intelligence (AAAI-90)*. 854–860.

Turney, P. D. 2000. Types of cost in inductive concept learning. *Proceedings of the Workshop on Cost-Sensitive Learning at the 17th International Conference on Machine Learning*.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. John Wiley & Sons.

Veeramachaneni, Sriharsha, Emanuele Olivetti, Paolo Avesani. 2006. Active sampling for detecting irrelevant features. *Proceedings of the 23rd international conference on Machine learning (ICML-2006)*. 961–968.

Williams, D., X. Liao, L. Carin. 2005. Active data acquisition with incomplete data. Technical report, Department of Electrical and Computer Engineering, Duke University.

Witten, I. H., E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Zheng, Z., B. Padmanabhan. 2002. On active learning for data acquisition. *Proceedings of IEEE International Conference on Data Mining*. 562– 569.

Zheng, Z., B. Padmanabhan. 2006. Selectively acquiring customer information: A new data acquisition problem and an active learning based solution. *Mamangement Science* **52**(5) 697–712.