Active Feature-Value Acquisition

McCombs research paper series No. IROM-08-06, September 2007

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—the "training data"—from which the model is induced. 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 acquisition *policies*—procedures for producing acquisition *schedules*. An acquisition schedule is a ranking of potential information acquisitions, in this case currently unknown feature values. An ideal acquisition schedule would rank most highly those acquisitions that would yield the largest improvement in model quality per unit cost. We assume that model quality can be assessed objectively, for example as some function of the model's estimated predictive performance.

Similar to the sequential decision problem studied by Mookerjee and Mannino (Mookerjee and Mannino 1997, Mannino and Mookerjee 1999) we also address sequential acquisitions. Mannino and Mookerjee address the 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 missing values are acquired to improve predictive model *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 (a recommender system). Often, the 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, acquiring feedback from all consumers about all the products they have purchased is infeasible. A better acquisition policy would determine which ratings from which customers would be most cost-effective to acquire via costly incentives, so as 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. Similar to transformation in other information goods, information on consumers need not be acquired in bundles. Effective acquisition policies will enable firms to identify and 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.

This paper makes several contributions. In Active Feature-value Acquisition (AFA) for classifier induction (Melville et al. 2005), the cost of features may vary feature-to-feature and instance-to-instance. Given (i) a set of prospective acquisitions, (ii) the cost of acquiring each specific feature value, and (iii) an arbitrary classification-model induction algorithm, the general AFA problem is to produce an acquisition schedule so as to improve the induced model's performance for a minimum acquisition cost. To our knowledge, no prior work[1] addresses general AFA.

The primary contribution of this paper is the introduction of a method for solving AFA problems. We propose an acquisition policy that produces acquisition schedules based on estimates of the 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 the model's generalization accuracy, profit in a particular setting, or the expected cost of model error. Finally, the expected-utility 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. Because 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

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

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

2.1. Task Definition and Framework

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 matrix F, where $F_{i,j}$ corresponds to the value of the j-th feature of the i-th instance. 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. At any given time, the induction process may choose to acquire the value of $F_{i,j}$ at the cost $C_{i,j}$ to improve a given performance objective. The Active Feature-value Acquisition (AFA) task is to identify the feature value that would result in the highest improvement in performance per unit cost.

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; b – size of query batch; C – cost matrix for all instance-feature pairs;

1. Initialize $AcquisitionCost$ to cost of F
2. Initialize set of possible queries Q to $\{q_{i,j} : i = 1, ..., m; j = 1, ..., n;$ such that $F_{i,j}$ is missing$\}$
3. Repeat until stopping criterion is met
4. Induce a classifier, $M = L(T)$
5. $\forall q_{i,j} \in Q$ compute $score(q_{i,j}, L, T)$
6. Select a subset S of b queries with the highest $score$
7. $\forall q_{i,j} \in S$
8. Acquire values for $F_{i,j}$
9. $AcquisitionCost = AcquisitionCost + C_{i,j}$
10. Remove S from Q
11. End Repeat
12. Return $M = L(T)$

An iterative framework for AFA is presented in Algorithm 1. The framework is independent of the classification modeling technique of choice; it is given a learner, which includes a model induction algorithm and a missing value treatment to allow for induction from the incomplete matrix F.[2] At each phase a "score" is

[2] 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.

estimated and assigned to each prospective acquisition, reflecting the estimated added value of each potential acquisition per unit cost. The acquisition with the highest score is selected and the corresponding feature value is acquired; a particular approach to assign scores to prospective acquisitions will be described in detail in Section 4. Once a value is acquired the training data and the information acquisition costs are appropriately updated and this process is repeated until some stopping criterion is met, e.g. a desirable model accuracy has been obtained.

3. 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. McCardle's decision maker decides whether to acquire new information that may improve the estimation of technology success or to act based on the current knowledge. The objective in the multi-armed bandit problem and the technology adoption decision problem is the estimation of frequency parameters. 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 or inference algorithms was treated by Mookerjee and Mannino (Mookerjee and Mannino 1997) (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. In section 6, 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 in Lizotte et al. (2003). First, the fundamental goals are different. AFA seeks acquisition 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 acquisition; 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 in 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 regardless of their class. We experiment with and discuss the implications of different cost structures in Section 5.6. 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*, particularly that of Single Feature Lookahead policy with a large lookahead, would be hopelessly inefficient. In Sections 4.3 and 6 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 for a logistic regression model, assume that the modeling performance (termed risk) and acquisition costs must be given

by the same units, and 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 processes examples incrementally and only acquires 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. This suggests a policy of uniform random sampling of queries, which acquires a representative sample of missing data. We use this policy as a baseline for comparison.

4. Active Feature-value Acquisition Policies

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 associated 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 is 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, and so on. 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:

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

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 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 derivation of an acquisition schedule by ranking acquisitions 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 to induction from the acquisition, $\mathcal{A}(\cdot)$. 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 which 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 for each missing value $F_{i,j}$, 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 is computationally very intensive and is infeasible for most interesting problems.

4.1. Estimating an acquisition's contribution to performance

Let us consider the measure to be used for estimating the value to induction from an acquisition, $\mathcal{A}(F, F_{i,j} = v_k)$. New information aims to improve the model's classification accuracy, thus 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 the desired properties of a utility measure for information acquisition it is necessary to acknowledge 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 affects 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:

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

where $\mathrm{var}\,\hat{f}$ 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 $\mathrm{var}\,\hat{f}$ 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.

This observation suggests that new feature-value acquisitions that increase the average probability estimation of the true class can help to reduce the likelihood of erroneous classification. Unfortunately, as shown by the dotted line in Figure 1 (described in more detail below), classification accuracy itself is a rather crude estimate of the quality of the probability estimation.

Let us consider an alternative measure, *Log Gain* (LG). For a model induced from a training set F, let $\hat{f}_{C,F}(y|x)$ be the probability estimated by the model that instance x belongs to the *correct* class. Let $LG = -\log_2 \hat{f}_{C,F}(y|x)$. For a data set of k training instances, let the value to induction from an acquisition resulting in a training set F be given by the sum of Log Gains for the set of training instances: $\mathcal{A}(F) = \sum_{j=1}^{k} \log_2 \hat{f}_{C,F}(y_j|x_j)$. 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.

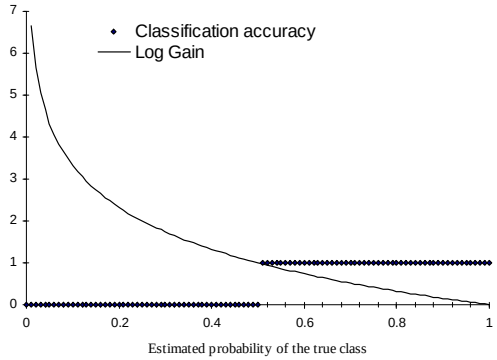


Figure 1 Log Gain and classification accuracy vs. the probability of the true class

Figure 1 compares Log Gain with classification accuracy. Specifically, For a binary classification problem, Figure 1 shows classification accuracy and Log Gain as a function of the model's estimated probability of the *true* class, assuming binary, maximum a posteriori classification. As shown, Log Gain monotonically decreases as the probability of the true class increases. Furthermore, the steepest changes in Log Gain occur when the initial estimate of the true class is *least* correct Thus, when the average probability estimation leads to incorrect classification, Log Gain assigns higher weights to acquisitions that improve the model's probability estimation toward a *correct* classification. By contrast, the measure of classification accuracy is very coarse (a step function)—it does not capture changes in the estimated class probability except when the model's estimated most-probable class changes.

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, $\mathrm{var}\,\hat{f}$, as long as the true class probability f and the average probability estimation $E\hat{f}$ lead to the same (correct) classification, then the likelihood of classification error is reduced the more extreme the average probability estimation $E\hat{f}$ is. This is because it is less probable for the estimation variance to cause an erroneous classification. Now, let us examine how Log Gain and classification accuracy capture these changes in the model's estimation following an acquisition. As shown in Figure 1 when the probability estimation already leads to a correct classification (i.e., $\hat{f}_{C,F}(y|x) > 0.5$) the difference in Log Gain before and after an acquisition is largest when the model is least confident in its prediction (i.e., the probability estimate is close to 0.5 initially). Thus, using Log Gain as a measure of utility also promotes acquisitions which are likely to decrease the risk of incorrect classification due to estimation variance when the average probability estimation already leads to correct prediction—and for which classification accuracy would remain unchanged.

In summary, Log Gain captures important changes that will allow the AFA policy promotes acquisitions which decrease the likelihood of classification error: It assigns higher weights to acquisitions that increase

the average probability estimation of the true class when, on average, the model's class prediction is incorrect, and it promotes 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 which promote these two objectives would benefit the AFA policy as well. In Section 5.4 we evaluate empirically the advantages conferred to the AFA policy when the value to induction is measured by Log Gain as compared to classification accuracy.

4.2. Estimating feature-value distribution

We now address the second term in equation 1. There are many ways to model the distribution of an unknown feature-value $F_{i,j}$. We would like to estimate the conditional probability distribution of a missing feature value given the known values. Specifically, for each feature j, the probability $P(F_{i,j} = v_k)$ in Eq. 1 is inferred from a model $M_{i,j}$, which maps feature values of instance *i* onto the distribution of the feature $F_{i,j}$.

The primary challenge to be addressed in this estimation stems from the *unknown* values in the data. Because of the AFA setting, the instances to which the model $M_{i,j}$ would be applied are likely to have one or many missing values. Thus predictors whose values are required for *inference* may not be available. While these missing values can be imputed with an estimation during inference, an imperfect imputation may lead to a different prediction than that inferred if the missing value(s) were known. Prior research has shown that popular treatments for estimating a test instance's missing predictors or their distribution to allow for inference, result in substantially inferior predictions than those obtained with models that employ only the *known* predictors of the corresponding instance (Saar-Tsechansky and Provost 2007). In particular, a model that incorporates a predictor that will be missing at inference time at best adds an irrelevant variable, increasing prediction variance; a model that includes an *important* variable that would be missing at prediction time is significantly worse unless the value can be imputed with a highly accurate estimate. This is because incorporating an important predictor for model induction would tend to reduce the effectiveness at capturing predictive patterns involving other predictors that will be available at inference time and that can subsequently minimize the loss.

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. However, different instances may include different sets of known feature values. To predict the value distribution of feature j for different instances with an arbitrary modeling techniques, it would be necessary to induce a different model for each encountered set of known feature. This is clearly a very computationally intensive task. For the empirical evaluations in this paper we employ one predictor, the class variable, to estimate the distribution of a missing value, because it 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 5.4 we validate empirically the benefits of relying on known predictors exclusively. Our results are consistent with prior research—a straightforward application of more complex modeling that relies on the interactions among predictors and which require the estimation of unknown predictors for inference does not benefit AFA performance. In addition, a simpler model that does not condition the missing feature distribution performs poorly because it does not take into account any instance-specific information in the estimation.

4.3. 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 computatinally fast approach to reduce the set considered to a sub-sample of all the possible queries. We consider two approaches, both of which are based on an exploration parameter α ($1 \leq \alpha \leq \frac{mn}{b}$) which controls the complexity of the search. (Recall that b is the batch size, m the number of examples, and n the number of features.) To select a batch of b queries, first a sub-sample of αb queries is selected from the available pool, and then the expected utility of each query in this sub-sample is evaluated. 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.

The first approach, *Uniform Sampling (US)*, identifies a representative subset of missing feature-values via a uniform random sample of queries. The expected utility from each feature-value in the set is estimated and the feature values with the highest utility values are acquired.

An alternative approach is to try to limit the consideration set to a subset 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". This notion of uncertainty 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, the lack of discriminative patterns results in uncertain predictions where the model assigns similar likelihoods for class membership in 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 *Error Sampling* and *Uniform Sampling* are used to reduce the set of instances whose missing values are considered for acquisition *Sampled Expected Utility-ES* (SEU-ES) and *Sampled Expected Utility-US* (SEU-US), respectively.

5. Experimental Evaluation

We now present a comprehensive set of experiments demonstrating the efficacy of AFA. We first show that features can be acquired cost-effectively for modeling, using a set of web-usage data sets and common benchmark data sets. We then report studies on the measures we employ as well as on sensitivity analyses to the design choices (Langley 2000, Hevner et al. 2004).

5.1. Data Sets

We begin by evaluating the proposed policies on a set of data sets from a variety of domains. Four data sets,[3] *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 determine 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. In the initial experiments we simply assume for all data sets all feature costs are equal. In Section 5.6, we experiment with different cost structures.

5.2. Methodology

In the empirical evaluations that follow the performance of each acquisition policy is evaluated over 10 random replications of 10-fold cross-validation as follows. In each replication, the data set was randomly partitioned into 10 folds. In each iteration of cross-validation, all policies were provided with the same subset of initial feature-values, drawn uniformly at random from the training partition of the data. 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

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

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 that were harder to learn, 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 error of A over B at each acquisition phase and report the average over all acquisition phases. We refer to this average as the *average percentage error reduction*. 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 across all the acquisition phases according to a paired t-test ($p < 0.05$). At the beginning of the learning curve, when a large number of the feature-values are missing, useful acquisitions often have a significant impact on learning. To capture this, we also report the average percentage error reduction over the 20% of points on the learning curve with the largest improvements (Saar-Tsechansky and Provost 2004). We refer to this as the *top-20% average error reduction*.[4] The learning curves that we present below show average performance of each policy over 10 experiments of 10-fold cross-validation.

5.3. Results

We begin with an evaluation of the classification performance as a function of acquisition cost, obtained by the policies SEU-ES, SEU-US, and a policy in which a representative set of feature-values acquisitions are drawn uniformly at random. Summary results of the experimental comparison along with the beta values used in the experiments throughout the paper and the size of the initial sample are presented in Table 1. In this and subsequent tables each significant value is marked with an asterisk (*). Figure 2 presents results on four data sets that exhibit the different patterns of performance we observe. As shown in Table 1, for all

[4] To interpret the top-20% average error reduction, it is important to examine the learning curves to ensure that one curve is indeed consistently above the other, which is the case for these experiments, in order to ensure that the measure is not simply favoring the higher-variance curve.

Data Set			SEU-US		SEU-ES	
	Beta	Initial Sample (Instances)	Average Percent Error Reduction	Top-20% Average Error Reduction	Average Percent Error Reduction	Top-20% Average Error Reduction
audiology	100	147	14.61*	19.04*	19.72*	24.50*
car	50	1033	10.93*	19.01*	11.11*	17.1*
eToys	100	125	19.11*	27.95*	49.18*	61.31*
expedia	100	350	10.04*	15.78*	16.61*	21.66*
lymph	20	38	7.02*	10.71*	3.05*	7.67*
priceline	100	75	10.69*	17.36*	1.24†	11.07*
qvc	100	225	3.76*	7.80*	14.82*	22.20*
vote	10	59	18.3*	29.42*	8.33*	22.47*

*$p<0.05$ † $p<0.06$

Table 1 Error reductions obtained with SEU-US and SEU-ES as compared to uniform query sampling

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 of 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 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 sample 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*, *Error Sampling* draws a superior subset of potential acquisitions than those drawn on average via uniform query sampling. As a result SEU-ES performs significantly more informative acquisitions than SEU-US. In other data sets, e.g. *priceline*, SEU-US is preferable. Recall, that *Error Sampling* assigns a higher preference to instances that exhibit patterns inconsistent with those captured by the current model. *Error Sampling* ranks entire *instances* and does not evaluate individual feature-values. As such, it may rank instance A above B, based on the fact that acquiring *all* the missing values for A will result in a better model. However, an

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

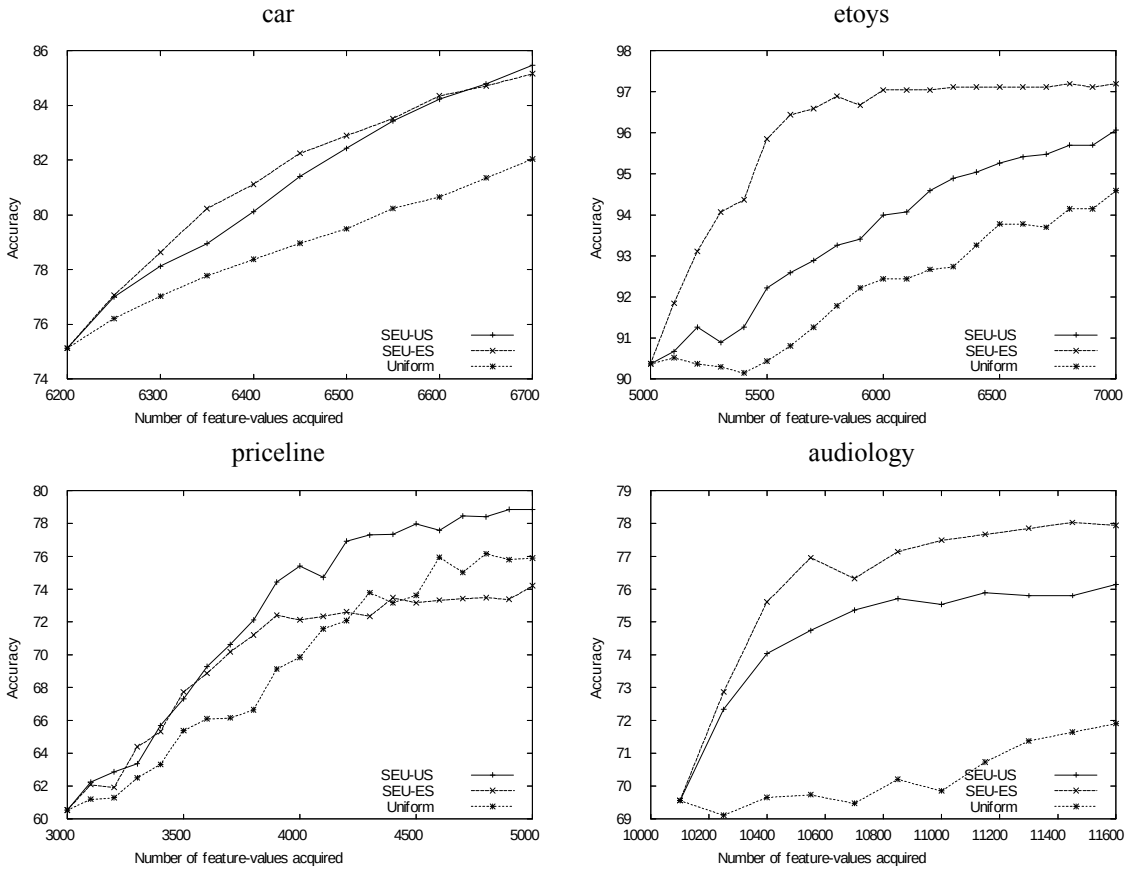


Figure 2 Classification performance as a function of the number feature-values acquired, assuming uniform feature costs.

individual feature-value for instance B may still contribute more to learning than each of the individual missing feature-values for instance A.

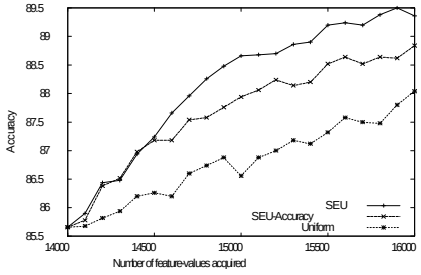
The average error reduction obtained with SEU-US over all acquisition phases ranges between 3.76% and 19.11%. The top-20% average error reduction obtained with SEU-US ranges between 7.81% and 29.42%. SEU-ES often results in even more substantial savings, but its performance is more varied than that of SEU-US. The average error reduction obtained by SEU-ES ranges between 1.24% and 49.18%, and its top-20% error reduction can be as high as 61.31%. As we note above, Error-Sampling may sometimes fail to select instances with a highly informative feature-value if the entire instance is 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.

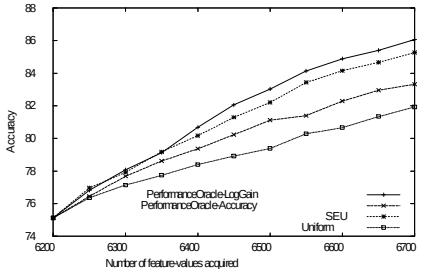
	SEU vs. SEU-Accuracy		Performance Oracle-LogGain vs. Performance Oracle-Accuracy	
Data Set	Avg Err Red	Top-20% Err Red	Avg Err Red	Top-20% Err Red
audiology	3.45*	7.46*	2.1*	5.68*
car	6.81*	13.73*	8.34*	15.57*
etoys	10.18*	21.01*	0.7	12.75*
expedia	3.94*	6.54*	3.45*	9.89*
lymph	1.3	5.57*	6.83*	11.24*
priceline	6.39*	12.63*	2.45*	4.8*
qvc	3.82*	6.91*	2.94*	6.72*
vote	9.73*	19.93*	6.52*	18.51*

$^*p < 0.05$

(a)



(b)



(c)

Figure 3 SEU with classification accuracy as a measure of value to induction

5.4. Empirical Evaluation of the AFA Measures

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 ($P(F_{i,j} = v_k)$). As we discussed in Section 4.1, these estimations are quite challenging because of the missing values in the data and because of the need to capture changes following a single value acquisition. In this section, we examine the effect on performance of the measures we propose and compare them with several alternatives.

In Section 4.1 we showed analytically that *Log Gain* captures changes in the estimated class probabilities which reduce the likelihood of erroneous classification. By contrast, classification accuracy only captures changes in the probability estimation when the estimated class membership also changes. To evaluate this effect on SEU performance empirically we first compare the SEU policy performance when Log Gain and classification accuracy estimated over the training set are used to measure the value of prospective acquisitions. We refer to the latter SEU policy as *SEU-Accuracy*. Figure 3(a) presents the average and top-20% average error reduction obtained when using SEU and SEU-Accuracy. Figure 3(b) shows the performance of each policy as well as of uniform sampling for the *expedia* data set.

As shown in Figure 3(a) for all the data sets SEU results in statistically significantly higher or comparable classification accuracy for a given acquisition cost than that obtained with SEU-Accuracy. The improvements obtained with Log Gain can be substantial, ranging from 5.57% to 21.01% on the top-20% average

error reduction metric. By capturing changes in the probability estimation Log Gain is able to detect significantly more informative feature values that lead to better models on average.

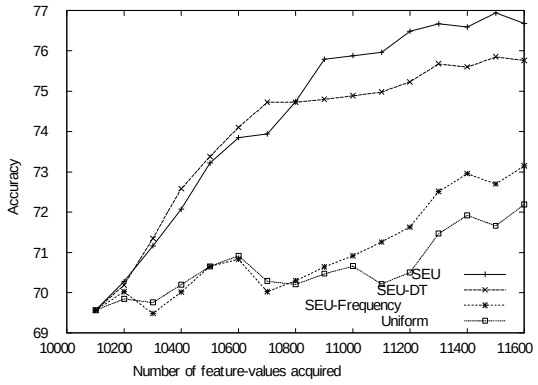
A possible reason for the inferior performance obtained by SEU-Accuracy may be the difficulty of precisely estimating classification accuracy using only the training data: differences between *training* and *test* data often result in differences in performances measured on each set. While in practice only the training data are available to the SEU policy, it is important 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 evaluated the SEU policy's performance when 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-Log Gain and Performance Oracle-Accuracy, respectively. Figure 3(a) presents the error reduction obtained with Performance Oracle-Log Gain 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 *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. Other measures that can detect the changes in the model's class probability estimation are also likely to be effective for the SEU policy.

Now, recall from Eq. 1 and the associated discussion that in order to estimate expected utility, we not only need to estimate the utility from various values that might result from a potential acquisition, we also must estimate the probability distribution over those values. As we discuss in Section 4.2, prior research has concluded that employing only predictors whose values are known during inference improves prediction significantly. Based on these findings, in the experiments so far, 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* distribution 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 decision tree, 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 4.2, 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 rely on unknown values or estimate a simple unconditional distribution, this approach conditions the estimation on the known label. Consistent with our

Data Set	SEU vs. SEU-Frequency		SEU vs. SEU-DT	
	Average	Top-20%	Average	Top-20%
audiology	11.6*	17.25*	3.45*	7.86*
car	7.94*	14.87*	-0.16	2.65
eToys	18.61*	28.24*	3.48	9.31*
expedia	10.38*	14.89*	-1.06	2.29
lymph	7.12*	10.66*	-1.2	2.89
priceline	10.55*	17.32*	0.96	5.83*
qvc	3.68	7.15*	3.48	11.62*
vote	15.84*	24.45*	0.02	11.32*

*$p<0.05$

(a)



(b)

Figure 4 SEU with different methods to estimate a missing feature value's distribution

approach are other methods that may capture arbitrarily complex predictive patterns, but which incorporate only predictors whose values will be known at inference time.

Figure 4(a) presents reductions in error, when using conditional distributions (SEU) as compared to using unconditional frequency estimation (SEU-Frequency) and decision trees (SEU-DT). For the *audio* data set, Figure 4(b) shows the performance of SEU with each estimation approach. The results suggest that unconditional distributions provide poor estimates of the feature-value distribution as compared to the conditional distributions. Specifically, the top-20% average error reduction obtained with SEU employing conditional distributions as compared to SEU-Frequency, ranges from 7.15% to 28.24%. The poor estimates produced by unconditional distributions lead to acquisitions that are only marginally better and often comparable to those obtained with uniform query sampling. We also find that SEU is often better or comparable to SEU-DT which captures more complex patterns but relies on predictors that may be unknown at inference time. As was shown in prior work, errors in estimating missing predictors or their distributions cumulatively contribute to prediction error. Furthermore, because the induction technique implicitly assumes all predictors will be available during inference, it does not capture alternative predictive patterns involving feature-values that will be available during inference.

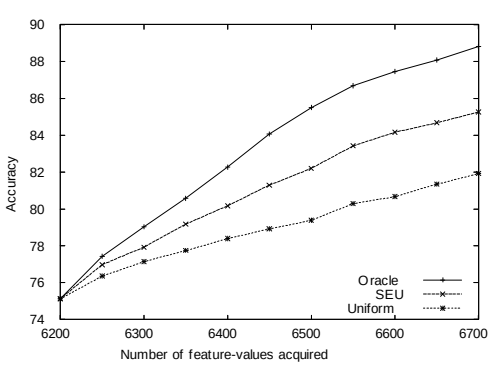
5.5. Upper bound Performance: A Comparison with an Oracle

We now explore SEU's performance as compared to an upper bound. To derive an upper bound for SEU performance, we assume an oracle knows the true values of missing features to determine which acquisitions of feature-values 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 in SEU, to render the evaluation feasible, rather than evaluate all possible acquisitions the oracle selects the best acquisition among a sample of αb prospective acquisitions. Both policies select prospective acquisitions from the same set of prospective acquisitions.

Data Set	Oracle vs. SEU		
	Average Percent Error Reduction	Top-20% Average Error Reduction	SEU Error Reduction as % of Oracle's
audiology	9.47*	15.92*	49.21
car	13.11*	23.09*	53.17
eToys	3.11*	11.97*	40.51
expedia	4.09*	11.9*	42.89
lymph	14.9*	23.39*	36.38
priceline	0.63	4.28*	56.54
qvc	2.42*	6.03*	47.31
vote	0.66	6.59*	62.43

*p<0.05

(a)



(b)

Figure 5 Oracle performance as compared to 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 to capture what proportion of the optimal improvement gained by the oracle is 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.

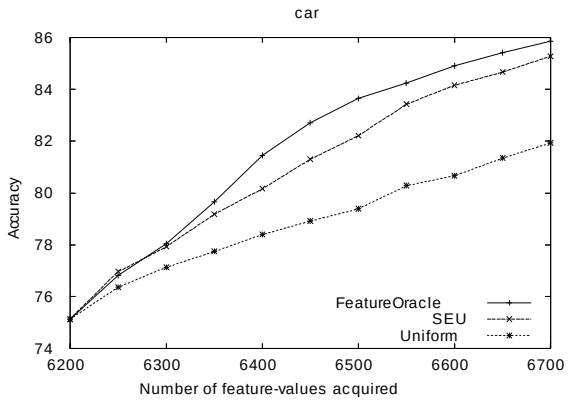
Let us now decompose the advantages conferred by the oracle over the imperfect estimations performed by SEU. Specifically, we decompose the relative advantages into the oracle's knowledge of the true model's performance measured over the held-out test data as compared to SEU's estimation over the training 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. These insights can suggest how improvements in each of SEU's estimations can contribute to SEU's overall performance so as to approximate the upper bound performance of the oracle.

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 each missing value 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, *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.

Figure 6(a) presents the average error reduction and top-20% error reduction obtained with the Feature

Data Set	FeatureOracle vs. SEU	
	Average Percent Error Reduction	Top-20% Average Error Reduction
audiology	-1.93	-0.31
car	3.87*	7.78*
eToys	-2.6	2.95
expedia	4*	8.43*
lymph	2.19*	5.63*
priceline	-2.24	0.26
qvc	-2.08	0.42
vote	0.64	8.35*

*$p<0.05$



(b)

Figure 6 A feature-oracle performance as compared to SEU

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 oracle's acquisitions lead to models that are between 2.19% and 4% more accurate than those produced by SEU, on average. Because SEU computes the expected value from prospective acquisitions, a more accurate estimation of each missing feature-value distribution can improve the policy's performance. Our results above show that, in principle, there is some room for improvements in feature value distribution estimation. As we show in Section 5.4, 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. More computationally intensive approaches than the one we employ, may be considered as well. For example, it is possible to induce a different model to estimate the distribution of each missing value, using the complete set of known predictors in the corresponding instance.

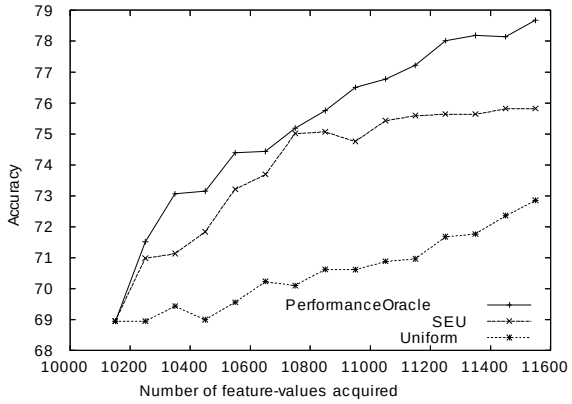
To evaluate the benefits from assessing the model's accuracy directly over the test data we compare SEU's performance to that of a *Performance Oracle*—a policy in which the improvement in Log Gain is computed on the test data. We fix all other components of the polices so that the Feature Oracle and SEU employ the same measure to estimate feature-value distribution and evaluate the same set of prospective acquisitions at each acquisition phase. Figure 7(a) presents the average percent error reduction obtained with Performance Oracle as compared to SEU. Figure 7(b) shows the performance of models induced with Performance Oracle, SEU, and the Uniform policy for the *audiology* data set. With the exception of a single data set, the *Performance Oracle*'s ability to rank potential acquisitions by their impact on predictions over the test data results in models that are substantially and statistically significantly better than those induced

SEU with Performance Oracle vs. SEU

Data Set	Average Percent Error Reduction	Top-20% Average Error Reduction
audiology	5.63*	10.66*
car	2.74*	5.07*
etoys	6.29*	18.41*
expedia	3.56*	11.05*
lymph	2.17*	5.77*
priceline	2.64*	5.91*
qvc	-0.18	3.15*
vote	5.05*	12.5*

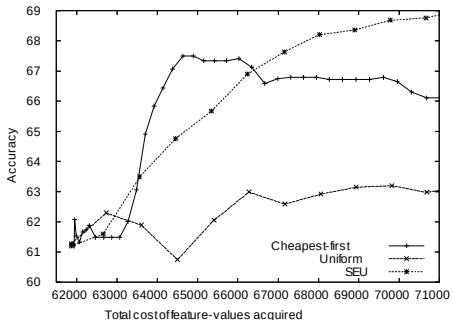
*$p<0.05$

(a)

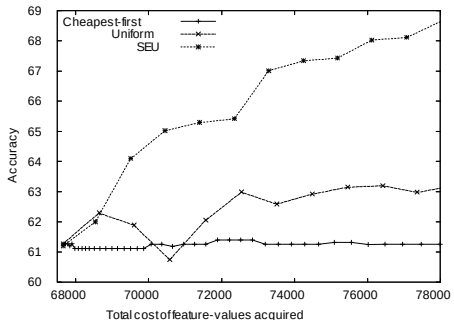


(b)

Figure 7 SEU as compared to Predormance Oracle



(a) Feature cost structure 1



(b) Feature cost structure 2

Figure 8 Comparing different policies on artificial data under different cost structures

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 the most significant advantage to its performance over SEU. However, in practice, this information cannot be available to an acquisition policy. The 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, in principle, it may be possible to improve it so as to exhibit a performance comparable to that of the Feature Oracle.

5.6. Non-uniform feature costs

For some of the data sets we employ (e.g., the UCI data sets) features already have been preprocessed carefully for relevance. However, in practice data often have many irrelevant attributes. In these situations, the power of SEU would be particularly useful in order to avoid investment in non-informative acquisitions. In this section we explore SEU's performance under such a setting. 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 different assignments of feature costs.

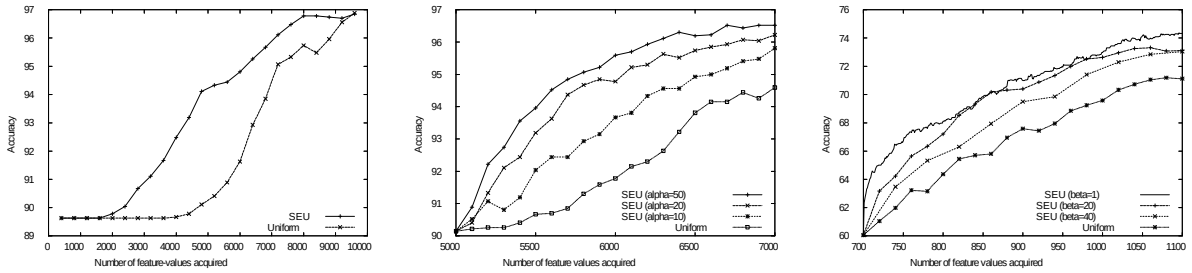
Since uniform sampling does not take feature costs into account, we also compare SEU with a baseline strategy that does. This approach, *Cheapest-first*, selects feature values for acquisition in order of increasing cost. The results for all randomly assigned cost structures show that for the same cost, 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 feature-values per unit cost is more critical when there are features of varying information values and costs.

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., the left half of Figure 8a), 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.

5.7. Sensitivity Analyses

The SEU policy incorporates several parameters that affect its performance. These parameters include the sample size of feature values whose contributions to learning are evaluated by the SEU policy, the number of feature values acquired in each acquisition phase, and the size of the initial sample provided to the SEU policy for evaluating the contributions of prospective acquisitions. We now explore in turn how each of these parameters affects SEU's 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 (i.e., for each data set, the number of initial features values is equal to that of 10 instances of the corresponding 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



(a) Exhaustive acquisition of values (b) SEU with different alpha values (c) SEU with different beta values

Figure 9 Sensitivity analyses

both policies require additional data to produce predictive patterns and improve performance. However, the acquisitions made by the SEU policy are significantly more informative and thus SEU acquires fewer features to achieve a given level of performance as compared to the uniform sampling policy. Thus, even when only a very small number of values are available initially, it is preferable to employ SEU. In addition, as 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 alpha, is likely to affect the SEU policy's performance. 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 alpha 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 is clearly 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 performance with batch acquisitions is 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.

6. Active Information Acquisition

In this section, we aim to further demonstrate 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 induce a customer's response model 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 efficient acquisitions to result in a desired performance.

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 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 5.4. We refer to this policy as Sampled Expected Utility-AIA (SEU-AIA).

6.1. Determining the Consideration Set for AFA

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 missing class labels than feature values and a uniform sample would tend to reflect this. Such a bias could be quite 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 the information conveyed by this value for model induction (described next), normalized by the value's cost. These weights then guide the sampling of the consideration set.

We evaluate the contribution to induction of all feature-values of a given a *feature* F_i by the *information gain* $IG(F_i, L)$ of the feature F_i and class variable L, 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 are also more valuable to induction. All missing feature values of a given feature are assigned the same weight: 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.

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 feature 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.

6.2. SEU-AIA Performance

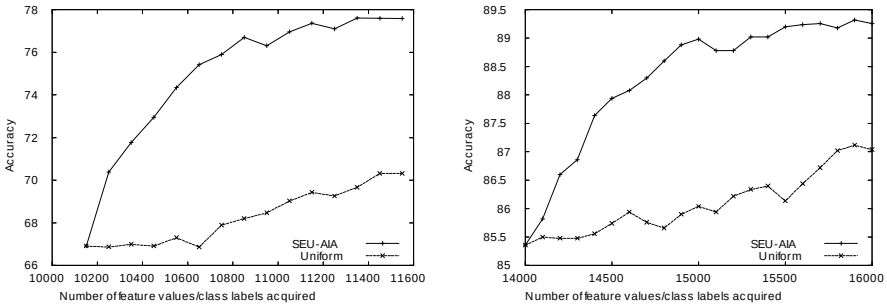
To assess the performance of SEU-AIA for this general information acquisition task, we remove class labels and feature values from the training data uniformly at random. We compare the performance of SEU-AIA to acquiring missing values uniformly at random. For this set of experiments we assume that all 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 achieved by using SEU-AIA over uniform sampling. SEU-AIA consistently acquires informative values for modeling that 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.

Data Set	SEU-AIA vs. Uniform	
	Average	Top-20%
audiology	21.29*	26.28*
car	0.52*	2.58*
eToys	39.71*	50.40*
expedia	15.97*	21.23*
lymph	16.42*	21.35*
priceline	28.54*	40.79*
qvc	23.47*	29.17*
vote	20.05*	28.25*

*p<0.05

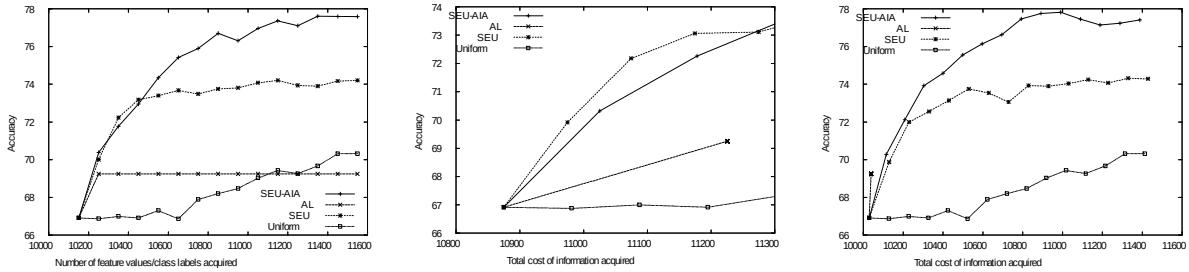
(a)



Audiology Expedia

(b)

Figure 10 Accuracy obtained with Active Information Acquisition and with Uniform Random Sampling



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

Figure 11 A comparison of different acquisition policies under different cost structures

To gain further insight into the SEU-AIA policy's choice of acquisitions we analyze experimentally the performance of SEU-AIA to that of an active learning (AL) policy, which 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 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 EU-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.

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) not one type of information is consistently more cost-effective than the other and it is therefore 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 best policy, 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 structures in Figure 11(a) and (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 preferable, allowing one to consistently identify different types of cost-effective acquisitions.

7. 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.

- The current formulation does not address "lookahead" in the selection of examples. Perhaps being less myopic will allow better estimations of the relevant quantities, and therefore better selection of the examples that will lead to the steepest learning curve. Further, if batches of examples are going to be selected anyway, it may be possible to optimize their composition.

- 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's 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—queries can also correspond to arbitrary feature-value sets.

Clearly, estimating the value distribution of all possible assignments for a given set and computing the value to induction from each assignment would be 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.

- In this paper we condition the distribution of a prospective acquisition on the class variable, because it is guaranteed to be known at inference time. However, in principle, one can model this distribution using any set of *known* predictors. As we note earlier, this approach can be hopelessly 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. Using Bayes' theorem the conditional probability of the dependent variable C, given a set of *known* (observed) independent variables $v_1, v_2, ..., v_n$, is proportional to the joint probability $P(C, v_1, v_2, ..., v_n) = P(C)P(v_1, v_2, ..., v_n \mid C)$. While this is highly costly to compute, given the "naïve" assumption that all independent variables are independent of each other given the dependent variable's value, the joint probability can be written as $P(C) \cdot \prod_{j=1}^{n} P(v_j \mid C)$. The naïve assumption is attractive for the AFA setting, because it allows inference without the need to estimate the values or the distributions of missing predictors. In addition, rather than induce a different model for each possible set of missing predictors, it is trivial to marginalize any missing variables by not including them during inference. The benefits of this approach remain to be validated empirically; the Naïve Bayes tends to produce biased (extreme) probability estimates and as these probabilities are critical to SEU's expectation calculation, this drawback may outweigh the benefits.
- Our policies employ a "wrapper" approach to identify informative acquisitions for a given induction procedure including its missing value treatment of choice. Clearly, the quality of the models induced can benefit from a missing value treatment that performs well under a potential bias in the missingness pattern created by a selective (non-uniform) acquisition policy. In the empirical evaluations in this paper we employ the C4.5 induction algorithm, and it would be useful for future work to explore the performance with other missing value treatments or modeling techniques.
- The *Expected Utility* framework allows one to incorporate performance objectives other than accuracy. For example, when the benefits from making different accurate predictions and the costs of different errors are specified, *Expected Utility* can be, in principle, applied to identify acquisitions that improve the expected *benefit* from model use per unit of information acquisition cost.

8. Conclusions

We have proposed 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 introduce a measure for the

utilities of different possible feature-value query results that captures the benefit from acquisitions 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 scarce data. We show 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, is shown experimentally to consistently yield better models per unit acquisition cost, when compared to uniform query sampling. This result holds for uniform and non-uniform feature acquisition costs. The technique's advantage is particularly apparent when feature values have varying information value and incur different costs.

We also study SEU's component measures as compared to alternatives and to omniscient oracles, that know exactly the quantities being estimated. These studies reveal that SEU's measure of prospective feature value distributions is very effective, and that the greatest improvement in performance by the oracles is obtained when they have access to the "test" instances to compute the exact impact of different values that may be acquired. A sensitivity analysis of the method's parameters produces intuitive results—SEU performs best when the consideration set of prospective acquisitions whose expected utilities are evaluated is large and when the number of values acquired at each phase is small.

Finally, we show 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 only acquire feature values or only class labels.

As we discuss 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, allowing them to enhance their limited data on their customers with particularly informative values.

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

The paper was improved by substantive comments by the editor and the anonymous reviewers. We thank Duy Vu, Pryank Mishra, Hasrat Godil, Prateek Gupta, and Saurabh Baji for their help in software development for this research.

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.

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*.

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.