Data acquisition and cost-effective predictive modeling: targeting offers for electronic commerce

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

Prem Melville
IBM T.J. Watson Research Center
pmelvil@us.ibm.com

Maytal Saar-Tsechansky
Red McCombs School of Business
University of Texas at Austin
maytal@mail.utexas.edu

ABSTRACT

Electronic commerce is revolutionizing the way we think about data modeling, by making it possible to integrate the processes of (costly) data acquisition and model induction. The opportunity for improving modeling through costly data acquisition presents itself for a diverse set of electronic commerce modeling tasks, from personalization to customer lifetime value modeling; we illustrate with the running example of choosing offers to display to web-site visitors, which captures important aspects in a familiar setting. Considering data acquisition costs explicitly can allow the building of predictive models at significantly lower costs, and a modeler may be able to improve performance via new sources of information that previously were too expensive to consider. However, existing techniques for integrating modeling and data acquisition cannot deal with the rich environment that electronic commerce presents. We discuss several possible data acquisition settings, the challenges involved in the integration with modeling, and various research areas that may supply parts of an ultimate solution. We also present and demonstrate briefly a unified framework within which one can integrate acquisitions of different types, with any cost structure and any predictive modeling objective.

1. INTRODUCTION

Improved predictive modeling can lead to more efficient processes, higher levels of customer satisfaction, reduced operating costs, and higher returns on investment. Electronic commerce has been viewed as an ideal domain of application for data modeling technology [2, 24]: Data are plentiful and relatively reliable. In principle, one can have a "closed-loop" system that can mine data, take actions, and measure results. Integration with existing processes can be much smoother than when previously manual systems must be automated.

Electronic commerce also is likely to change the face of automated predictive modeling. Having a closed-loop system allows modeling systems to begin to make decisions affecting the acquisition of the data to be modeled. For example, consider supervised modeling to help in deciding which web-site visitors should be presented with a particular offer. For supervised modeling, training data must include "labels," known values for a target variable—in this case, whether or not particular visitors responded to the offer. Targeting a consumer with unwanted solicitations may incur the cost of reduced goodwill. Possibly more important, targeting a consumer with one offer may incur the opportunity cost of not presenting a known-to-be highly profitable offer. The costs of acquiring labels for training data can be reduced using techniques from active learning [6] and optimal experimental design [22, 12]; we describe active learning techniques below.

However, existing techniques for integrating modeling and data acquisition cannot deal with the rich environment that electronic commerce presents, for at least two reasons. First, there are many different sorts of data than can be acquired. Consumers' responses to offers can be acquired through direct solicitations, such as via experimental special offers, via customer surveys, and via interactions such as Amazon's on-line acquisition of product ratings. Firms also collect information indirectly, in the course of normal business interactions, for example by observing responses to offers or the results of everyday merchandising decisions. For supervised learning one not only can acquire labels, one can acquire values for attributes (i.e., the independent variables). For our example, one may acquire credit-bureau data, or psychographic data, or prior-purchase data from a business partner. Attribute values may be obtainable individually, or in particular batches.

Data acquisition should consider various sorts of data simultaneously [50]. Some of these types of data acquisition have been addressed individually (which we discuss below), but to our knowledge there is no well-accepted data-acquisition procedure for the general problem.

The second characteristic of the e-commerce environment that is not dealt with by existing techniques is that each data acquisition has a cost associated with it. For example, different means by which consumer preferences (i.e., labels) can be acquired entail different costs. Learning from normal business transactions and experimental offers entails opportunity costs. Similarly, for acquiring consumers' preferences via surveys it is often necessary to provide consumers with costly incentives to provide genuine feedback. Attribute values can also be acquired directly from third parties. For example, Acxiom[1] sells detailed consumer demographic and lifestyle data to firms in support of their marketing efforts; other firms such as Abacus[2] maintain and sell specialized consumer purchase information.

[1]http://www.acxiom.com
[2]http://www.abacus-us.com

Furthermore, the problem to which the resultant model will be applied has associated costs and benefits. Even for the case where the only data to be acquired are training labels, traditional methods acquire data in an attempt to build the statistically most accurate model. However, accuracy maximization is not necessarily the most cost-effective policy. For example, very often predictive models are used to support profit maximization and some costly improvements in prediction accuracy do not improve the underlying objective [53]. Thus, data acquisition must take costs and benefits into account—both the costs of data acquisition, and the costs and benefits of the alternative courses of actions and their possible outcomes [59, 43].

In this paper, we present in detail the issues involved with data acquisition for predictive modeling in settings, such as electronic commerce, where various data can be acquired—at a cost—to (potentially) improve the modeling. For illustration, we use the running example of choosing offers to display to web-site visitors, which captures the issues in a generally familiar setting. The issues apply more broadly to a diverse set of electronic commerce modeling problems, from personalization to customer lifetime value modeling. Throughout the presentation, we provide the reader with a high-level guide to various research areas that may supply parts of an ultimate solution.

2. TARGETING E-COMMERCE OFFERS

Consider the following problem faced by web sites, including electronic commerce sites. What offer should a visitor encounter when she requests a page? By "offer" we mean some part of the page, separate from the content that the visitor intended to request, for which there are various alternatives. Various offers may be presented on a page. For example, when a user visits Amazon.com, many offers are presented including internal advertisements, special discounts, and links to other areas of Amazon's web site. For clarity, we will ignore combinatorics and discuss the selection of a single offer (e.g., which should go in the upper right-hand offer area?).

The selection of "the best" offer involves a conditional modeling problem: given some representation of the customer (that may include demographics, prior behavior and the customer's activities during the current session), and the available offers, estimate the expected value of the offer. This, in turn, involves estimating (conditional) probabilities and values associated with the various possible offers. These must be estimated before making the offer; however, even retrospectively the computation of the value of an offer is not always straightforward. In order to estimate conditional probabilities and values, an offerer faces the time-honored problem of determining what exactly to condition on for any particular offer.

2.1 Estimating the value of an offer

How to compute value in different situations differs based on at least three factors. First, there are fundamentally different sorts of offers. An advertisement for a third party brings a different value structure than an enticement to visit a different area of the same e-commerce site. Second, different values are derived by different pricing models. For example, sponsored advertisements based on keyword bidding will take the bids into account in determining expected value.[3] When ranking sponsored advertisements for display along with organic search results, a procedure that is now standard [46] is to combine bid amounts with estimated probabilities of clickthrough to compute the expected revenue from each potential sponsored advertisement. This amounts to the immediate value expected to be generated from an ad:

$$E_{ad}[revenue] = p_{click}(ad) \cdot CPC_{ad} \quad (1)$$

where p_{click} is the probability of a clickthrough, and CPC_{ad} is the cost-per-click the advertiser agreed to pay.

Third, different ultimate notions of outcome value may be used, perhaps due to what exactly is measurable [32, 44]. Viewing success as a user clicking on the offer is perhaps naive, but it is easy to measure. A less naive strategy would be to look for a conversion "resulting" from clicking on the offer. This allows for a clearer notion of value in certain cases, but also may miss difficult-to-track conversions (such as subsequent off-line purchases) and aspects of value such as brand awareness[4] [21]. The notion of brand awareness reminds us that even in terms of measurable profit, the value of an acquired customer extends far beyond the immediate transaction. A firm may want to consider the effect of an offer on the lifetime value of a customer [48]. And looking even further, a firm should want not just to maximize the lifetime value of the customers independently, but the lifetime value of the population of (potential) customers. For example, from the data mining perspective, prior work [11, 17] provides views into the value of considering potential customers as a social network.

Of course, difficulty in measuring the offerer's true goals may not be sufficient reason to ignore them, and some firms try to incorporate notions of longer-term value into their explicit value calculations. For example, recently, Google began additionally to include (in the ranking function) estimates of the quality/relevance of the target page.[5] This could be viewed as an attempt to incorporate in its value estimate longer-term affects (such as searchers becoming disillusioned with sponsored ads if they find irrelevant results). In principle, to best select among alternative offers/ads it is necessary to estimate the expected value of each including long-term effects on consumers and offerers. Specifically, an ad placement decision may take into account possible impact of the decision on the customer's future behavior as discussed above, as well as the offerer's future bidding behavior (e.g., frequency, CPC, etc.), the impact on future behaviors of other offerers who are competing for the space, and the behaviors of other customers.

2.2 On what attributes to condition?

In order to estimate the expected value of the possible offers, an offerer must determine: on which, if any, attributes will it condition the estimates. For example, the probability of a visitor clicking on an offer for discounted cookware might be conditioned on the number of cookbooks the visitor previously purchased. Many different sorts of attributes might be considered for conditioning.

2.2.1 Atomic offers and undifferentiated visitors

In the simplest scenario, the offerer considers no specific information about the offer, the visitor, or the offering context. Clickthrough probability and value for each offer can be estimated unconditionally from all prior visitors to which the offer was made. To our knowledge, this is one of the two most frequently applied

[3] http://adwords.google.com

[4] Some on-line advertising pricing schemes use cost-per-impression, charging based on how often visitors are shown an ad [46].

[5] http://adwords.blogspot.com/2007/02/quality-score-updates.html

scenarios for offer modeling. Although this simple scenario provides no tailoring to specific users, it brings the advantage of a large amount data for certain offers and keywords. On the other hand, as discussed by Richardson et al. [46], for newer offers there may be little or no data at all on which to base estimates.

2.2.2 Visit attributes

Offers can be targeted based on attributes of the visit. For example, the other of the two most frequent scenarios for offer modeling is to condition based on the keywords most recently entered by a visitor. This is the conditioning used for most sponsored advertising accompanying search engine results. Improvements in estimation may be obtained by aggregating across somehow-similar keywords [45]. Other possible visit attributes include the content of the current page,[6] other pages visited, and more generally the clickstream of the current visit, such as purchasing an item, the removal of an item from the shopping cart [38, 40], as well as time of day, day of the week, time of year, etc. Visit attributes may help web sites to infer information need or commercial intent [39, 3, 8]. More generally, the context in which a user's activities are being made [28] has been noted as an important concept that can improve the predictability of consumers' on-line behavior [38]. However there is no agreement as to what "context" constitutes, and hence how to derive such context from data.

2.2.3 Offer attributes

Attributes of the thing being offered can improve the targeting substantially, especially for newer offers. For estimating clickthrough rate (CTR), Richardson et al. [46] discuss in detail, and compare empirically, a wide variety of attributes that could be used to describe offers. They show that by training a logistic regression with offer attributes, they can decrease the error in estimating CTR by 30-40%. Importantly for this paper, many of the potential attributes could be costly to compute (related-term CTR, landing page quality, reputation, term category entropy, external attributes based on encyclopedia or thesaurus lookups, etc.). For example, calculating category entropy to assess the specificity of a search phrase (how "targeted" it is) improves prediction of CTR substantially. They also propose that creating an attribute based on a quick human opinion, again at a cost for each keyword/offer instance, could improve modeling substantially.

Offer attributes may be particularly useful in combination with visitor attributes [16] (described next). For a product offer, the product may belong to a product hierarchy with various levels of generalization. For example, high-tech gadgets may appeal more to visitors with a high-tech "lifestyle." Attributes of the type of offer may also affect the estimation of expected value. For example, discount offers may appeal more or less to visitors in different income ranges. Additionally, expected value modeling may take into account the proposed location of the offer on the page.

2.2.4 Visitor attributes

In principle, personalized offers can be made by conditioning on attributes of the visitors, beyond those obtainable from the current session. For example, very coarse-grained conditioning (e.g., prior customers versus anonymous visitors) can retain the advantages of ease of application and massive data. A visitor's IP address also can give relatively reliable coarse-grained information, such as country-of-residence.[7] As with traditional targeted marketing [54] there are many different types and sources of data on which to condition: prior purchase (as with traditional "recency, frequency, monetary value (RFM)" analysis), geographic, demographic, psychographic, and lifestyle attributes.

E-commerce sites may have additional data from prior visits, such as referrer search terms, clickstreams, search history, and web-surfing history, from which useful attributes can be constructed. Nasraoui et al. [40] provide a list of references to research on web-usage mining to extract frequent patterns from clickstream history, as well as a case study of their use. The most visible use of visitor-specific attributes in offer conditioning is for explicit recommendations, which usually are based on the particular products previously purchased or rated by the visitor.

Acquiring values for visitor attributes can be more-or-less costly. Referring to the example above combining visitor and offer attributes, how could the offerer know that a visitor has a high-tech lifestyle? Or a high income? Lifestyle and demographic data can be purchased from syndicated data providers such as Acxiom. Behavior on other web sites may be very useful [41], and is known by, and potentially may be purchased from, business partners.

2.2.5 Contemporaneous context

Finally, what is going on in the world beyond the electronic commerce site in principle may have an effect on estimates of response probability or conditional value. What are currently popular items or topics that might encourage clickthrough or purchase? What is the current economic climate? To our knowledge, currently such attributes are taken into account only in the creative minds of the marketing staff. Such information could be acquired at a cost.[8]

3. MODELING WITH COSTLY DATA

So, both intuitively and based on prior research results we can conclude that conditional modeling can improve offer decision making. However, this potentially useful attribute information does not all come for free. For example, in order to gather data on CTR or value of a particular offer, one must make the offer to some visitors. This introduces various costs—e.g., learning from experimental offers incurs opportunity costs. Similarly, obtaining consumer feedback (such as for capturing product preferences) requires providing consumers with significant incentives [38].

Acquiring conditioning attributes also incurs various costs. For example, category-entropy [46], which improved the estimation of CTR substantially, involves running a search engine and processing the results, which at the least incurs significant opportunity costs. The same applies to using encyclopedia and thesaurus lookups.[9] Attributes that can be purchased from partners or third parties can incur actual monetary expense. Firms may incur a cost to acquire additional information, such as psychographic, consumption, and lifestyle data from third-party suppliers. Information about different consumers may be obtained from different data suppliers who may charge different amounts for different types of information.

[6] http://adsense.google.com

[7] Google uses the visitor's IP address in targeting advertisements via AdSense.

[8] For example, consider using Amazon's "mechanical turk" to buy these data and others; http://www.mturk.com.

[9] Richardson et al. [46] note that "more advanced techniques have been proposed that would have been infeasible to do for every ad in our data set."

Furthermore, in principle, firms may also find it useful to purchase information about a visitor's activities at other sites [61].

Although we have been discussing offer-making in order to provide an in-depth look at a specific problem, even within electronic commerce there are many model-building tasks that require similar costly information acquisition, from recommendations [29], to the selection of ad keywords [60], to the prediction of commercial intent [39, 3, 8], and beyond.

4. BASIC (COSTLY) DATA ACQUISITION

In this section we discuss different settings in which researchers and practitioners have considered costly data acquisition for predictive modeling. We briefly describe the related acquisition tasks that arise and the main techniques for guiding acquisition. In Section 5 we describe in detail a solution for the general framework that subsumes most of the settings described in this section.

For each setting we describe, we will consider the task of inducing a classification model from a set of conditioning attributes (features). For example, the task could be to predict whether an offer will receive a positive response. In this case, each offer is a data instance that can be described by a fixed set of features, such as the visit attributes, offer attributes, etc., and the class label can be *positive* or *negative* depending on the response observed. Given a dataset consisting of m n-dimensional instances, we represent it by an m-by-n data matrix X, where x_{ij} corresponds to the value of the j-th feature of the i-th instance. For simplicity, we treat the class label as the n-th feature.

4.1 Real-time (automated) experimentation

A basic form of costly data acquisition is gaining popularity in electronic commerce: running real-time, controlled experiments to gather data on the effectiveness of alternatives; Kohavi et al. [23] provide a detailed practical guide and a wealth of pointers into the literature. For estimating the effectiveness of offers, most often (to our knowledge) automated experimentation is done under the atomic-offer/undifferentiated-visitor scenario.

Once a firm is performing automated experiments, it makes economic sense to carefully manage the trade-off between exploring new offers and exploiting those that have been proved very effective in the past. The atomic-offer/undifferentiated-visitor scenario with a static set of possible offers corresponds to the classic multi-armed bandit problem [47]: given k choices (slot machines) each of which will produce stochastically some (unknown) reward (when the lever is pulled), maximize the cumulative (discounted) reward over a series of decisions (which lever to pull next). Every decision simultaneously generates a reward and generates data that improves the decision modeling.

With additional conditioning attributes, this on-line offer decision problem more generally would benefit from the application of reinforcement learning methods [57, 19]. Its importance notwithstanding, in this paper we will not consider further the on-line exploration/exploitation trade-off; the interested reader should see [42]. Instead, we will look more deeply at the problem of acquiring data for improving the performance of a model in settings where there will be explicit training and testing phases. The general ideas apply more broadly.

4.2 Active learning

The mostly broadly studied costly acquisition setting is that of traditional active learning [6], which is depicted in Figure 1(a). In these figures the gray boxes represent known information and the white boxes represent information that may be acquired at a cost. In the active-learning setting all feature-values are known or missing features can be dealt with (e.g., via imputation). For example, we may only consider building models on simple offer attributes that may be readily available for all offers, such as the length of the offer description [46]. For some, perhaps small, set of these instances we have labels collected from past user interaction. The rest of the instances are unlabeled, but they can be selected for labeling. Labeling each instance comes at a cost—this could be just the cost of acquiring a label, or could include the opportunity cost of not displaying a more profitable offer. Given the cost associated with labeling, the task of active learning is to select the best next instance to be labeled so as to build a good model at a low cost. Most prior work in active learning has focused on selecting instances to maximize classification accuracy, though some recent work has dealt with alternative objectives, such as class probability estimation [52, 36].

The active-learning setting has received a substantial amount of research attention, resulting in methods that have been successfully applied to different learning algorithms, such as neural networks [9, 6], decision tree induction [27], Hidden Markov Models [7], SVMs [58, 5] and nearest neighbor classifiers [30].[10] The challenge in active learning is to determine which unlabeled instance(s) should be selected, such that labeling it and adding it to the current data will increase the model's predictive performance the most. Most popular methods use a variant of the following general ideas:

- *Uncertainty sampling* [27, 26], which selects instances on which the current model has the greatest uncertainty in its predicted label;
- *Query-by-Committee* (QBC) [55, 13, 1, 33, 14], which selects instances on which a committee of classifiers most disagree; and
- *Estimation of error reduction* [49, 30], which selects instances, that once labeled and added to the training set, are expected to result in the lowest error on future test instances.

It has been demonstrated that active learning can significantly reduce the amount of labeled data required to build accurate models in some e-commerce related domains, such as direct marketing [53] and identifying internet ads [18].

4.3 Active feature-value acquisition

In the active learning setting we assume we have unlabeled instances, and the learner dynamically selects the instances to be labeled. Consider instead the following scenario: we are given a set of offers that have already been displayed and the corresponding responses have been recorded. We now want to build a model to predict the response on new offers. To do this, we may have some features describing the offers already made, which can be used to build a model. However, we may be able to improve the modeling by using additional features which were not available when the data was being labeled—such as demographic, pyschographic and lifestyle attributes of visitors. Such data can be purchased from

[10] See also work on optimal experimental design [22, 12].

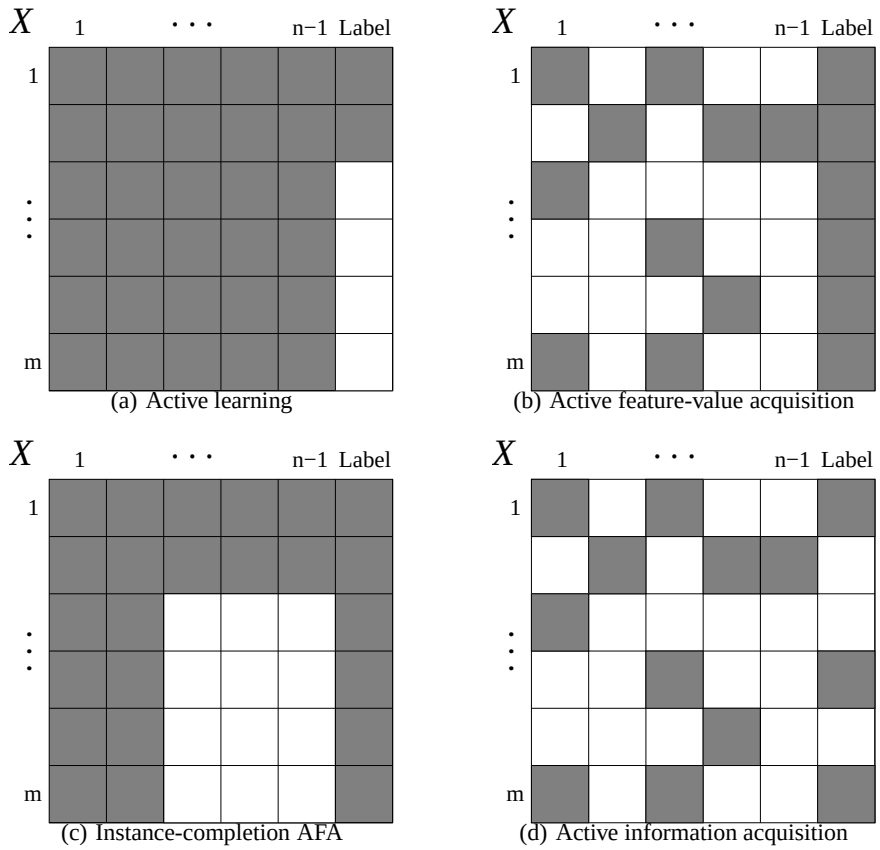

Figure 1: Different data acquisition settings. Gray boxes represent known information and white boxes represent information that may be acquired at a cost.

syndicated data providers or business partners, but they come at a varying costs. Acquiring all features for all instances may be prohibitively expensive and unnecessary, while acquiring a random subset of feature values may be sub-optimal.

Therefore, we want to select *incrementally* feature values that are most cost-effective for improving the modeling. The most general form of this active feature-value acquisition (AFA) setting is shown in Figure 1(b). In this setting, Melville et al. [35] present a procedure that ranks alternative feature-value acquisitions based on an estimation of the expected improvement in model performance per unit cost. We elaborate on this notion in more detail in Section 5. Lizotte et al. [31] study AFA under a fixed *budget*: total cost to be spent towards acquisitions is determined a priori and the acquisition procedure must identify the best *set* of acquisitions for this budget. In contrast to the incremental setting, in the budgeted setting the user is not given the option to stop the acquisition process at any time; thus the order in which acquisitions are made is not important. We will continue our discussion assuming the incremental setting.

In some situations, feature values for an instance may naturally be available in sets; e.g. a set of demographic information, such as education level and income, may be available at a single cost. This variation on the active feature-value acquisition task presents the challenge of estimating the value of sets of acquisitions, which increases the computational complexity of expected-value estimation combinatorially. Recently, there have been methods proposed to use efficient data structures to ameliorate the complexity of set-value estimation by reusing shared computations [4]. As an alternative to the computationally intense optimal solution, there have been studies that present heuristic solutions for a special case of this setting, in which one set of features is known for all instances, and the task is to select instances for which the remaining features can be acquired in a batch [61, 34]. This *instance-completion* setting is shown in Figure 1(c), and we briefly describe two approaches to AFA in this setting below.

The first approach, Dual Objective Data Acquisition (DODA) [61], assigns to each instance a heuristic score, which is an average of two measures. The first measure aims to capture the contribution of the instance to learning, and the second tries to capture the potential contribution of the instance to imputation models induced to predict the missing values. DODA does not employ incomplete instances for induction, hence the first measure is intended to capture the value of adding a complete instance to the training set. While the second measure is intended to estimate the contribution of a new instance to imputation, which is employed in the evaluation of subsequent acquisitions.

The second approach to instance-completion AFA, *Error Sampling* [34], is based on the conjecture that a set of feature values is more likely to have an impact on subsequent model induction if the acquired values belong to an instance the current model classifies

incorrectly. Such a set of feature values embed predictive patterns that are not consistent with the current model, and hence may be more informative to acquire. Motivated by this reasoning, *Error Sampling* identifies informative instances as those that the current model misclassifies. Next, it ranks correctly classified instances in order of decreasing uncertainty in the model's prediction.

The effectiveness of the active feature-value acquisition methods described in this section has been demonstrated for online customer conversion prediction, where they lead to a substantial reduction in the amount of data required to build accurate models.

5. ACTIVE INFORMATION ACQUISITION

A natural extension of the settings described in Section 4 is where both class labels and feature values may be missing and can be acquired at a cost. This is more realistic for offer targeting than the prior settings, since many conditioning attributes defining offers will be missing; we are faced with the choice of acquiring more feature values for previously labeled instances (Tell me more about that customer who responded...) or selecting incomplete instances to be labeled (Let's see whether this customer will respond...). Different attributes will have different costs of acquisition, which all will be different from the cost of labeling. In this section we briefly summarize a framework and solution proposed by Saar-Tsechansky et al. [50] for dealing with these complex trade-offs.

This setting (active information acquisition (AIA)) is shown in Figure 1(d): arbitrary elements of the data matrix X may be missing. Considering the target variable simply to be another "feature," for each missing feature x_{ij}, there is a corresponding cost C_{ij} at which it can be acquired. Let q_{ij} refer to the query for the value of x_{ij}. Then, the general task of active information acquisition is the problem of selecting the information query (instance plus value)[11] that will result in the largest increase in model quality per unit cost.

5.1 General AIA framework

The overall framework for the generalized AIA problem is presented in Algorithm 1. Since AIA is defined as an iterative task, at each step the learning algorithm is trained on the current (incomplete) dataset and ranks all possible queries based on their expected contribution to model quality normalized by cost. The highest-ranking query is then selected, and the feature value corresponding to this query is acquired. The dataset is appropriately updated, and this process is repeated until some stopping criterion is met, e.g., desirable model quality has been achieved. To reduce computational costs, multiple queries can be selected at each iteration, resulting in batch acquisitions, as often is done in active learning settings for classification tasks.

While the overall framework of Algorithm 1 is straightforward, the crux of the problem lies in ranking queries by their expected contributions to model quality. In subsequent sections, we discuss the challenges involved in performing this estimation accurately and efficiently.

5.2 Estimating expected utility

At every step of the AIA algorithm, the next best feature to acquire is the one that will result in the highest improvement in model quality per unit cost.[12] Since the true values of missing features are unknown prior to acquisition, it is necessary to estimate the potential impact of every acquisition for all possible outcomes. Hence, the optimal policy is to ask for feature values which, once incorporated into the data, will result in the highest increase in model quality in *expectation.* This *Expected Utility* approach is based on defining a *utility function* $\mathcal{U}(x_{ij} = x, C_{ij})$ which quantifies the anticipated benefit arising from obtaining a specific value x for feature x_{ij} via the corresponding query q_{ij} at cost C_{ij}. Then, the expected utility for query q_{ij}, $EU(q_{ij})$, is defined as the expectation of the utility function over the marginal distribution for the feature x_{ij}:

$$EU(q_{ij}) = \int_x \mathcal{U}(x_{ij}{=}x, C_{ij}) P(x_{ij}{=}x) \qquad (2)$$

While ranking queries using the expected utility defined above is the optimal acquisition strategy, the true marginal distribution of each missing feature value is unknown. Instead, an empirical estimate of $P(x_{ij} = x)$ in Eq. (2) can be obtained using probabilistic classifiers. For example, in the case of discrete (categorical) data, for each feature j, a naïve Bayes classifier M_j could be trained to estimate the feature's probability distribution based on on the values of other features of a given instance. Then, when evaluating the query q_{ij}, the classifier M_j is applied to the corresponding instance x_i to estimate the distribution of possible values for the missing feature $\hat{P}_j(x_{ij} = x|x_i)$, conditioned on all known feature values for the instance. Then, the expectation in Eq. (2) can be easily computed by piecewise summation over the possible values. For continuous attributes, computation of expected utility can be performed using Monte Carlo methods, or by discretizing and using probabilistic classifiers as described above.

5.3 Computing the utility function

Algorithm 1 Active Information Acquisition

Given:
X – initial (incomplete) instance-feature matrix
$\mathcal{L}$ – learning algorithm
b – size of query batch
C – cost matrix for all instance-feature pairs
Output:
$M = \mathcal{L}(X)$ – final model trained on dataset incorporating acquired values

1. Initialize set of possible queries $Q = \{q_{ij} : x_{ij}$ is not known$\}$.
2. Repeat until stopping criterion is met
3. Generate a classifier, $M = \mathcal{L}(X)$
4. $\forall q_{ij} \in Q$ compute $score(M, q_{ij}, \mathcal{L}, X)$
5. Select a subset S of b queries with the highest $score$
6. $\forall q_{ij} \in S$,
7. Acquire values for x_{ij}: $X = X \wedge x_{ij}$
8. Remove S from Q
9. Return $M = \mathcal{L}(X)$

[11] The possible (costly) information queries that can improve modeling can include additional structure, or can extended beyond the elements of this data matrix [43].

[12] We consider here a myopic notion of optimality, where only the current acquisition is considered. The framework applies more generally.

Selecting an appropriate utility function $\mathcal{U}$ to estimate the benefits of possible acquisition outcomes in Eq. (2) is a critical component of the AIA framework. The choice of utility function should be determined by the value to be optimized, for example, the marginal improvement in accuracy per unit of acquisition cost:

$$\mathcal{U}(x_{ij}=x, C_{ij}) = \frac{\mathcal{A}(X \wedge x_{i,j}=x) - \mathcal{A}(X)}{C_{i,j}} \qquad (3)$$

where $\mathcal{A}(X)$ is the accuracy of the current classifier; $\mathcal{A}(X \wedge x_{i,j} = x)$ is the accuracy of the classifier induced from X augmented with $x_{i,j} = x$; and $C_{i,j}$ is the cost of acquiring $x_{i,j}$.

Usually, maximizing simple classification accuracy is not the primary objective. In the case of sponsored search with cost-per-click pricing, we are more interested in calculating expected revenue from a potential advertisement as in Eq. (1). To do this we can build a model to predict p_{click}, the clickthrough rate (CTR), which can be cast as a class-probability estimation task. In this scenario, instead of measuring classification accuracy $\mathcal{A}(.)$ in Eq. (3), we could measure the mean squared error or the average Kullback-Leibler divergence [25] between the model's predicted CTR and the true CTR on the labeled data. In the *Expected Utility* formulation, the utility measure can be redefined to describe any other objective—e.g., revenue from alternative pricing schemes for online advertising, such as cost-per-impression or cost-per-action.

The *Expected Utility* approach therefore corresponds to selecting the query that will result in the estimated largest increase in the model utility of choice, per unit cost, in expectation. If all feature costs are equal, this corresponds to selecting the query that would result in the model with the highest expected performance. Otherwise, *Expected Utility* allows several small, high-margin acquisitions to be selected instead of one larger acquisition with less expected improvement per unit cost.

The advantage of the general AIA 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. By using this common measure one can rank acquisitions of different types—in this case, acquisitions of class labels as well as of feature values.

Another attractive feature of this approach is that it generalizes prior approaches that consider more restricted settings. When only class labels are missing, acquisition costs are uniform, and classification accuracy is the criterion of interest (i.e., traditional active learning), it is equivalent to a generalization of the method used by [49]. Their method has been shown to be effective in this setting. Analogously, in the case that class labels are given, but other feature values may be missing, this formulation is equivalent to active feature-value acquisition as in [35].

5.4 Efficiency considerations

A major challenge in implementing active information acquisition is the computational complexity of evaluating all potential acquisitions. Expected utility $EU(q_{ij})$ cannot be computed in closed form for arbitrary numeric attributes, and even for discrete distributions $P(x_{ij}=x)$ the computation could require re-training models on the dataset for every possible value of every missing feature x_{ij}. Thus, selecting the best from *all* available queries would require the computation over all outcomes for $O(mn)$ possible queries. Therefore, exhaustive selection of a query that maximizes the expected utility is computationally infeasible for datasets of even moderate size. This selection can be made tractable by limiting the search space to a subsample of the available queries.

For example [50], one can compute the information gain (IG) [37] of each of the n features (the IG of the class label being 1). Informally, the information gain gives an indication of how discriminative each feature is in terms of the class label. Each instance-feature query can then be scored based on the IG of the feature divided by its cost, and a sample taken from the top-scoring queries, and the more intensive expected utility computation applied only to this sub-sample. By adjusting the sample size one can control the trade-off between the amount of time spent and the effectiveness of the selection scheme.

5.5 Demonstration of results

Past work [50] has demonstrated the effectiveness of the *Expected Utility* approach to active information acquisition for the e-commerce task of customer conversion prediction, and we highlight that demonstration here. We consider four e-commerce data sets from a related study by Zheng and Padmanabhan [61]. These data sets contain information about web users and their visits to large retail web sites. The target (dependent) variable indicates whether or not the user made a purchase during a visit. The predictors describe visitors' surfing behaviors at the site as well as at other sites over time. Induced models estimate whether a purchase will occur during a given session and employ the *Expected Utility* approach to determine which unknown values are most cost-effective to acquire so as to improve the models' predictions.

To assess the performance of the *Expected Utility* approach to the active information acquisition task, class labels and feature values are removed from the training data uniformly at random. The performance of AIA is compared to acquiring missing values uniformly at random. Here we present the set of experiments which assume that all features and class labels have the same cost. The purpose of this comparison is to verify that *Expected Utility* 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. For results on experiments for different cost distributions see [50].

Figure 2 demonstrates the substantial impact achieved by using AIA over uniform sampling. AIA consistently acquires informative values for modeling that result in models superior to those obtained by uniform acquisition. By evaluating and comparing the different types of information effectively, AIA provides a significant lift in predictive performance. The magnitude of this impact can be seen in Table 1, which summarizes the percentage reduction in error of AIA over uniform sampling averaged over all points of the curve.

Data Set	Average Percentage Error Reduction
etoys	39.71
expedia	15.97
priceline	28.54
qvc	23.47

Table 1: Error reduction produced by using Active Information Active over Uniform Random Acquisition.

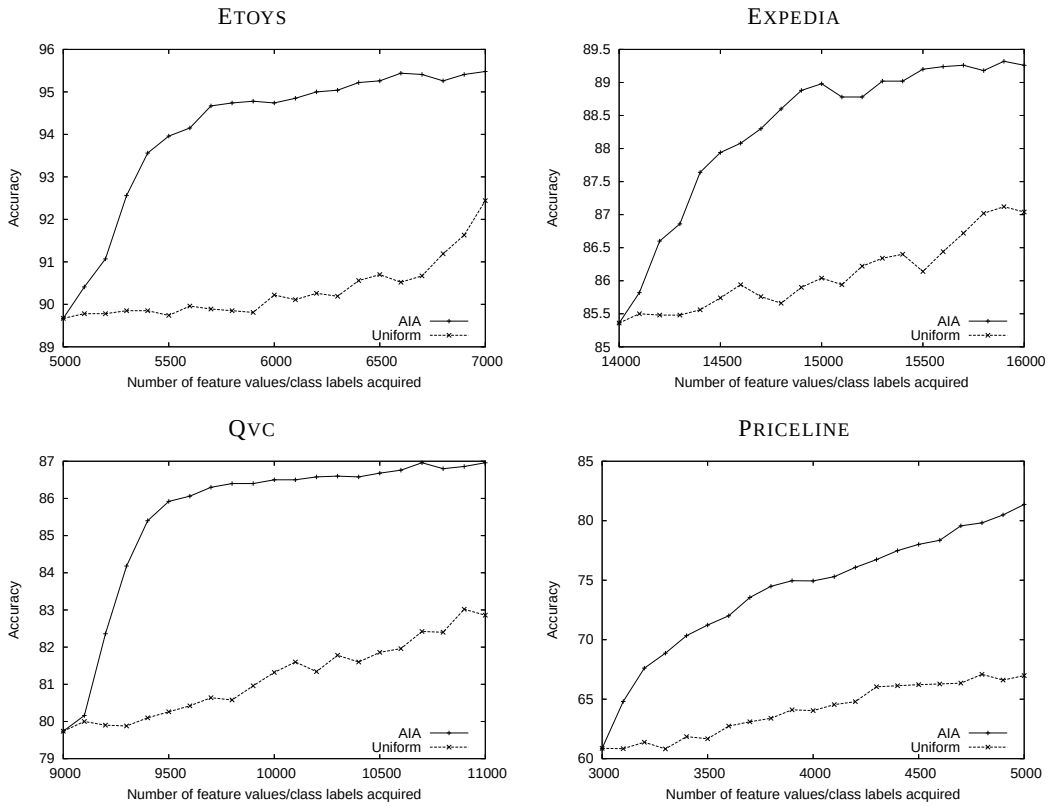

Figure 2: Comparing Active Information Acquisition vs. Uniform Random Acquisitions

6. ACQUIRING DATA WHEN USING A MODEL

In Sections 4 and 5 we have discussed different settings for data acquisition in the process of *training* a model. The focus has been acquiring data in a cost-effective manner while building an accurate model. However cost-effective data acquisition can also be a critical concern during model use, i.e., when the learned model is used for prediction on a new instance. New (test) instances may also have missing features values that can be acquired before making a prediction. For example, suppose we have built a model to predict the response to offers on a site, which is conditioned on both offer and visitor attributes. Now, if we have a returning visitor to the site we may already have all her information (demographic, psychographic, etc.), which we can use to predict her response to a new offer. However, if we have a new visitor, we may choose to make predictions solely on the offer information we have, or we could choose to acquire more visitor information to possibly make a better-informed decision. As before, this additional information usually comes at cost, and one must decide between the cost of obtaining more information and the possible loss in revenue from making an ill-informed decision. Importantly, acquisition of information must be evaluated against the best possible decision based on the available information. The "available information" includes possible estimations of the missing values (for imputation), and the use of models that do not require at all values or that employ only a subset of the features [51]—choosing between these methods for dealing with missing data can make a large difference in model performance [51].

As with model induction, there are several settings for data acquisition during model testing. Sheng and Ling [56] study the setting of feature acquisition for testing when a model has already been built. In their setting, given a test case with missing values, they attempt to determine which feature values should be acquired, and in which order, such that the sum of the feature-acquisition cost and expected misclassification cost is minimized.

Greiner et al. [15] analyze the problem of learning an optimal *active classifier*, i.e., a classifier which, given a partially specified instance, returns either a class label or specifies which feature value to acquire next. In their setting, it is assumed that during classifier induction, the learner has access to all feature values for training instances.

A logical extension of these settings is one in which there are acquisition costs both during training and testing. This is, in fact, straightforward to incorporate within the AIA framework. In Algorithm 1, for the learner $\mathcal{L}$ we can use a bounded active classifier learner, such as a bounded-depth decision tree [10]. This will now enable us to acquire feature values and class labels that will lead to the greatest increase in expected performance of the active classifier per unit cost. In related work, Kapoor et al. [20] explored approximate solutions to combining active feature-value acquisition with active classification in the *budgeted learning* setting.

7. CONCLUSION

Electronic commerce is revolutionizing the way we think about data mining, by making it possible to seamlessly integrate the processes of data acquisition and model induction. In this paper, we discussed several possible data acquisition settings, along with the challenges involved in developing solutions for each setting. We also presented a unified framework for active information acquisition, under which one can consider acquisitions of different types,

with any cost structure, for any modeling objective. We demonstrated the effectiveness of the proposed solution for the task of predicting online customer conversion, which shares the opportunity for costly information acquisition with other e-commerce tasks, such as targeting online offers. Employing active information acquisition can allow the building of predictive models at significantly lowered costs. Furthermore, if information is acquired cost-effectively, a modeler may be able to explore new sources of information, that previously were ignored because of the prohibitive cost of acquiring *all* information from each source. Using more and richer sources of data can also boost the predictive performance of models, which, in electronic commerce, can translate to more efficient processes, higher levels of customer satisfaction, reduced operating costs, and higher returns on investment.

Acknowledgments

We thank Mikhail Bilenko, Panos Ipeirotis, and Ronny Kohavi for enlightening discussions of predictive modeling and data acquisition for electronic commerce.

8. REFERENCES

[1] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)*, pages 1–10, 1998.

[2] S. Ansari, R. Kohavi, L. Mason, and Z. Zheng. Integrating e-commerce and data mining: Architecture and challenges. In *Proceedings of WEBKDD*, 2000.

[3] A. C. Bemmaor. Predicting behavior from intention-to-buy measures: The parametric case. *Journal of Marketing Research (JMR)*, 32(2), May 1995.

[4] M. Bilgic and L. Getoor. Voila: Efficient feature-value acquisition for classification. In *AAAI '07: Proceedings of the 22nd National Conference on Artificial Intelligence*, July 2007.

[5] K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of 20th International Conference on Machine Learning (ICML-2003)*, pages 59–66. AAAI Press, 2003.

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

[7] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML-95)*, pages 150–157, San Francisco, CA, 1995. Morgan Kaufmann.

[8] H. Dai, L. Zhao, Z. Nie, J. Wen, L. Wang, and Y. Li. Detecting online commercial intention (OCI). In *Proceedings of the 15th International World Wide Web Conference (WWW-06)*, 2006.

[9] D. T. Davis and J. N. Hwang. Attentional focus training by boundary region data selection. In *International Joint Conference on Neural Networks*, volume 1, pages 676–81. IEEE, 1992.

[10] D. Dobkin, D. Gunopoulos, and S. Kasif. Computing optimal shallow decision trees. In *International Workshop on Mathematics and Artificial Intelligence*, 1996.

[11] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[12] V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, NY, 1972.

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

[14] R. Gilad-Bachrach, A. Navot, and N. Tishby. Query by committee made real. In *Advances in Neural Information Processing Systems*, 2005.

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

[16] P. Haase, M. Ehrig, A. Hotho, and B. Schnizler. Personalized information access in a bibliographic peer-to-peer system. In *Proceedings of the AAAI Workshop on Semantic Web Personalization*, 2004.

[17] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 22(2), May 2006.

[18] V. S. Iyengar, C. Apte, and T. Zhang. Active learning using adaptive resampling. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pages 91–98, 2002.

[19] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[20] A. Kapoor and R. Greiner. Learning and classifying under hard budgets. In *Proceedings of the European Conference on Machine Learning (ECML-05)*, Porto, Portugal, October 2005.

[21] K. L. Keller. Conceptualizing, measuring, managing customer-based brand equity. *Journal of Marketing*, 57(1):122, 1993.

[22] J. Kiefer. Optimal experimental designs. *J. R. Stat. Soc.*, series B 21:272–304, 1959.

[23] R. Kohavi, R. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2007)*, 2007.

[24] R. Kohavi and F. Provost. Applications of data mining to electronic commerce. *Data Mining and Knowledge Discovery*, 5(1/2), 2001.

[25] S. Kullbak and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 1951.

[26] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of Seventeenth International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-94)*, 1994.

[27] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning (ICML-94)*, pages 148–156, San Francisco, CA, July 1994. Morgan Kaufmann.

[28] H. Lieberman and T. Selker. Out of context: Computer systems that adapt to, and learn from, context. *IBM Systems Journal*, 39, 2000.

[29] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, page 7680, 2003.

[30] M. Lindenbaum, S. Markovitch, and D. Rusakov. Selective sampling for nearest neighbor classifiers. In *Proceedings of*

the Sixteenth National Conference on Artificial Intelligence (AAAI-99), pages 366–371, 1999.

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

[32] P. Manchanda, J.-P. Dub, K. Y. Goh, and P. K. Chintagunta. The effect of banner advertising on internet purchasing. *Journal of Marketing Research (JMR)*, 43(1):98–108, Feb 2006.

[33] P. Melville and R. J. Mooney. Diverse ensembles for active learning. In *Proceedings of 21st International Conference on Machine Learning (ICML-2004)*, pages 584–591, Banff, Canada, July 2004.

[34] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM-04)*, pages 483–486, 2004.

[35] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. An expected utility approach to active feature-value acquisition. In *Proceedings of the International Conference on Data Mining*, pages 745–748, Houston, TX, November 2005.

[36] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. Mooney. Active learning for probability estimation using Jensen-Shannon divergence. In *Proceedings of the European Conference on Machine Learning (ECML-05)*, pages 268–279, Porto, Portugal, October 2005.

[37] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

[38] B. Mobasher and S. S. Anand. Intelligent techniques for web personalization. In *Lecture Notes in Artificial Intelligence (LNAI 3169)*. Springer, 2005.

[39] W. W. Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1/2), 2003.

[40] O. Nasraoui, C. Rojas, and C. Cardona. A framework for mining evolving trends in web data streams using dynamic learning and retrospective validation. *Journal of Computer Networks- Special Issue on Web Dynamics*, 50(10):1425–1652, July 2006.

[41] B. Padmanabhan, Z. Zheng, and S. O. Kimbrough. Personalization from incomplete data: what you don't know can hurt. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 154–163, 2001.

[42] E. Pednault, N. Abe, and B. Zadrozny. Sequential cost-sensitive decision making with reinforcement learning. In *Eigth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, 2002.

[43] F. Provost. Toward economic machine learning and utility-based data mining. Invited talk at the KDD-05 Workshop on Utility-Based Data Mining: http://storm.cis.fordham.edu/ gweiss/ubdm05/Provost-slides.pdf.

[44] N. V. Raman and J. D. Leckenby. Factors affecting consumers' Webad visits. *European Journal of Marketing*, 32(7/8), 1998.

[45] M. Regelson and D. Fain. Predicting click-through rate using keyword clusters. In *Proceedings of the 2nd Workshop on Sponsored Search Auctions*, 2006.

[46] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, 2007.

[47] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55, 1952.

[48] S. Rosset, E. Neumann, U. Eick, and N. Vatnik. Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery*, 7(3):321–339, 2003.

[49] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of 18th International Conference on Machine Learning (ICML-2001)*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[50] M. Saar-Tsechansky, P. Melville, and F. Provost. Active feature-value acquisition. Technical report, McCombs Research Paper Series No. IROM-08-06, 2006.

[51] M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*. Forthcoming.

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

[53] M. Saar-Tsechansky and F. Provost. Decision-centric active learning of binary-outcome models. *Information Systems Research*, 18(1), 2007.

[54] D. R. Self and R. F. Lusch. Direct response marketing: a comparative review. *Journal of Marketing*, 50(1), 1986.

[55] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the ACM Workshop on Computational Learning Theory*, Pittsburgh, PA, 1992.

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

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

[58] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, 2000.

[59] P. Turney. Types of cost in inductive concept learning. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning (WCSL at ICML-2000)*, 2000.

[60] W. Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th International World Wide Web Conference (WWW-06)*, 2006.

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