Selectively Acquiring Customer Information: A New Data Acquisition Problem
and an Active Learning Based Solution

Zhiqiang Zheng
A.G. School of Management, University of California, 18 Anderson Hall, Riverside, CA 92521
+1(951)827-3779, eric.zheng@ucr.edu

Balaji Padmanabhan
The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104
+1(215)573-9646, balaji@wharton.upenn.edu

Abstract

This paper presents a new information acquisition problem motivated by business applications where customer data has to be acquired with a specific modeling objective in mind. In the last two decades there has been substantial work in two different fields - optimal experimental design and machine learning – that has addressed the issue of acquiring data in a selective manner with a specific objective in mind. We show in this paper that the problem presented here is different from the classic model-based data acquisition problems considered thus far in the literature in both fields. Building on the work in optimal experimental design and in machine learning we develop a new active learning technique for the information acquisition problem presented in this paper. We demonstrate that the proposed method perform well based on results from applying this method across twenty Web usage and machine learning datasets.
1. Introduction

Consider the following decision problems:

1. A CRM manager for an online firm wants to build a model predicting a given visitor’s probability of making a purchase at the store. The manager has access to detailed Web browsing data on all visitors to the store, but may have business reasons to believe that there are unobserved, yet important, variables that significantly influence the act of purchase. Such unobserved variables could be the visitor’s intent, demographics, perhaps some measure of Web savvy and even the visitor’s activities at other Web sites. Indeed recent research (Padmanabhan et al 2005, Park & Fader 2004) has shown that such additional data can help predict online purchases better. It may be possible to acquire this information, perhaps by asking visitors to complete an online survey. But this is an expensive task since appropriate incentives will have to be provided. When the firm can only afford to target a sample of visitors, which visitors should the firm obtain the additional information from?

2. A risk manager in the account management department of a credit card firm wants to build a model of customer default and to use the model in the process of approving credit limit increases for existing customers. The bank has historical transaction data on all customers, but the manager wants to use additional selected data from customer credit reports for this task. There is strong academic support for doing so - Kallberg and Udell (2003) show that knowing how customers behave with competitors can reduce prediction error of default significantly (from 20.6% down to 12.8% in their empirical study). Dunn and Kim (2004) show empirically that a few variables not readily available to a firm (e.g. the number of credit cards on which a customer has reached the borrowing limit) have significant impact on modeling default. However, each request to a credit reporting agency for any such information has a real cost. Which customers should the manager acquire this information for?

The above two scenarios have the following common characteristics:

1. There is some readily available data generated as a natural consequence of a firm doing its business. We refer to this data as “local data” in this paper. For example, for credit card firms, all customer transactions conducted with their card constitute local data.
2. There is a specific modeling objective, and the target variable (e.g. “purchase prediction”, “customer default”) related to this is also part of the readily available data - online firms observe which customers purchased and credit card companies observe which of their customers defaulted. We use the term “local models” to refer to models for this dependent variable that are built on the local data.

3. There is additional useful information that is not readily available, but it is known what this information is. For example, the CRM manager knows that customer demographics and browsing behavior across sites is useful, and the risk manager knows that the number of credit cards held is useful. We use the term “global data” to represent local data augmented with the additional information and the term “global models” to refer to models built on the global data.

What can be done in such cases? A trivial solution for this is to acquire the additional data for all customers. In this case models built on the global data will outperform those built on local data. However this can be expensive, and as we show, even unnecessary. A second possibility is to randomly pick some customers to acquire this data from. This does not use the specific modeling objective, but provides a good benchmark.

In this paper we address the problem of determining how to selectively acquire additional customer data to build better models than those using local data alone. There has been substantial work in two different fields - optimal experimental design (OED) and machine learning – that has addressed the issue of acquiring data in a selective manner with a specific objective in mind. However as we describe later, the problem presented here is significantly different from the standard selective data acquisition scenarios considered in OED and machine learning. Building on the work in these two fields we develop a new active learning heuristic for the information acquisition problem presented in this paper. Empirically we show that this new method performs significantly better than random data acquisition for a variety of real datasets.

On a theoretical note, a more common phenomenon across many fields is the separation between data selection and data modeling. For example, in most experimental designs data is collected without having in mind a specific model (Smith 1993). Surveys are often designed without assuming how the survey results are going to be analyzed. In data mining too, data miners often first collect data independently and then decide what data mining models to build on the data. Smith (1993) in his influential presidential speech at the Royal
Statistical Society questioned this common practice and argued that statistical inference is affected by the way one selects the sample. Since inference on the population is the main objective of statistics, and inference is based on sample selected from a population, Smith (1993) argued that it is important to know how the sample is selected and vice versa. In OED various data selection methods have been developed that are specific to the modeling task. Simon and Lea (1974), in the context of machine learning for model building, also asserts that “learning should involve simultaneous search over two spaces: the hypothesis space and the instance space. The search of the hypothesis space (or the knowledge of the model) can affect how the instance space is searched.” For building case-based systems, Mookerjee and Mannino (1997) observed and cautioned that separating concept formation (model building) and retrieval strategy formation (acquisition) is a major design limitation. We follow this stream of research and propose a goal-oriented data acquisition strategy that integrates data acquisition with model building.

In Section 2 we formulate the data acquisition problem. Section 3 reviews related work and Section 4 describes a general method for selective data acquisition. The algorithm described in Section 4 is generic and can be used with various parametric and non-parametric models. We then describe, in Section 5, one specific instantiation of the general method where we develop a Bayesian method for selective data acquisition for logistic regression. We present the results of this specific method in Section 6. We then discuss the findings, limitations and future work in Section 7 followed by conclusions in Section 8.

2. Problem Formulation

We first propose, in Section 2.1, a generic model that integrates data acquisition with model building. In Section 2.2 we then present the specific data acquisition problem considered in this paper.

2.1 The Generic Data Acquisition Problem

The data acquisition problem can be treated as a decision problem. We follow the approach used in OED, particularly in Lindley (1972), Atkinson (1996) and Chaloner & Verdinelli (1995), that specifies a utility function reflecting the purpose of a model. The data acquisition problem is then to select data to maximize
the expected utility. Without loss of generality suppose there is a parametric model of the form $Y = f(X, \theta)$, where $\theta$ represents the unknown parameters for the model from $X$ to $Y$. Assume that $D$ is the dataset consisting of $n$ records. Also suppose that there is a data selection mechanism $W$ (which will be the key decision variable in this formulation) that decides which data to choose as follows:

$$W = \left\{ \begin{array}{c} d_1, d_2, \ldots, d_n \\ w_1, w_2, \ldots, w_n \end{array} \right\}$$  \hspace{1cm} (1)$$

In (1), $d_1, d_2, \ldots, d_n$ represents individual data points of $D$, and $w_1, w_2, \ldots, w_n$ represents the weights (or probabilities of selection) that the mechanism assigns to the data points. A selection mechanism, hence, is one that specifies the probability of each individual instance of $D$ being selected for modeling. We use $W(D)$ to represent the sample of data selected when applying $W$ to $D$. The parameters of the model $\theta$ are estimated from the selected data $W(D)$. Next define a utility function $U(\theta, W, D)$ to reflect the goal of building the model, such as minimizing the prediction error or minimizing the variance of the parameter estimates.

Hence the decision problem becomes how to develop the best data selection mechanism $W$ to maximize the expected utility as shown in (2) below, where the only decision variable is $W$. The integrand in (2) can be interpreted from right to left as (i) first applying selection mechanism $W$ to obtain a sample of data $D$ (i.e. $p(D|W)$), (ii) then estimating the parameters based on the selected data (i.e. $p(\theta/D, W)$) and (iii) computing the utility of the model (i.e. $U(\theta, W, D)$)

$$U(W^*) = \max_{w} \int \int U(\theta, W, D) \times p(\theta/D, W) \times p(D/W) d\theta dD$$ \hspace{1cm} (2)$$

Notice that equation (2) integrates data selection with model building and the formalism is used as follows. First the decision maker determines the optimal $W^*$ (by solving eq. (2)). Then she determines what data is to be selected (by applying $W^*$ to the data) and builds the model on the data subsequently.

The central issue in this decision problem is specifying an appropriate utility function reflecting the goal of the model. A common goal for parametric modeling is to estimate the parameter $\theta$ as accurately as possible. If there is prior knowledge on the parameter $\theta$, then Bayesian updating can be used (after choosing
to derive the posterior probability about \( \theta \). A well-known Bayesian updating mechanism is the Kullback-Leibler distance (Kullback 1959), maximizing the difference between posterior and prior as follows.

\[
W^* = \arg \max \int \int \log \frac{p(\theta | D, W)}{P(\theta)} \ p(\theta, D | W) d\theta dD
\]

In (3), notice that \( p(\theta, D | W) = p(\theta | D, W) \times p(D | W) \).

### 2.2 The Specific Data Acquisition Problem

In this section we present the specific data acquisition problem studied in this paper. Assume that there exists a specific target variable, \( Y \). For example, \( Y \) can be defined as whether or not a user transacts at a Web site within a session. Let \( N \) be the number of total data points. Let \( D_L \) be the local data that consists of attributes \( X_1, X_2, \ldots, X_M, Y \). Assume that the values of these attributes are known for all points. Let \( D_G \) be the global data that consists of attributes \( X_1, X_2, \ldots, X_M, X_{M+1}, \ldots, X_P, Y \). Here assume that the values for attributes \( X_{M+1}, \ldots, X_P \) are all unknown initially, but can be acquired. We assume that initially only incomplete information (the local data) is available for all \( N \) records, and after some data acquisition procedure, complete information is acquired for \( K \) of these records where \( K \leq N \).

![Scenario 1: Local Model](image1.png)  ![Scenario 2: Global Model](image2.png)

**Figure 2.1:** Two scenarios in model building

In particular we assume that a firm has local data and can build a local model as shown in Scenario 1 in Figure 2.1. Further, as motivated in Section 1, we assume that the decision maker knows the set of global variables that are useful, and wants to build a global model for this problem. Given that global data on only \( K \) records can be obtained, the problem addressed here is identifying which \( K \) records to acquire given the
specific modeling objective. The interesting tradeoff between scenarios 1 and 2 is that the model built in scenario 2 uses complete information, but on fewer points.

The utility function that only involves the global data becomes:

\[ U(W^*) = \max_{w} \int \int U(\theta, W, D_g) \times p(\theta / D_g, W) \times p(D_g / W) d\theta dD_g \]  

(4)

Note that this paper addresses the problem of information acquisition (of \( D_g \)) during the course of model building – i.e. the goal is to selectively acquire global data in order to build a global model. This can be distinguished from another information acquisition problem, where sequential information acquisition is done in the course of problem solving. An illustrative scenario is a doctor making a diagnosis. “Global” variables in this case may consist of readings such as a patient’s blood pressure and temperature, as well as (expensive) readings such as the results of an EKG or an MRI scan. In this case, the doctor may start with a blood pressure reading. Based on what this (acquired) value turns out to be, additional tests are ordered, and the process continues in this manner. A decision tree model for diagnosis (perhaps even implicit in the doctor’s mind) can permit such sequential information acquisition during problem solving. However in this example note that it is assumed that the decision model already exists. When sequential information acquisition is done in the course of model-building (as in this paper) the scenario where this is applicable is the following. We assume that a new case will show up “as a vector of variables”, but that we only currently have limited (local) data on several training examples. The idea of sequential information acquisition here is to build a good global model that can help make predictions for these new cases (such as the ones described in the introduction). To illustrate this in the context of the medical example above, consider the following hypothetical scenario (similar to the ones described in the introduction). Assume that all patient records are available electronically, and that a doctor’s office can automatically retrieve the complete medical history of any patient who comes in with a problem. For such patients a goal is to make predictions based on current symptoms, the results of tests plus summary variables based on the entire medical history of the patient. The office may have a large amount of historical data on patients’ visits and the outcomes, but the historical data (collected before electronic patient records were available) does not contain the complete medical histories at
the times of the visits. A sequential information acquisition procedure here helps to build a global model by sequentially acquiring some of the past patients’ medical histories in order to build a good global model.

This paper only addresses the latter acquisition scenario for scope reasons. Integrating these two approaches is an interesting possibility, and we discuss this in Section 7 of the paper.

3. Related Work

The selective information acquisition problem studied in this paper is a novel problem and to our knowledge, there is no related work that directly addresses this. However the broader selective acquisition concept has been addressed in several fields including Information Systems (IS), model-based survey sampling, optimal experimental design (OED) and machine learning.

In IS, selective information acquisition has been studied in the form of decision-theoretic acquisition. A decision-theoretic information acquisition framework was first proposed by Moore and Whinston (1986, 1987) and then applied in Moore et al. (1996), Dos Santos and Mookerjee (1993), Mookerjee and Mannino (1997). Moore and Whinston (1986, 1987) investigated a fundamental information acquisition problem in decision support systems, where there is opportunity to acquire additional information to reduce the uncertainty associated with the final decision. The paper presents a general framework including two conceptual problems facing the decision maker: 1) developing an optimal information gathering strategy and 2) developing an optimal decision strategy. Further the paper proposes a decision-theoretic model to gather information to optimize the final decision. Moore et al. (1997) then applied this model to address the problem of acquiring information on agents’ preferences for a manager to allocate scarce resources in a socially optimal manner. The information gathering (acquisition) strategy is designed to efficiently acquire information in such a way as to maximize the expected payoff. Mookerjee and Dos Santos (1993) developed an optimal sequential information acquisition strategy for inductive expert systems, where information is acquired to reduce uncertainty only if the benefits gained from acquiring the information exceed its cost. Mookerjee and Mannino (1997) empirically show that separating concept formation (model building) and retrieval strategy formation (acquisition) is a design limitation for case-based systems. Further they propose a decision-theoretic solution in which concept formation and retrieval strategy are combined. The decision-
theoretic framework, which suggests linking information acquisition with the final decision (a goal-oriented feature), is generic in nature. Our proposed method, which ties information acquisition with model building, also falls into such a framework.

In conventional survey sampling (Chaudhuri et al. 1992, Cochran 1977, Singh and Mangat 1996), the focus of data acquisition is on drawing inferences about the population from collected survey responses. Survey sampling primarily uses random sampling. When the population consists of homogeneous groups, *stratified sampling* or *cluster sampling* (Chaudhuri et al. 1992) is often used, where the population is divided into subgroups and then each group is randomly sampled. Note that this type of survey sampling is usually goal-independent: the sampling procedure does not depend on how the data is to be used in model building. Smith (2001) points out that this is mainly due to the separate role of data sampler and data analysts. When the data analyst is the sampler, one can make use of the model directly during the sampling process. This is the essence of model-based survey sampling. The influential, yet once thought to be “provoking” (according to Smith 2001), work of Royall (1970), argues for using the models to select data. The suggestion was that models should be used first to predict the unobserved values. Then conditioning on the prediction, it can be decided where to sample. There are similar ideas in OED, active learning and in this paper. Royall (1976) argued in favor of a model-based inference by citing the fact that for randomization inferences the likelihood function is un-informative. After three decades of development of model-based survey sampling these methods were gradually accepted.

Understandably, the counterpart of the model-based approach, optimal experimental design (OED), also had a similar history. Experimental design in general deals with acquiring data through experiments when the data is not available in natural settings (Box et al. 1978, Cook and Campbell 1979). OED aims to generate a smaller sample in experimental design than regular randomized experimental design (Atkinson and Donev 1992, Atkinson and Bailey 2001, Cohn 1996). The idea behind OED was set out by Kiefer (1959). According to Atkinson and Bailey (2001), the paper had “a rough reception and the published discussions related to the paper were hostile”. Subsequently, however, OED has become one of the major sub-disciplines of
experimental design and has appeared in the standard textbooks (see Atkinson and Bailey (2001) for an interesting chronology).

OED is essentially an incremental approach for data acquisition. At each phase, OED approaches decide which subject to be experimented upon on the basis of how much the addition of each subject would contribute to the model (rather than randomly selecting from the pool). In doing so, OED uses optimization techniques to decide which subject to acquire data for. A variety of optimization techniques have been developed such as D-optimality, G-optimality, A-optimality and I-optimality (Atkinson and Donev 1992, Atkinson and Bailey 2001, Cohn 1996). Both D- and A-optimality criteria are related to the information matrix \( X'X \) for the design, where \( X \) is the matrix of independent variables. Roughly, a good design should "minimize" the variance \((X'X)^{-1}\), which is the same as "maximizing" the information matrix \( X'X \). A-optimality is based on the sum of the variances of the estimated parameters for the model, which is the same as the sum of the diagonal elements, or trace of \((X'X)^{-1}\). Like the determinant, the A-optimality criterion is a general measure of the size of \((X'X)^{-1}\). Both G-optimality and I-optimality (the average prediction variance) are based on the variance of prediction of the candidate points, which is proportional to \( x'(X'X)^{-1}x \), where \( x \) stands for the candidate points. As this formula shows, these two criteria are also related to the information matrix \( X'X \). Minimizing the average prediction variance has also been called I-optimality, the "I" denoting integration over the candidate space. Further reviews can be found in (Atkinson and Donev 1992, Atkinson and Bailey 2001, Cohn 1996). OED primarily focuses on parametric models, especially linear regressions. Active learning, a relatively more recent approach in data mining, addresses a similar problem as OED, but it has focused on non-parametric models such as decision trees and neural networks.

The concept of active learning was first proposed in Cohn (1996) to describe learning methods where learning models have control on what data to feed into the model for training. Tong and Koller (2001) made an interesting analogy suggesting that “a traditional learner is a student that sits and listens to a teacher while an active learner is a student that asks the teacher questions, listens to the answers and asks further questions based upon the teacher's response.” As pointed out in Saar-Tsechansky and Provost (2001), the notion of active learning can be traced back to Simon and Lea (1974) where they proposed that learning should involve
simultaneous search over two spaces: the hypothesis space and the instance space. The search of the hypothesis space can affect how the instance space is searched.

In data mining, active learning is often applied when “labeled” data (i.e. \( Y \) is known) is expensive but a lot of unlabeled data is readily available. One example is information retrieval. In order to train an information retrieval (IR) system it is expensive to get the target variable because labeling whether a document is relevant or irrelevant by human subjects could be very expensive; yet unlabeled documents are abundant. Active learning aims to minimize the number of labels needed for training.

In the recent past, there have been several new methods developed in the active learning area. The different active learning methods mainly differ in the way of determining the utility of a data point to the specific model to be built. They broadly fall into two categories: heuristic based and optimization based approaches (Hasenjäger and Ritter 1999). Query by Committee (QBC) (Engelson and Dagan 1999, Freund 1997) is the best-known heuristic-based approach. QBC employs several committee members (each of them is a model) and each member makes predictions on the unseen data. The data points chosen are those in which there is maximum disagreement among the committee. Another type of heuristic was proposed in Hasenjäger and Ritter (1999) and MacKay (1992) that is similar to boosting - selecting those data points that the model misclassifies. Optimization approaches employ an objective function and those data points that optimize this objective function are selected. Some well-known objective functions are variance-based objective functions and those based on some measure of information gain. In Cohn et al. (1996) data points are selected to minimize the overall prediction variance of a model. Saar-Tsechansky and Provost (2001) selects those data according to the variance of bootstrap predictions of class probability estimates. Among information gain based approaches, MacKay (1992) used a Shannon information criterion and suggested that for Gaussian distributions, this criterion is equivalent to the heuristics that choose data where the current model yields the largest error. In Tong and Koller (2001), a Kullback-Leibler (KL) entropy for the problem of Bayesian parameter estimation is used.

Though ideas in model-based survey sampling, OED and active learning have similarities to our problem, the methods proposed are not directly applicable to the data acquisition scenario considered in this
study for the following key reason: model-based survey sampling, OED and active learning all study a different context, where the aim is to acquire the target (dependent) variable, and the set of all explanatory variables is fixed and the values known. In this paper we attempt to acquire explanatory variables in the form of global customer data, and this is an important and novel extension. None of the prior work addresses this issue because it did not arise naturally in the traditional contexts studied in the fields.

4. The Generic Selective Data Acquisition Method
Motivated by the approaches in active learning and OED, in this section we present a generic data acquisition algorithm for the problem considered in this paper. It is “generic” in that although the selection method presented is clearly model dependent, it is not optimized for any specific type of model. In Section 5 we show how this generic approach can be used for a specific model (logistic regression).

Similar to OED and active learning, our solution is also iterative and is based on the spirit of Bayesian approaches. Given the current global model we aim to identify the data point (or points) likely to be of greatest utility to the model. Each time such a decision is made, the actual global data for that “most valuable point” (MVP) is acquired, and the process is repeated\(^1\). If the global data \((X_1, X_2, \ldots, X_{M+1}, \ldots, X_p, Y)\) is known for all the additional points, selecting the MVP is straightforward - e.g. add every point separately and re-build the model each time, and subsequently pick the point that improves the model the most. However, in our problem context for every additional data point we consider for acquisition, only local data \((X_1, X_2, \ldots, X_M, Y)\) is known. To address this issue, based on the currently available global data our approach builds an imputation model\(^2\), that imputes the unknown global data \((X_{M+1}, \ldots, X_p)\) for every point. As mentioned above, the MVP can then be selected once the global data is estimated for all the prospective points. Then after acquiring the real global data for the MVP, both the global model and the imputation model are re-built.

Note that the imputation model is a key component in estimating the MVP and it improves over time since the imputation model, like the global model, is continuously re-calibrated at the end of each iteration. Therefore every point that is acquired at the end of each iteration, has to be good for two different purposes:

\(^1\) Instead of acquiring just one point at a time, the algorithm allows acquiring \(sz\) points at a time, where \(sz\) (step size) is treated as a user-specified input.

\(^2\) e.g. multiple imputation or nearest neighbor substitution as is discussed in Little and Rubin (1987).
1. The acquired point should help improve the global model to be built.

2. The acquired point should help improve the imputation model so that subsequent imputations are more accurate.

This dual objective is the main distinguishing feature of our approach as compared to existing methods in OED and active learning. Prior work on multi-objective optimization (Cook and Wang 1994, Steuer 1986) suggested two approaches: 1) Constrained optimization and 2) a weighted average approach. In the first case, when the multiple criteria are of different priorities one approach is optimizing the higher priority criterion \( C_1 \) subject to the constraint of \( C_2 \) (Cook and Wang 1994). Although it may first appear that improving the global model is “more important”, this is not necessarily the case here since the global model will be improved by selecting good points – but determining which points are good needs a good imputation method too. Hence we adopt the weighted average approach because of its simplicity and the flexibility it may provide by permitting each objective to be weighted differently if needed. In the implementation of our method we use equal weights as a default. Determining the optimal weights for the objectives is an interesting problem but beyond the scope of this paper.

For each candidate data point \( d_G \), we compute two scores: 1) a score based on how the data point contributes to improving the global model, denoted as \( \text{Score}_M(d_G, \theta, W) \) where \( \theta \) represents the model of the target variable and \( W \) represents the selection mechanism as described in Section 2 and 2) a score based on how much the point contributes to improving the imputation model, denoted as \( \text{Score}_I(d_G, \theta, W) \). Given a user-specified weight \( \alpha (0<\alpha<1) \) on \( \text{Score}_M(\cdot) \), the utility of \( d_G \) is

\[
U(d_G, \theta, W) = \alpha \cdot \text{Score}_M(d_G, \theta, W) + (1-\alpha) \cdot \text{Score}_I(d_G, \theta, W)
\] (6)

and \( d_G^* = \text{argmax} U(d_G, \theta, W) \)

Since eq. (6) involves both the model of the target variable and the imputation model, it is hard to derive a closed form solution for \( W^* \) for general models. Below we describe a computational approach that details the optimal selection mechanism. Let \( D_L \) denote the set of all known local data (\( D_L \) corresponds to the shaded area in Scenario 1 of Figure 2.1). Each record in \( D_L \) therefore consists of \( ID, X_1, X_2, \ldots, X_M, Y \) where \( ID \) is an index ranging from 1 to \( N \) and \( |D_L| = N \). \( D_L \) remains constant throughout the procedure. Let \( D_G \) denote the set
of known global data. Each record in $D_G$ therefore consists of $ID, X_1, X_2, \ldots, X_M, X_{M+1}, \ldots, X_P, Y$. $|D_G| = 0$ initially and $|D_G| = K$ at the end of the data acquisition process ($D_G$ corresponds to the shaded area in Scenario 2 of Figure 2.1). The algorithm $DODA$ (Dual Objective Data Acquisition) is presented in Figure 4.1. The inputs are (a) the local data (b) a classifier $C$ (c) a scoring function $Score_M$ that computes how good a new data point is to the global model using classifier $C$ (d) an imputation model $IM$ (e) a scoring function $Score_{IM}$ (f) a weight $\alpha$ to combine the two scores (g) a stopping criterion for the sequential acquisition process and (h) a minimum step size in which the data is acquired.

Input: local data $D_L$, a stopping criterion, step size $sz$, classifier $C$, score function for the classifier $Score_M$, imputation method $IM$, score function for the imputation model $Score_{IM}$ and weight $\alpha$

Output: $K$ points for which global data is acquired

1. $N = |D_L|$
2. $I_L = \{1, 2, \ldots, N\}$ /* index of all points in $D_L$ */
3. $D_G = \{}$ /* known global data, initially empty */
4. $I_G = \{}$ /* index of all points in $D_G$ */
5. $S \leftarrow$ randomly select $sz$ integers from $I_L-I_G$
6. repeat {
7. Forall ($j \in S$) {
8. Get $d_g = \{ID, X_1, \ldots, X_P, Y\}$ for the element in $D_L$ where $ID=j$
9. $D_G = D_G \cup d_g$
10. $I_G = I_G \cup \{j\}$
11. }
12. Build an imputation model $IM$ from $D_G$
13. $UID = I_L-I_G$ /* current set of IDs for which global data is unknown */
14. $U_G = \{t \mid t \in D_L$ and $t.ID \in UID\}$
15. Forall ($t \in U_G$) {
16. $d_G = IM(\text{record } t)$ /* impute record $t$ using $IM$ to get $d_G$ */
17. $Score(t) = \alpha \times Score_M(t) + (1-\alpha) \times Score_{IM}(t)$
18. }
19. $S \leftarrow$ select $sz$ IDs in $U_G$ with the best scores
20. } until the stopping criterion is met
21. Output: $D_G$

Figure 4.1: Algorithm $DODA$ – Dual Objective Based Data Acquisition

The algorithm acquires data points to maximize the performance of a model of the target variable for a given classifier $C$. Assume that at some point in the process the algorithm has a subset of known global data. The idea is to choose the next point to maximize a composite score based on the dual objectives. In order to do this, for each candidate point the algorithm imputes its global data first and then adds this record to the known global data, and considers the goodness of the classifier and the goodness of the imputation model and chooses the best candidate point based on a weighted average score.
Specifically, the set $I_G$ incrementally maintains the list of indices in the local data for which global data is acquired. Initially this set consists of $sz$ random points for which global data is acquired (steps 1-11). Once an initial set of global data is acquired, step 12 builds an imputation model from the known global data. Steps 13 and 14 identify the set of all transactions for which global data is unknown ($U_G$). For each of the records in $U_G$, the unknown values are imputed (step 16) and the score for this data point is computed in Step 17 using a weighted average of the two scores. Next Step 19 selects the best $sz$ points to acquire. This entire process is continued until a user-specified stopping criterion is met. Below we discuss some possible stopping criteria that may be used.

An important consideration in choosing a stopping condition is the computational cost it imposes. This is a particular concern in active learning, which is known to be a computationally expensive process. The simplest stopping rule is to specify the desirable number ($K$) of customers to acquire data from. The actual choice of $K$ may stem from different considerations, such as resource constraints. For instance, in such a cost-based perspective, given the cost of acquiring a customer’s information and an available budget a decision maker determines the affordable number of customers to acquire accordingly. From a computational standpoint this criterion will not add any additional burden on DODA. Another stopping rule is for the user to specify a desirable performance measure such as an upper bound for the prediction error. However testing the stopping criterion in a loop iteratively will involve building a model and testing it on out of sample data each time, and this is computationally expensive. A more unified approach is to examine the real cost-benefit tradeoff, and to stop data acquisition when the cost exceeds the benefit. In this scenario, the user ought to have a good understanding of both the cost and the expected benefits of acquiring additional customer information. Our experiments (Section 6) shed some light on the expected benefits. Perhaps such an approach can be used to determine an optimal $K^*$ to maximize the overall benefit derived from the process. This needs a thorough understanding of the domain and a careful investigation into the cost-benefit tradeoff. For the experiments reported in this paper we assume a user specifies a desirable number of points as the stopping condition and vary this number to construct standard active learning performance curves, based on which the algorithm is evaluated.
5. The Score Functions

As discussed above, each data point for which the global data is unknown needs to be scored based on how well it helps the imputation model and the global model. In Sections 5.1 and 5.2 we specify the two scoring functions.

5.1 Score Function for the Imputation Model

We estimate the value of a new data point for imputation by applying a bootstrapping technique \textit{SVID} (Scoring based on Variance of the Imputed Data). The main idea behind \textit{SVID} is that if a data point cannot be imputed well, then it should be useful to acquire to build a better imputation model. Similar heuristics have been proposed in the Query By Committee (QBC) theory of Freund et al. (1997) and the uncertainty region of Hasenjäger and Ritter (1999). Freund et al. (1997) showed that this heuristic indeed can lead to information gain and can decrease prediction error exponentially. Specifically, \textit{SVID} estimates the variance of imputation by using several bootstrap samples to build different imputation models. For every candidate point the imputation models provide predictions based on which the variance of the prediction can be computed. Points for which the imputed global data has higher variances, are considered as good candidates for which true global data can be acquired. The score function is computed as follows. Suppose there exists an imputation model \textit{IM} (e.g. those described in Little and Rubin 1987, Schafer and Olsen 1998) and a number of bootstrap samples \textit{B}. We first generate a set of bootstrap samples \textit{S}_k, k=1, ...B from \textit{D}_G; then \textit{IM} is applied on each \textit{S}_k to generate imputation model \textit{IM}_k, for the purpose of predicting the missing global data \textit{U}_G. For each data point \textit{X}_i, i \in \textit{U}_G, we then impute the value \textit{X}_{ij} for each variable \textit{j}, \textit{j}=M+1, ..., \textit{P}. Finally the score of \textit{X}_i according to how useful it is for imputation is given by

\[
\text{Score}(i) = \sum_{j=M+1}^{P} \sqrt{\frac{1}{B} \sum_{k=1}^{B} (x_{ijk} - \mu_{ij})^2 / B}
\]

(7)

where \( \mu_{ij} = \left( \frac{1}{B} \sum_{k=1}^{B} x_{ijk} \right) / B \)

Equation (7) computes the variance of imputation over the \textit{B} imputation models with regard to data point \textit{X}_i, where \textit{x}_{ijk} stands for the value of the \textit{i}^{th} record, \textit{j}^{th} variable under the \textit{k}^{th} imputation (among the total \textit{B} multiple imputations). Computing the scoring function is presented formally in Figure 5.1.
Given a dataset $D_L$, current available global data $D_G$, imputation method $IM$, and number of bootstrap samples $B$,

**Input:**
- $D_L$, the local data set.
- $D_G$, the current available global data set.
- $IM$, the imputation method.
- $B$, the number of bootstrap samples.

**Output:**
Scores for each data point $X_i$, $i \in U_G$.

1. $N = |D_L|$ /* index of all points in $D_L$ */
2. $I_L = \{1, 2, ..., N\}$ /* index of all points in $D_L$ */
3. $D_G \leftarrow$ known global data
4. $I_G \leftarrow$ index of all points in $D_G$
5. Generate $B$ bootstrap samples from $D_G$
6. Build $B$ imputation models $IM_1, IM_2, ..., IM_B$ by applying $IM$ to $B$ bootstrap samples of $D_G$
7. $UID = I_L - I_G$ /* current set of IDs for which global data is unknown */
8. $U_G = \{i \mid i \in D_L$ and $i.ID \in UID\}$
9. For all $i \in U_G$
   10. $x_{ij} \leftarrow$ the imputed value for variable $j$ of record $i$ using imputation model $IM_k$ ($M+1 \leq j \leq P, 1 \leq k \leq B$).

The scoring function $SVID$ (Scoring based on Variance of Imputed Data) is defined as:

$$Score(i) = \sum_{j=M+1}^{P} \left( \frac{\sqrt{\sum_{k=1}^{B} (x_{ijk} - \mu_{ij})^2}}{B} \right) \mu_{ij} \left( \frac{\sum_{k=1}^{B} x_{ijk}}{B} \right)$$

**Figure 5.1** Scoring Function $SVID$ (Scoring based on Variance of Imputed Data)

### 5.2 Score Function for the Classifier

In general, a natural method of computing this score is based on adding each data point, building a global model and then scoring each point based on how the global model is improved by adding this point. One possible approach along these lines is to test how the point can improve prediction. While this is appealing, the complexity is high since scoring each point requires re-building a global model and testing it on out of sample data. If the number of points that need to be acquired is large this may be prohibitive, but future research is needed to examine if there are efficient ways of doing this. The alternative is to use scoring functions that do not require rebuilding models when evaluating data points. In particular, in this section we present one such scoring function specific to logistic regressions.

Our scoring function uses a Bayesian approach and is based on minimizing the posterior variance of the estimated parameters. This idea can be applied to generic parametric models, the goal of which is often on estimating parameters with low variance. For the linear regression, it is well known that there exists an optimal criterion (e.g. D-optimality) with respect to selecting data to minimizing parameter variance.

---

3. In an extreme case corresponding to a sequential acquisition with step size of 1, the above steps may have to be repeated $K \times N - K(K-1)/2$ times where $N$ is the amount of global data that can be potentially acquired and $K$ is the amount of data that is acquired. (i.e to pick the first point $N$ points need to be considered, to pick the second $(N-1)$ points need to be examined and so on until $K$ points are acquired). If there are 100,000 possible records and we need to acquire 1000 then this may need almost 100 million models that need to be built and tested.
(Atkinson and Donev 1992). The OED field has developed a variety of optimality criteria for linear models as discussed previously in Section 3. Developing the optimal criteria for non-linear models is harder and unlike the linear regression, a closed-form solution does not always exist. So far the OED studies on non-linear models have been mainly restricted to generalized linear models. Chaloner and Verdinelli (1995) derived the criteria for a simple logit model with only one independent variable. Kuhfeld et al. (1994) proposed using the same criterion for linear regression (Max |X'X|) as the approximation for logistic regression and argued that it was efficient for a variety of applications. Little work has been done to study the Bayesian optimal criteria for general logistic regression with more than one variable.

Below we derive a Bayesian score function for logistic regressions. Assume at a certain acquisition stage, K data points have been acquired among the total N points. A logit model is then built on these K records and \( \hat{\beta}_k \) is the vector of parameter estimates. An imputation model is also built and the global data of the remaining \( N-K \) points are imputed from the model. Thus the utility of a candidate \((K+1)^{th}\) point would be equal to the expected reduction of the variance of parameter estimates after acquiring this point. Denote as \( X \) the input vector consisting of \((K+1)\) points, among which \( K \) points were acquired and the \((K+1)^{th}\) one was imputed\(^4\). Let \( \hat{Y} \) be a \((k+1) \times 1\) output vector where all the \( Y \) values are predicted using \( \hat{\beta}_k \). The Bayesian score function of the logistic regression for a candidate \((K+1)^{th}\) point is given in Proposition 1.

**Proposition 1:** If the objective is to acquire data to minimize the variance of posterior parameter estimation for logistic regression, then the candidate point for which the score function \( \text{Score-Log} = | \hat{Y}'(1-\hat{Y})X'X | \) is maximum should be acquired.

The proof is presented in Appendix A. We note that this score function is related to the well-known \( d \)-optimality criteria for a linear regression. The \( d \)-criterion aims to maximize \(|X'X|\), the determinant of the covariance matrix (Atkinson and Donev 1992). For a simple linear regression (with one independent variable only), the variance matrix \( X'X \) degenerates into \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \), which is proportional to the variance in \( X \).

\(^4\) Note that imputing this \((K+1)^{th}\) point is necessary, because we don’t have the global data for this record before we actually decide to acquire it.
We prefer the variance in $X$ to be large because the wider the range of $X$ examined, the more confident we can be about the estimation of the coefficients (Greene 2000). This is illustrated in Figure 5.2. Suppose two linear models have been fitted to two data sets as shown in Figure 5.2. Assume the prediction variances on $Y$ of the two linear models are the same (both are within the range of $y_1$ and $y_2$ in Figure 5.2). Clearly the variance of $x$ is larger in the right chart, and it is also clear that the linear curve of the right chart is better.

![Figure 5.2 Graphical Illustration of Score-Log](image)

The other component $\hat{Y} (1 - \hat{Y})$ represents the variance of prediction since $Y$ is a binary variable. $\hat{Y} (1 - \hat{Y})$ yields the highest value when logit probability $\hat{Y}$ equals to 0.5, i.e., the model is least certain about how to classify the datum. Hence in summary, the score function $|\hat{Y} (1 - \hat{Y})X'X|$ for logistic regression suggests that we want to examine not only the data points that represents a wide range in $X$, but also at the same time those points that have higher predictive variance $\hat{Y} (1 - \hat{Y})$.

**5.3 Instantiating and Evaluating DODA**

We presented two score functions with respect to the dual-objectives in Sections 5.1 and 5.2. We use these two score functions to instantiate DODA to generate a specific data acquisition algorithm, DODA-Log, for logistic regression. We choose a weight $\alpha = 0.5$ to combine the two score functions. The baseline imputation method used in the algorithms is nearest neighbor substitution (Little and Rubin 1987). We chose this because it is one of the most effective imputation methods when little is known about the distribution of the data (Little and Rubin 1987, Schafer and Olsen 1998). The number of bootstrap samples generated for multiple imputation is chosen to be 5. To examine the performance of DODA-Log we need to run the sequential acquisition procedure over the entire range of data, from say $K=1\%$ all the way until $K=100\%$. 


where $K$ is the percentage of global data acquired. Hence in the experiments we use a user-specified $K$ as the stopping criterion and vary this in increments of 2%.

In evaluating $DODA-\text{Log}$, our goal is to examine how much it reduces the amount of data needed for training models (i.e. logistic regression) without compromising the model performance. A common benchmark used in active learning is the random selection method. An active learning method is considered to be efficient if it achieves the same level of model performance with significantly fewer data than random selection does. In the active learning field, this is often done experimentally, i.e., by examining how the algorithm performs on real datasets. There are also some approaches to demonstrate the efficiency of active learning methods theoretically using the PAC learning criterion (Schuurmans and Greiner 1995, Greiner and Grove 2002). A PAC-learning criterion specifies the number of examples a model needs to achieve an error rate smaller than $\varepsilon$ with probability at least $1-\delta$ (Valiant 1984). However the PAC-learning criterion needs to make specific assumptions on data distributions and it is often limited to a few simple modeling problems (e.g. the concept learning problem in Valiant 1984). For example, Schuurmans and Greiner (1995) described an optimal sequential data acquisition procedure that can reduce the number of examples needed from $1/\varepsilon \ln(1/\delta)$ for a random method down to $1/\varepsilon$. However, the procedure is only for the specific half-space concept-learning problem and for uniformly distributed data. For more general learning problems it can be intractable to demonstrate data efficiency theoretically using the PAC-learning criterion. In the next section we follow the convention of active learning and demonstrate data efficiency experimentally.

6. Results

We first discuss the datasets used, the benchmarks against which we compare the algorithm, the performance measures and two relevant scenarios for comparison. We then present the results of evaluating $DODA-\text{Log}$.

6.1 Experiment Setup

We tested $DODA-\text{Log}$ on 20 datasets all of which have a binary target variable. Ten of them are Web usage datasets where the target variable is whether the user made a purchase in the remainder of the session and the explanatory variables are features of a session. For details about the data preprocessing algorithms and the
metrics the reader is referred to Padmanabhan et al. (2005). For these ten datasets the selection of global versus local data is natural – the data captured by individual Web sites about user browsing behavior at that site is “local” information while features about across-site behavior are part of the global data. In each of these datasets, 15 of 40 explanatory variables are local variables. The global data include 25 additional explanatory variables representing users’ browsing behaviors across sites (e.g. number of sites visited in this session). In addition to the Web usage datasets, we also tested DODA-Log on datasets maintained at the UCI repository for machine learning research (Blake et al. 1998). For the UCI datasets, each dataset as a whole is treated as the global data, and the local data is generated by randomly hiding one-third of the variables. The number of variables in the UCI datasets considered ranges from 4 to 16.

In the absence of global data, the only available data are all the local variables for all \( N \) customers (scenario 1 in Figure 2.1). Based on this local data, we build a local model and treat this as the first benchmark against which global models are compared. The next two benchmarks are based on global models with alternate data selection strategies. The second benchmark we use is the global model built using random data acquisition. The third benchmark is a greedy heuristic which acquires the data points where the current model performs worst. More specifically, at a certain iteration, a global model built on \( K \) available global data points is used to make predictions on the \((N-K)\) unknown points, and the data point that the current model makes the worst prediction on is chosen. This type of heuristic has been used in (Hasenjäger and Ritter 1999, MacKay 1992) for different problems.

In order to compare DODA-Log with these three methods, we define a critical mass metric as the percentage of complete data that is required to build a model that is better than the local model. This metric is computed as \( K_{\text{min}} / N \) where \( K_{\text{min}} \) is the minimum number of global data points needed to build a global model that outperforms the local model. For all models we use the mean squared error on a random 50% out of sample data to measure the model performance.

A subtle aspect specific to our problem context is that there are two types of out of sample data that are relevant to consider for the following reason. In practice, there are two types of customers (or just data points) who may be encountered after the entire modeling process is completed (i.e. after acquiring \( K \) points and
building the final global model). First, there are friends, customers for whom \(X_1, X_2, \ldots, X_P\) are all known and the task is to predict \(Y\). For the credit risk example if the customer has applied for a new card the company is often authorized to access his or her credit report. Second, there are strangers, customers for whom only local data \((X_1, X_2, \ldots, X_M)\) will be available when these customers are encountered. Suppose the credit card company wants to score potential customers in order to select some to send pre-approved cards to. According to FTC regulation\(^5\) in the US, the firm is not allowed to access the credit report unless the customer applies for a card. In this case, only local data will be available and these customers represent “strangers”. In order to apply the global model on such out of sample data, \(X_{M+1}, \ldots, X_P\) will have to be imputed using the final imputation model built during data acquisition, and then the final global model can be applied to \(X_1, X_2, \ldots, X_P\) in order to make the prediction. Clearly, doing well on strangers is more difficult that doing well on friends, and our experimental results show this. In this section we only present the results for the friends scenario. Additional results for the strangers scenario is in the electronic supplement to this paper.

### 6.2 Results

We vary the desired amount of global data \((K)\) from 0% to 100% of the data in order to observe the performance of each method over the entire range. Due to space constraints, we do not present plots for all the 20 datasets. Figures 6.1 and 6.2 present two examples. The X-axis represents the percentage of acquired global training data and the Y-axis represents the MSE (mean squared error) of the models on the out of sample data. Each learning curve shows how MSE decreases as more global data are acquired. As mentioned before, the benchmark local model is built using the local variables in the entire training data and thus represents a straight line in this graph. Note that the point when all methods acquire 100% of the global data corresponds to the MSE of a global model built using global variables in the entire training data.

Consider the performance on the Penndigits1 dataset (Figure 6.1). \textit{DODA-Log} performs uniformly better than the random and greedy data acquisition methods. Observe that from just 18% of acquired data based on \textit{DODA-Log}, a better model can be built than from using the entire local data. In contrast, the random method and the greedy method need to acquire 32% and 68% of global data respectively to outperform the local

---

\(^5\) See FTC regulation on credit report at http://www.ftc.gov/bcp/conline/edcams/fcra/index.html
model. The performance on the Amazon dataset is shown in Figure 6.2. DODA-Log is still uniformly better, though the benefit over random and greedy acquisition methods is lower. In this case the critical mass is 12%, 18% and 16% for DODA-Log, random and greedy methods respectively.

![Figure 6.1: Performance on the Pendigits1 dataset](image1)

<table>
<thead>
<tr>
<th>DataSet</th>
<th>Area over local</th>
<th>Critical Mass % (CM)</th>
<th>Area after CM</th>
<th>Critical Mass % (CM)</th>
<th>Area after CM</th>
<th>Critical Mass % (CM)</th>
<th>Area after CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>0.109</td>
<td>18</td>
<td>0.073</td>
<td>16</td>
<td>0.069</td>
<td>12</td>
<td>0.083</td>
</tr>
<tr>
<td>B&amp;N</td>
<td>0.102</td>
<td>26</td>
<td>0.063</td>
<td>36</td>
<td>0.057</td>
<td>14</td>
<td>0.069</td>
</tr>
<tr>
<td>CDNow</td>
<td>0.135</td>
<td>52</td>
<td>0.048</td>
<td>86</td>
<td>0.011</td>
<td>48</td>
<td>0.047</td>
</tr>
<tr>
<td>Expedia</td>
<td>0.093</td>
<td>36</td>
<td>0.047</td>
<td>22</td>
<td>0.038</td>
<td>22</td>
<td>0.05</td>
</tr>
<tr>
<td>Travelocity</td>
<td>0.055</td>
<td>62</td>
<td>0.012</td>
<td>66</td>
<td>0.01</td>
<td>24</td>
<td>0.041</td>
</tr>
<tr>
<td>BMG</td>
<td>0.035</td>
<td>62</td>
<td>0.009</td>
<td>60</td>
<td>0.011</td>
<td>32</td>
<td>0.016</td>
</tr>
<tr>
<td>BUY</td>
<td>0.065</td>
<td>30</td>
<td>0.024</td>
<td>12</td>
<td>0.03</td>
<td>14</td>
<td>0.043</td>
</tr>
<tr>
<td>QVC</td>
<td>0.044</td>
<td>34</td>
<td>0.017</td>
<td>50</td>
<td>0.016</td>
<td>32</td>
<td>0.024</td>
</tr>
<tr>
<td>Priceline</td>
<td>0.125</td>
<td>36</td>
<td>0.086</td>
<td>40</td>
<td>0.094</td>
<td>18</td>
<td>0.082</td>
</tr>
<tr>
<td>Etoys</td>
<td>0.087</td>
<td>90</td>
<td>0.005</td>
<td>82</td>
<td>0.007</td>
<td>78</td>
<td>0.016</td>
</tr>
<tr>
<td>Iris1</td>
<td>0.92</td>
<td>28</td>
<td>0.69</td>
<td>34</td>
<td>0.49</td>
<td>18</td>
<td>0.75</td>
</tr>
<tr>
<td>Iris2</td>
<td>1.00</td>
<td>26</td>
<td>0.693</td>
<td>18</td>
<td>0.776</td>
<td>24</td>
<td>0.725</td>
</tr>
<tr>
<td>Cancer1</td>
<td>0.12</td>
<td>48</td>
<td>0.06</td>
<td>68</td>
<td>0.03</td>
<td>30</td>
<td>0.09</td>
</tr>
<tr>
<td>Cancer2</td>
<td>0.053</td>
<td>70</td>
<td>0.013</td>
<td>90</td>
<td>0.007</td>
<td>48</td>
<td>0.029</td>
</tr>
<tr>
<td>Liver1</td>
<td>0.06</td>
<td>52</td>
<td>0.02</td>
<td>42</td>
<td>0.02</td>
<td>36</td>
<td>0.03</td>
</tr>
<tr>
<td>Liver2</td>
<td>0.085</td>
<td>48</td>
<td>0.033</td>
<td>38</td>
<td>0.048</td>
<td>38</td>
<td>0.051</td>
</tr>
<tr>
<td>Pima1</td>
<td>0.08</td>
<td>14</td>
<td>0.05</td>
<td>58</td>
<td>0.03</td>
<td>14</td>
<td>0.06</td>
</tr>
<tr>
<td>Pima2</td>
<td>0.107</td>
<td>18</td>
<td>0.071</td>
<td>20</td>
<td>0.057</td>
<td>14</td>
<td>0.073</td>
</tr>
<tr>
<td>Pendigits1</td>
<td>0.34</td>
<td>32</td>
<td>0.11</td>
<td>68</td>
<td>0.01</td>
<td>18</td>
<td>0.23</td>
</tr>
<tr>
<td>Pendigits2</td>
<td>0.202</td>
<td>46</td>
<td>0.067</td>
<td>60</td>
<td>0.035</td>
<td>26</td>
<td>0.1</td>
</tr>
<tr>
<td>Average</td>
<td><strong>0.191</strong></td>
<td><strong>41.4</strong></td>
<td><strong>0.110</strong></td>
<td><strong>48.3</strong></td>
<td><strong>0.092</strong></td>
<td><strong>28.0</strong></td>
<td><strong>0.130</strong></td>
</tr>
<tr>
<td>average WEB</td>
<td><strong>0.085</strong></td>
<td><strong>44.6</strong></td>
<td><strong>0.038</strong></td>
<td><strong>47</strong></td>
<td><strong>0.034</strong></td>
<td><strong>29.4</strong></td>
<td><strong>0.047</strong></td>
</tr>
<tr>
<td>average UCI</td>
<td><strong>0.297</strong></td>
<td><strong>38.2</strong></td>
<td><strong>0.18</strong></td>
<td><strong>50</strong></td>
<td><strong>0.150</strong></td>
<td><strong>26.6</strong></td>
<td><strong>0.214</strong></td>
</tr>
</tbody>
</table>

![Figure 6.2: Performance on the Amazon dataset](image2)

In order to draw general conclusions, we compared metrics across all datasets. From each chart (dataset) we construct the following two metrics for the random, greedy and DODA-Log methods:

1. Critical Mass (percentage of global data needed to outperform the local model). Critical mass only reveals when the global model outperforms the local model, however, it does not show by how much. Hence we developed another metric – the area after critical mass, described next.

![Figure 6.3: Performance on the 20 datasets](image3)
2. Area after critical mass - the area of gains over the local model’s performance (in percentage) after the critical mass. In the chart this is the area under the line that represents the local model’s performance.

Figure 6.3 tabulates results from each dataset for each method. All the results are the averages of 5 runs, based on different random starting points of the initially available data records. As a reference point, we report the gain of the global model (built on the entire global data) over the local model (column 2) as the upper bound that each method can attain. The average critical mass across the 20 datasets considered, are 41.4%, 48.3% and 28.0% for Random, Greedy and DODA-Log respectively. In terms of the area after critical mass, the averages are 0.110, 0.092 and 0.130 for Random, Greedy and DODA-Log respectively.

In terms of comparative performance between the data acquisition methods, Figure 6.4 shows that based on a paired t-test across 20 datasets, the critical mass of the random method is not significantly different from the greedy method at a 0.05 level (p=0.096). However the critical mass of DODA-Log is significantly lower than that both of the random and greedy approaches (both p-values are less than 0.001). For the other metric (area after critical mass), DODA-Log significantly outperforms both random and greedy methods (Figure 6.5). There is no significant difference between the random and greedy approaches (p = 0.169).

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Greedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>0.096(1.74)</td>
<td></td>
</tr>
<tr>
<td>DODA-LOG</td>
<td>&lt;0.001 (-6.33)</td>
<td>&lt;0.001 (-5.01)</td>
</tr>
</tbody>
</table>

**Figure 6.4**: Significance of Critical Mass Comparisons

<table>
<thead>
<tr>
<th></th>
<th>Global</th>
<th>Random</th>
<th>Greedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>&lt;0.001(-4.51)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greedy</td>
<td>&lt;0.001(-4.01)</td>
<td>0.169(-1.43)</td>
<td></td>
</tr>
<tr>
<td>DODA-LOG</td>
<td>&lt;0.001(-4.19)</td>
<td>0.003(3.37)</td>
<td>0.031 (2.33)</td>
</tr>
</tbody>
</table>

**Figure 6.5**: Significance of Area after Critical Mass Comparisons

7. Discussion

The results above suggest that active learning can be a promising approach for selective customer information acquisition. While the experiments above demonstrated DODA-Log’s performance compared to alternative acquisition techniques they did not shed light on performance relative to some “optimal MSE” for the different datasets. In this section we first present experimental results that approximately measure how
close to optimal DODA-Log is. Section 7.2 discusses issues relating to why the method works, and Section 7.3 describes some key limitations and associated opportunities for future work.

7.1 Closeness to Optimality

While Section 6 compares DODA-Log with other possible data acquisition methods it does not provide a measure of how close to optimal the method is. However obtaining the optimal subset by exhaustive enumeration is prohibitive due to the combinatorics even for very small datasets. For example, selecting 10 records from a dataset with 100 records leads to \( \binom{100}{10} = 1.73 \times 10^{13} \) possible combinations (and models) that need to be examined. However what we can certainly do as a first order approximation is to find the best subset among \( Q \) different random samples where \( Q \) is a large number. In this case \( \min(MSE_1, MSE_2, \ldots, MSE_Q) \) can give us an approximate measure of the optimal MSE, where MSE\(_i\) represent the mean squared error of the global model built on \( i^{th} \) random subset acquired.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MSE on different percentages of acquired data</th>
<th>Average MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>Iris1</td>
<td>DODA-Log</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0</td>
</tr>
<tr>
<td>Iris2</td>
<td>DODA-Log</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0</td>
</tr>
<tr>
<td>Cancer1</td>
<td>DODA-Log</td>
<td>0.258</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.169</td>
</tr>
<tr>
<td>Cancer2</td>
<td>DODA-Log</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.167</td>
</tr>
<tr>
<td>Liver1</td>
<td>DODA-Log</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.445</td>
</tr>
<tr>
<td>Liver2</td>
<td>DODA-Log</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.445</td>
</tr>
<tr>
<td>Pima1</td>
<td>DODA-Log</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.396</td>
</tr>
<tr>
<td>Pima2</td>
<td>DODA-Log</td>
<td>0.452</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.393</td>
</tr>
<tr>
<td>Pendigits1</td>
<td>DODA-Log</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.076</td>
</tr>
<tr>
<td>Pendigits2</td>
<td>DODA-Log</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.088</td>
</tr>
<tr>
<td>Average</td>
<td>DODA-Log</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>Optimal(_{100})</td>
<td>0.218</td>
</tr>
</tbody>
</table>

Figure 7.1: Comparison of MSE between DODA-Log and Optimal\(_{100}\)
We computed this measure for $Q=1000$ for the 10 (smaller) UCI datasets. Further for each UCI dataset we computed this measure based on acquiring $K=10\%, K=20\%,..., K=100\%$ of the data. For each given $K$ we compare DODA-Log’s performance with $\min(MSE_1, MSE_2,...,MSE_{1000})$, which we refer to as Optimal$_{R,1000}$ in Figure 7.1. On average across all the runs DODA-Log is within 14% of Optimal$_{R,1000}$ (i.e. $0.252-0.217)/0.252$). Also from this analysis it is clear, as expected, that the gap narrows as more data is acquired. The MSE gap is much higher when the amount of data is below 30% than above (DODA-Log is within 9% of Optimal$_{R,1000}$ when acquiring 30% or more data). This corresponds well with our results in Figure 6.3 which suggest that the critical mass for the UCI datasets is close to 30%. While DODA-Log is significantly better than alternative data acquisition techniques, as these results suggest there is still room for improvement.

7.2 Understanding why and when selective data acquisition works

Why and when does selective data acquisition work? These questions have been posed in the literature but exact conditions are elusive. Mackay (1992), Greiner and Grove (2002) and Hasenjäger & Ritter (1999) suggest that active learning works because it eliminates redundancy among data and consequently active learning is best applied to the data that contains redundancy. According to Hasenjäger & Ritter (1999), not all data are equally valuable for model building. After the model has seen enough data, the utility of “seeing” additional data diminishes. Active learning only selects the most informative data and thus a redundant datum will be of less utility to the model. This is best reflected in the score function of our algorithm DODA-Log. The $X'X$ component weighs down redundant data points and thus a similar record has a lesser chance to be chosen (e.g. two identical points will make $|X'X|$ to be 0).

However is DODA-Log (and more generally, any active learning method) simply a redundancy-reduction approach? Below we attempt to answer the question experimentally for our approach. We first constructed a purely redundancy reduction method in the following manner. Suppose at a certain data acquisition stage, $K$ data points have been acquired and we aim to acquire the $(K+1)^{th}$ one from the remaining $(N−K)$ records. A redundancy reduction method is to acquire the $(K+1)^{th}$ data point that is furthest from the existing $K$ points. For this method, we only compute the distance based on the local data.
Figure 7.2: Performance of a purely redundancy reduction method

Figure 7.2 shows that this redundancy reduction method is better than random (which does not eliminate any redundancy) and is worse than DODA-Log (see Figure 6.3). This suggests that DODA-Log is more than just a redundancy reduction method. Our dual-objective method takes into account both objectives: building a good model for the target variable as well as a good imputation model. First, with respect to the objective of building a good model, the score function \( |\hat{Y}(1 - \hat{Y})X^T X| \) for logistic regression suggests that we want to not only reduce redundancy (through the component \(X^T X\)), but at the same time to also acquire those points that the model currently is uncertain of (through the component \(\hat{Y}(1 - \hat{Y})\)). Second, as we argue earlier in this paper, imputation is a critical component for the data acquisition scenarios considered here. Neither the random method nor the greedy method recognizes the importance of imputation. DODA-Log takes into account all these factors: redundancy reduction, model building and imputation, all of which contribute to the improved performance.

### 7.3 Limitations and Future Extensions

There are a several issues not addressed in this paper that need to be studied in future work, and we describe five specific opportunities below:

1. Most selective acquisition methods, including DODA, are not immune to outliers. For example, many active learning methods (Cohn et al. 1996, Mackay 1992, Freund et al. 1997, Saar-Tsechansky and Provost 2002) select data that the current model is least certain about. For noisy data, it might well be the case that these methods favor outliers. In view of this, Freund et al. (1997) argue that their QBC method is best applied when the data are noiseless, and DODA may similarly perform better in less noisy data. The good performance on out-of-sample data suggests that outliers do not significantly affect our model, but the effect of noise on performance needs to be studied more carefully.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Iris1</th>
<th>Iris2</th>
<th>Cancer1</th>
<th>Cancer2</th>
<th>Liver1</th>
<th>Liver2</th>
<th>Pima1</th>
<th>Pima2</th>
<th>Pendigits1</th>
<th>Pendigits2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Mass % (CM)</td>
<td>12</td>
<td>24</td>
<td>46</td>
<td>74</td>
<td>76</td>
<td>20</td>
<td>18</td>
<td>12</td>
<td>30</td>
<td>50</td>
<td><strong>36.2</strong></td>
</tr>
<tr>
<td>Area after CM</td>
<td>0.83</td>
<td>0.72</td>
<td>0.073</td>
<td>0.002</td>
<td>0.004</td>
<td>0.038</td>
<td>0.053</td>
<td>0.067</td>
<td>0.163</td>
<td>0.03</td>
<td><strong>0.198</strong></td>
</tr>
</tbody>
</table>
2. We do not fully consider various levels of aggregation at which data may be available. In practice customer data may be offered by a vendor in “blocks”. This will be an interesting and relevant extension.

3. There is yet another level at which we can study selectivity. We consider the case where global data for a single record is acquired either completely or is not acquired. In this sense our acquisition algorithms are “selective” with respect to customers, but not with respect to features. Being selective on features could be another interesting, and practical, extension. A recent article by Perlich and Provost (2004) partially addresses this problem by means of relation aggregation where they study how to choose features.

4. In Section 2, we distinguished between two kinds of information acquisition – one that occurs during the course of model building, and the other that occurs during the course of problem solving. We did not integrate these two approaches in this paper for the following reasons. The model used in this paper is logistic regression, which requires all the (explanatory) variables before it can be used to provide a prediction, and the model also does not allow for obtaining the values of variables one at a time. A natural solution is to build \textit{DODA-Tree} (like \textit{DODA-Log}), which uses sequential acquisition to build a decision tree model, which is fundamentally an approach that lends itself well to sequential information acquisition during problem solving. There are active learning methods that use decision trees as the underlying model, and in future work we will work on developing an instantiation of DODA for this case. However, since decision tree is a non-parametric method, it is not obvious how a closed-form score function (like Score-Log) can be derived. Without such a score function, one has to resort to either 1) a heuristic approach such as acquiring data which the decision tree cannot predict well or 2) a computationally expensive approach that selects data by directly building a tree and evaluating how good the tree is with the data. Both approaches have their drawbacks and need further investigation.

5. In this paper we focused on building a good global model, but did not make use of the local data to make the final prediction. There are two reasons for this (i) Indeed building a good global model is a good end in itself. Our approach here focuses on how to best build such a global model. (ii) It is not clear how best to combine local data (on N-K points) with global data (on K points). However using all the available data to
make predictions can improve the quality of the predictions made. Below we outline two approaches to doing this and discuss the inherent challenges.

The first approach is to treat the global variables not acquired as missing data. We can first impute these missing values using models such as KNN, MI (Rubin and Little 1987, Schafer and Olsen 1998), and then build a training model on all the records. This is a promising approach when the percentage of missing data is small, and when good imputation methods are available. However for problems when only a small portion of the data is acquired (as common in active learning) it is not clear how good the results will be since the amount of data that has to be imputed is large compared to the amount of data available. Further compounding the problem, specific to the scenarios considered in our paper the number of unknown global variables can be large compared to the number of known local variables. For example the Web usage datasets we used contain 25 global variables and 15 local variables. Our preliminary experimental results testing such a combined approach yielded mixed results, with the combined approach being worse in some cases and not significantly better in others.

Another approach is to build separate global and local models, and to then combine the two model predictions using a weighted scheme. This approach makes use of all available data by means of model combination and has the advantage that imputation is not required. Model combination is a well-studied topic in data mining and numerous methods have been proposed in the literature (see Dietterich 2000 and Zheng & Padmanabhan 2005 for detailed references). The setup in our paper is, in general, equivalent to the model combination approach for the following reason. In our setup the performance of the local model is fixed since the local data (on all N customers) is fixed. Hence better global models should yield better combined models, and this paper addresses the issue of building good global models. However one extension to this that needs to be investigated may be to modify the selection procedure given that the combined model is used in this manner. This may then select different data points during the iterative process and is certainly a promising extension that can be looked at.
8. Conclusions

In this paper, we presented a new problem of selective data acquisition, one motivated by practical decision contexts faced by managers. Toward this end, we presented a general method and a specific algorithm for selective data acquisition. The generic method is based on optimizing a dual-objective that takes into account both a classifier and an imputation model. Specific to the logistic regression we develop a Bayesian score function that minimizes the posterior variance of the parameters. The experimental results show that the methods perform well and indicate that selective data acquisition can be effective. Our main contributions here include presenting a new problem and developing new generic and specific data acquisition algorithms, based on extant work in OED and machine learning.

In the introduction we described Web browsing and credit card applications where there are natural situations in which firms do not have complete information and would benefit significantly by acquiring additional customer data. While these applications form a “top-down” motivation for this research, there are several recent trends, that relate to organizations actively seeking to acquire more complete customer information, that provide a “bottom-up” motivation. There is a trend toward acquiring complete information by integrating databases and conducting surveys. At the core of Customer Relationship Management is the idea of understanding customers well by integrating customer data across various “touch-points”. There are surveys on the Web that are becoming more common. Bizrate seeks customer information using online surveys after a customer purchases a product. Dell conducts surveys, when customers exit the site, to learn what users were looking for and whether they were satisfied. In the US, at the heart of security considerations today is a proposal of linking user data across multiple resources to gain total information awareness of individuals. Our work is related to this trend in that it addresses mechanisms that can help understand customers better by selective information acquisition.

More generally in this paper we presented a novel data acquisition problem, developed active learning algorithms and reported empirical results. The problem studied is inherently interdisciplinary and

---

approaches from machine learning, OED, more general Bayesian approaches and optimization-based techniques can be used. This is a rich domain in that there are several interesting extensions and much future work is needed to study the various approaches and explore the relevant extensions adequately.

Reference


Kuhfeld, W., R. Tobias and M. Garratt. 1994. Efficient experimental design with marketing research applications. *Journal of Marketing Research*. **31** (11), 545-557


Mookerjee, V., B. Dos Santos. 1993. Inductive expert system design: maximizing system value. *Information System Research*. **4** (2), 111-131


Appendix A: Derivation of Proposition 1

Below we develop a score function to determine the value of data with the goal of minimizing the variance of parameter estimation. Logistic regression can be represented as:

\[
\log \frac{\hat{Y}}{1-\hat{Y}} = \hat{\beta}'X
\]  

(8)

In (8) we use \(\hat{Y}\) to represent the logit probability, \(X\) is the independent variable, \(\beta\) is the coefficient. Assume the prior, i.e. the estimated coefficients from the current available data \((X_0, Y_0)\), follows a normal distribution\(^7\) \(\beta \sim N(\beta_0, \Sigma_0)\) where \(\beta_0\) is the coefficients estimated from current data and \(\Sigma_0\) is the variance-covariance matrix. Let \(\sigma\) be the variance of the data. The asymptotic variance-covariance matrix \(\Sigma\) of \(\beta\) of the logistic regression is

\[
\Sigma = [\hat{Y}'(1-\hat{Y})X'X]^{-1}\sigma^2
\]  

(9)

Thus asymptotically \(\Sigma_0\) can be expressed as \(\Sigma_0 = [\hat{Y}'_0(1-\hat{Y}_0)X'_0X_0]^{-1}\sigma_0^2\). Still assume that parameter estimation for the new data \((X, Y)\) follow a normal distribution \(N(\beta, \Sigma)\). Then the posterior distribution would be a normal distribution also. From Greene (1997) and Box & Tiao (1992), we can update the prior as follows:

\[
E(\beta / \sigma^2, X, Y) = [\Sigma_0^{-1} + \sigma^{-2}\hat{Y}'(1-\hat{Y})X'X]^{-1}[\Sigma_0^{-1}\beta_0 + \sigma^{-2}\hat{Y}'(1-\hat{Y})X'X\hat{\beta}]
= F\beta_0 + (1-F)\hat{\beta}
\]

(10)

where

\[
F = [\Sigma_0^{-1} + \sigma^{-2}\hat{Y}'(1-\hat{Y})X'X]^{-1}\Sigma_0^{-1}
= [(prior\ variance)^{-1} + (conditional\ variance)^{-1}]^{-1}[prior\ variance]^{-1}
\]

and

\[
Var(\beta / \sigma^2, X, Y) = [\Sigma_0^{-1} + [\sigma^{-2}(\hat{Y}'(1-\hat{Y})X'X)^{-1}]^{-1}]
= [\hat{Y}'_0(1-\hat{Y}_0)X'_0X_0\sigma_0^{-2} + \hat{Y}'(1-\hat{Y})X'X\sigma^{-2}]^{-1}
\]

(11)

\(^7\) Since for non-linear models, closed-form Bayesian optimal criteria do not always exist, approximations typically are used (Chaloner and Verdinelli 1995). Most approximations suggested in the OED literature involve using a normal approximation to the posterior distribution.
Based on (11), if the goal is to minimize the posterior variance of the parameters, one should maximize 
\[ \hat{\text{Y}}_0' (1 - \hat{\text{Y}}_0) X_0' X_0 \sigma^{-2} + \hat{\text{Y}}' (1 - \hat{\text{Y}}) X' X \sigma^{-2}, \]
or more conveniently maximizing the determinant of this matrix (Atkinson and Donev 1992). For simplicity we assume that the model behaves well: for the old data and new acquired data, the error terms follow the same distribution \( N(0, \sigma) \), i.e., \( \sigma_0 = \sigma \). Then we can simplify the Bayesian criterion for data selection by dropping \( \sigma \). Thus for a sequential data acquisition procedure, suppose there are \( K \) existing global data points on which the global model was built, the \((k+1)\)th data point should be chosen such that the determinant \( | \hat{\text{Y}}_K' (1 - \hat{\text{Y}}_K) X_K' X_K + \hat{\text{y}}_{k+1}' (1 - \hat{\text{y}}_{k+1}) x_{k+1}' x_{k+1} | \) is maximized (we use the lower case notation \((x, y)\) to represent a single data point). Equivalently, this determinant is the score for each unknown data point.

Further, assuming that after acquiring a single data point the parameter estimates remain approximately the same (Cohn 1996), we can predict \( \hat{\text{y}}_{k+1} \) of the \((k+1)\)th data point using \( \hat{\beta}_k \), the estimated coefficients derived from existing \( K \) available data points. This approximation technique is necessary and often used in OED (Chaloner and Verdinelli 1995) in order to reduce computation complexity: rather than rebuilding a logit model incorporating this \((K+1)\)th point, we only need to compute this criterion. Based on this approximation, the criterion can be further compressed as follows:

\[
\hat{\text{Y}}_K' (1 - \hat{\text{Y}}_K) X_K' X_K + \hat{\text{y}}_{k+1}' (1 - \hat{\text{y}}_{k+1}) x_{k+1}' x_{k+1} = \begin{bmatrix} \hat{\text{Y}}_K' (1 - \hat{\text{Y}}_K) \\ \hat{\text{y}}_{k+1}' (1 - \hat{\text{y}}_{k+1}) \end{bmatrix} \begin{bmatrix} X_K' X_K \\ x_{k+1}' x_{k+1} \end{bmatrix} \]

\[
= [\hat{\text{Y}}_{K+1}' (1 - \hat{\text{Y}}_{K+1})] [X_{K+1}' X_{K+1}] \tag{12}
\]

In (12), \( X_{K+1} \) represents a \((K+1) \times P\) matrix and \( \hat{\text{Y}}_{K+1} \) represents a \((K+1) \times 1\) vector where all the \( Y \) values are predicted using \( \hat{\beta}_k \). Dropping the subscripts of equation (E) for simplicity and denote \( \text{Score}_{k+1} = | \hat{\text{Y}}' (1 - \hat{\text{Y}}) X' X | \), where \( X \) is a \((k+1) \times P\) matrix and \( \hat{\text{Y}} \) is a \((k+1) \times 1\) vector. Then the Bayesian score function can be simplified as

\[
\text{Score}_{k+1} = | \hat{\text{Y}}' (1 - \hat{\text{Y}}) X' X | \tag{13}
\]
1. Results on “Strangers”

For strangers, Figure 1 and Figure 2 below show that DODA-Log performs better than both random and the greedy acquisition methods over the two same two datasets (Pendigits1 and Amazon). However, the critical mass of DODA-Log goes up to 30% (18% for friends) for Pendigits and 22% for Amazon (12% for friends).

Figure 1: Performance on Strangers – Pendigits1  
Figure 2: Performance on Strangers – Amazon

Figure 3 tabulates results from each dataset for each method for the “stranger” scenario. Here the results are the averages of 3 runs, based on different random starting point of the initially available data records. The results in Figure 3 show that a fairly large portion of the global data is needed to outperform local models. The average critical masses across the 20 datasets are 60.3%, 76.0 and 49.9% for random, greedy and DODA-Log respectively. For two Web datasets, Travelocity and Etoys, even acquiring 100% the training data is not enough to outperform the local models. This clearly indicates that the “strangers” scenario is significantly harder than the “friends” scenario because we need to impute the missing global values in the out-of-sample data for strangers. Note that the number of missing variables needed to be imputed for the Web datasets is 25 - from just 15 local variables. This makes the problem challenging for imputation methods and the average critical mass for the 10 Web datasets is 55.6%. On the other hand, for UCI datasets, only one third of the variables need to be imputed. And in those cases, DODA-Log performs significantly better (average critical mass is 44.2%).
<table>
<thead>
<tr>
<th>DataSet</th>
<th>Global Area over local Critical Mass % (CM)</th>
<th>Random Area after CM Critical Mass % (CM)</th>
<th>Greedy Area after CM Critical Mass % (CM)</th>
<th>DODA-LOG Area after CM Critical Mass % (CM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>0.082</td>
<td>0.051</td>
<td>0.036</td>
<td>0.055</td>
</tr>
<tr>
<td>B&amp;N</td>
<td>0.084</td>
<td>0.033</td>
<td>0.051</td>
<td>0.045</td>
</tr>
<tr>
<td>CDNow</td>
<td>0.12</td>
<td>0.038</td>
<td>0.015</td>
<td>0.057</td>
</tr>
<tr>
<td>Expedia</td>
<td>0.019</td>
<td>0.004</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>Travelocity</td>
<td>-0.012</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BMG</td>
<td>0.01</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>BUY</td>
<td>0.021</td>
<td>0.009</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>QVC</td>
<td>0.014</td>
<td>0.009</td>
<td>0.002</td>
<td>0.008</td>
</tr>
<tr>
<td>Priceline</td>
<td>0.069</td>
<td>0.028</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>Etoys</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iris1</td>
<td>0.63</td>
<td>0.35</td>
<td>0.14</td>
<td>0.39</td>
</tr>
<tr>
<td>Iris2</td>
<td>0.417</td>
<td>0.068</td>
<td>0.102</td>
<td>0.243</td>
</tr>
<tr>
<td>Cancer1</td>
<td>0.06</td>
<td>0.027</td>
<td>0.002</td>
<td>0.034</td>
</tr>
<tr>
<td>Cancer2</td>
<td>0.02</td>
<td>0.004</td>
<td>0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>Liver1</td>
<td>0.006</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>Liver2</td>
<td>0.016</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Pima1</td>
<td>0.079</td>
<td>0.065</td>
<td>0.028</td>
<td>0.065</td>
</tr>
<tr>
<td>Pima2</td>
<td>0.021</td>
<td>0.01</td>
<td>0.004</td>
<td>0.015</td>
</tr>
<tr>
<td>Pendigits1</td>
<td>0.094</td>
<td>0.028</td>
<td>0.014</td>
<td>0.037</td>
</tr>
<tr>
<td>Pendigits2</td>
<td>0.023</td>
<td>0.002</td>
<td>0.001</td>
<td>0.06</td>
</tr>
<tr>
<td>Average</td>
<td>0.089</td>
<td>0.037</td>
<td>0.021</td>
<td>0.053</td>
</tr>
<tr>
<td>average WEB</td>
<td>0.042</td>
<td>0.018</td>
<td>0.011</td>
<td>0.020</td>
</tr>
<tr>
<td>average UCI</td>
<td>0.137</td>
<td>0.06</td>
<td>0.030</td>
<td>0.086</td>
</tr>
</tbody>
</table>

**Figure 3:** Comparative Results on “Strangers” over 20 Datasets

The comparative performance in terms of critical mass (Figure 4) show that **DODA-Log > Random > Greedy** and all the pairwise difference is significant. In terms of area after critical mass, **DODA-Log significantly outperforms the greedy approach (p = 0.029).** The difference between **DODA-Log and random, though, is not significant (p = 0.097) at a 0.05 significance level.**

**Figure 4:** Significance of Critical Mass Comparisons Across Three Methods

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Greedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>0.001 (3.84)</td>
<td></td>
</tr>
<tr>
<td>DODA-LOG</td>
<td>0.004(-3.26)</td>
<td>&lt;0.001 (-6.06)</td>
</tr>
</tbody>
</table>

**Figure 5:** Significance of Area after Critical Mass Comparisons Across Three Methods

To summarize, there are two findings from the above experiments. The main result is that the dual-objective method **DODA-Log significantly outperforms the other two methods (random and greedy) both for “friends” and “strangers”, and appears to be a promising selective data acquisition technique.** The second finding, not surprisingly, is that all three data acquisition methods perform much worse in the “strangers” scenario than in
the “friends” scenario. Clearly doing well for strangers is a harder problem and more research is needed to study how the performance for this case can be improved.

2. Complexity and Consistency of Score-Log

**Proposition 2:** The computation complexity of Score-Log is $O(m^3N)$, where $m$ is the number of points to be acquired and $N$ is the number of records in the dataset.

Proof Sketch: Score-Log needs to compute $|\hat{Y}'(1-\hat{Y})X'X|$ at each acquisition phase. Suppose there are $N$ customers in total. Assume that the number of variables, $P$, is relatively small as compared to $N$. Then we can ignore the computation cost associated with computing the determinant of the $P\times P$ matrix. For a sequential procedure, at phase $K$, we need to evaluate Score-Log for $N-K$ candidate customers. For each candidate, we need to compute the matrix $\hat{Y}'(1-\hat{Y})X'X$, the time complexity of which is $O(K^2)$. Thus the overall time complexity of the procedure of acquiring all $N$ customers sequentially would be $\sum_{K=1}^{m} (N-K)\times K^2$, or $O(m^3N)$. □

One downside of active learning approaches is that often they are computationally expensive. In the above case this cost is manageable if $m \ll N$, which is often the case.

In general it is unclear whether parameter estimation for non-linear models following a sequential design remains consistent$^8$ (Wu 1985, Hu 1998). As pointed out by Rosenberger & Hu (2002), sequential designs may induce dependence among the data, and the covariance structure is often complex and intractable. Consequently, it is not always clear that maximum likelihood estimators will have the usual property of asymptotic normality, allowing for the usual standard errors and tests. This problem was first recognized by Ford & Silvey (1980). Despite the popularity of sequential designs in practice, no general solution has been found in the OED field (Atkinson and Bailey 2001). Some special conditions have been

---

$^8$ Consistency is defined as parameter estimator $\hat{\beta}$ tends to the true parameter $\beta$ with probability 1.
proposed under which the estimates remain consistent in Wu (1985), Hu (1998) and Rosenberger & Hu (2002). These conditions require the Martingale property (Wu 1985) of the sequential design. Denote $\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_i$ the sequence of error of the model built following a sequential design and let $\delta^2$ be the model variance. A sequential acquisition procedure is said to be a martingale procedure if it satisfies the following two assumptions:

\begin{align*}
E(\varepsilon_i / \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1) &= 0 \quad \text{and} \\
E(\varepsilon_i^2 / \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1) &= \delta^2 < \infty
\end{align*}

Equation (14) entails that the expected error $\varepsilon_i$ at stage $i$ given a sequence of errors $\varepsilon_1, \ldots, \varepsilon_{i-1}$ at earlier stages should be 0. Equation (15) states that the expected model variance at stage $i$ is bounded given the sequence of errors $\varepsilon_1, \ldots, \varepsilon_{i-1}$. Hu (1998) shows that sequential design of the generalized linear model (GLM) has the Martingale property as follows. Let $(X_i, Y_i)$ be the pair of observed vectors of independent and dependent variables at acquisition stage $i$. A GLM with a link function $\mu$ specifies the relationship between $y$ and $x$ through $E(y_i / \beta, y_1, \ldots, y_{i-1}) = \mu(x_i\beta)$. For the logistic regression, $\mu$ is the logit function. Then the error can be defined as

$$
\varepsilon_i = y_i - \mu(x_i\beta) = y_i - E(y_i / \beta, y_1, \ldots, y_{i-1})
$$

(16)

It is easy to verify that equation (16) satisfies the two Martingale assumptions specified in (14) and (15). First, $E(\varepsilon_i / \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1) = E(y_i) - E(E(y_i / \beta, y_1, \ldots, y_{i-1})) = 0$. Second we denote $V(\varepsilon_i)$ as the variance of $\varepsilon_i$ and let $V(y_i) = \delta < \infty$ be the variance of $y_i$. Then the expected variance of $\varepsilon_i$ can be expressed as $E(\varepsilon_i^2 / \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1) = V(\varepsilon_i) = V(y_i) = \delta^2$. Thus we arrive at proposition 3.

**Proposition 3** Parameter estimators of the logistic regression are consistent following the acquisition procedure of Score-Log.