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, Riverside,

Balaji Padmanabhan
The Wharton School, University of Pennsylvania

Electronic Companion Pages (Online Appendix)

1. 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:

$$Log\frac{\hat{Y}}{1-\hat{Y}} = \hat{\beta}' X \tag{8}$$

In (8) we use $\hat{Y}$ to represent the logit probability, X is the independent variable, β is the coefficient. Assume the prior, i.e. the estimated coefficients from the current available data (X_0, Y_0), follows a normal distribution[1] $\beta \sim N(\beta_0, \Sigma_0)$ where β_0 is the coefficients estimated from current data and Σ_0 is the variance-covariance matrix. Let σ be the variance of the data. The asymptotic variance-covariance matrix Σ of β of the logistic regression is

$$\Sigma = [\hat{Y}'(1-\hat{Y})X'X]^{-1}\sigma^2 \tag{9}$$

Thus asymptotically Σ_0 can be expressed as $\Sigma_0 = [\hat{Y}_0'(1-\hat{Y}_0)X_0'X_0]^{-1}\sigma_0^2$. Still assume that parameter estimation for the new data (X, Y) follow a normal distribution $N(\beta, \Sigma)$. Then the posterior distribution

[1] 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.

would be a normal distribution also. From Greene (1997) and Box & Tiao (1992), we can update the prior as follows:

$$E(\beta/\sigma^2, X, Y) = [\Sigma_0^{-1} + \sigma^{-2}\hat{Y}'(1-\hat{Y})X'X]^{-1}[\Sigma_0^{-1}\beta_0 + \sigma^{-2}\hat{Y}'(1-\hat{Y})X'X\hat{\beta}] \\ = F\beta_0 + (1-F)\hat{\beta} \tag{10}$$

where

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

and

$$Var(\beta/\sigma^2, X, Y) = \{\Sigma_0^{-1} + [\sigma^2((\hat{Y}'(1-\hat{Y})X'X)^{-1}]^{-1}\}^{-1} \\ = [\hat{Y}_0'(1-\hat{Y}_0)X_0'X_0\sigma_0^{-2} + \hat{Y}'(1-\hat{Y})X'X\sigma^{-2}]^{-1} \tag{11}$$

Based on (11), if the goal is to minimize the posterior variance of the parameters, one should maximize $[\hat{Y}_0'(1-\hat{Y}_0)X_0'X_0\sigma_0^{-2} + \hat{Y}'(1-\hat{Y})X'X\sigma^{-2}]$, 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 σ. 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 $|\hat{Y}_K'(1-\hat{Y}_K)X_K'X_K + \hat{y}_{k+1}'(1-\hat{y}_{k+1})x_{k+1}'x_{k+1}|$ 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 $\hat{y}_{k+1}$ 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:

$$\hat{Y}_K{}'(1-\hat{Y}_K)X_K{}'X_K + \hat{y}_{k+1}'(1-\hat{y}_{k+1})x_{k+1}'x_{k+1} = \begin{bmatrix} \hat{Y}_k{}'(1-\hat{Y}_k) \\ \hat{y}'_{k+1}(1-\hat{y}_{k+1}) \end{bmatrix} \times \begin{bmatrix} X'_k X_k \\ x'_{k+1} x_{k+1} \end{bmatrix} \quad (12)$$

$$= [\hat{Y}'_{K+1}(1-\hat{Y}_{K+1})][X'_{K+1}X_{K+1}]$$

In (12), X_{K+1} represents a $(K+1) \times P$ matrix and $\hat{Y}_{K+1}$ 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 $Score_{k+1} = |\hat{Y}'(1-\hat{Y})X'X|$, where X is a $(k+1) \times P$ matrix and $\hat{Y}$ is a $(k+1) \times 1$ vector. Then the Bayesian score function can be simplified as

$$Score_{k+1} = |\hat{Y}'(1-\hat{Y})X'X| \quad (13)$$

2. Results on "Strangers"

For strangers, Figure 6.1 and Figure 6.2 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 6.1: Performance on Strangers – Pendigits1

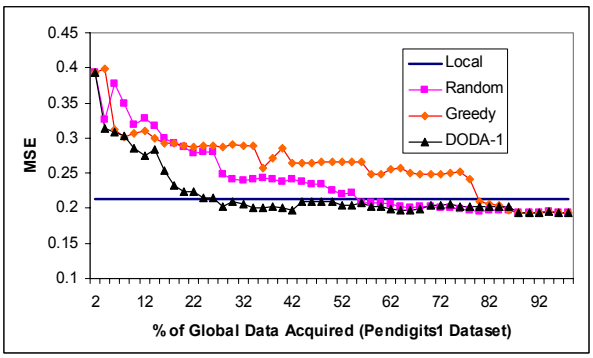

Figure 6.2: Performance on Amazon

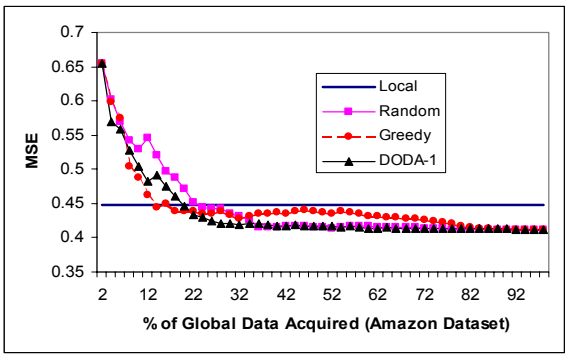

Table 6 tabulates results from each dataset for each method for the "stranger" scenario. Here the results are the averages of 3 runs, based on different random starting point of the initially available data records. The results in Table 6 show that a fairly large portion of the global data is needed 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 6 Comparative Results on "Strangers" over 20 Datasets

	Global	**Random**		**Greedy**		**DODA-LOG**	
DataSet	Area over local	Critical Mass % (CM)	Area after CM	Critical Mass % (CM)	Area after CM	Critical Mass % (CM)	Area after CM
Amazon	0.082	26	0.051	20	0.036	22	0.055
B&N	0.084	30	0.033	38	0.051	18	0.054
CDNow	0.12	54	0.038	90	0.015	50	0.057
Expedia	0.019	54	0.004	76	0.001	44	0.006
Travelocity	-0.012	100	0	100	0	100	0
BMG	0.01	58	0.004	60	0.005	48	0.006
BUY	0.021	48	0.009	76	0.003	58	0.005
QVC	0.014	40	0.009	88	0.002	40	0.008
Priceline	0.069	58	0.028	94	0.001	76	0.004
Etoys	0	100	0	100	0	100	0
Iris1	0.63	44	0.35	78	0.14	40	0.39
Iris2	0.417	78	0.068	68	0.102	62	0.243
Cancer1	0.06	68	0.027	92	0.002	48	0.034
Cancer2	0.02	78	0.004	96	0.001	50	0.01
Liver1	0.006	96	0.002	96	0.002	78	0.003
Liver2	0.016	76	0.004	58	0.005	56	0.005
Pima1	0.079	18	0.065	56	0.028	16	0.065
Pima2	0.021	42	0.01	58	0.004	26	0.015
Pendigits1	0.094	56	0.028	82	0.014	30	0.037
Pendigits2	0.023	82	0.002	94	0.001	36	0.06
Average	**0.089**	**60.3**	**0.037**	**76.0**	**0.021**	**49.9**	**0.053**
average_WEB	0.042	56.8	0.018	74.2	0.011	55.6	0.020
average_UCI	0.137	63.8	0.06	78	0.030	44.2	0.086

The comparative performance in terms of critical mass (Table 7) show that *DODA-Log* > Random > Greedy and all the pairwise difference is significant. In terms of area after critical mass (Table 8), *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.

Table 7 Significance of Critical Mass Comparisons Across Three Methods

	Random	Greedy
Greedy	0.001 (3.84)	
DODA-LOG	0.004(-3.26)	<0.001 (-6.06)

Table 8 Significance of Area after Critical Mass Comparisons Across Three Methods

	Global	Random	Greedy
Random	0.021 (-2.52)		
Greedy	0.020 (-2.53)	0.151 (-1.50)	
DODA-LOG	0.020 (-2.55)	0.097 (1.75)	0.029 (2.36)

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.

3. 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 and assume that the number of variables P is small 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 $\hat{Y}'(1-\hat{Y})X'X$, 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 $m << N$, which is often the case.

In general it is unclear whether parameter estimation for non-linear models following a sequential design remains consistent[2] (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 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 $\varepsilon_1,...,\varepsilon_{i-1},\varepsilon_i$ the sequence of error of the model built following a sequential design and let δ^2 be the model variance. A sequential acquisition procedure is said to be a martingale procedure if it satisfies the following two assumptions:

$$E(\varepsilon_i / \varepsilon_{i-1},\varepsilon_{i-2},...,\varepsilon_1) = 0 \text{ and} \qquad (14)$$

$$E(\varepsilon^2{}_i / \varepsilon_{i-1},\varepsilon_{i-2},...,\varepsilon_1) = \delta^2 < \infty \qquad (15)$$

Equation (14) entails that the expected error ε_i at stage i given a sequence of errors $\varepsilon_1,...,\varepsilon_{i-1}$ at earlier stages should be 0. Equation (15) states that the expected model variance at stage i is bounded given the sequence of errors $\varepsilon_1,...,\varepsilon_{i-1}$. 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 μ specifies the relationship between y and x through $E(y_i / \beta, y_1,..., y_{i-1}) = \mu(x_i\beta)$. For the logistic regression, μ is the logit function. Then the error can be defined as

$$\varepsilon_i = y_i - \mu(x_i\beta) = y_i - E(y_i / \beta, y_1,..., y_{i-1}) \qquad (16)$$

[2] Consistency is defined as parameter estimator $\hat{\beta}$ tends to the true parameter β with probability 1.

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

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