On Active Learning for Data Acquisition

Zhiqiang Zheng and Balaji Padmanabhan
Operations and Information Management,
The Wharton School, University of Pennsylvania
{zhengzhi, balaji}@wharton.upenn.edu

Abstract

Many applications are characterized by having naturally incomplete data on customers – where data on only some fixed set of local variables is gathered. However, having a more complete picture can help build better models. The naïve solution to this problem – acquiring complete data for all customers – is often impractical due to the costs of doing so. A possible alternative is to acquire complete data for "some" customers and to use this to improve the models built. The data acquisition problem is determining how many, and which, customers to acquire additional data from. In this paper we suggest using active learning based approaches for the data acquisition problem. In particular, we present initial methods for data acquisition and evaluate these methods experimentally on web usage data and UCI datasets. Results show that the methods perform well and indicate that active learning based methods for data acquisition can be a promising area for data mining research.

1. Introduction

Many data mining applications are characterized by the collection of naturally incomplete data in which the application only has data on some fixed set of "local" variables due to reasons such as data ownership, business issues and technological issues. Credit card companies have data on customer transactions with their cards, but do not have data on customer transactions with other cards. There are examples in the online world too where the inherent incompleteness of collected data shows up. For example, consider two users who browse the web for air tickets. Assume that the first user's session is as follows $Cheaptickets_1$, $Cheaptickets_2$, $Travelocity_1$, $Travelocity_2$, $Expedia_1$, $Expedia_2$, $Travelocity_3$, $Travelocity_4$, $Expedia_3$, $Cheaptickets_3$ where X_i represents some page i, at website X and in this session assume that the user purchases a ticket at Cheaptickets. Assume that the second user's session is $Expedia_1$, $Expedia_2$, $Expedia_3$, $Expedia_4$ and that this user purchases a ticket at Expedia (in the booking page $Expedia_4$, in particular). Expedia's local data would include the following:

User1: $Expedia_1$, $Expedia_2$, $Expedia_3$
User2: $Expedia_1$, $Expedia_2$, $Expedia_3$,, $Expedia_4$

In one case (user 2) the first three pages result in the user booking a ticket at the next page. In the other case (user 1), the first three pages result in no booking. Expedia sees the "same" initial browsing behavior, but with opposite results – one which resulted in a booking and one which did not. In [15] we showed that models built on such incomplete snapshots of web browsing data can result in significantly worse models, and sometimes even in erroneous conclusions.

Generalizing from these, there are many data mining applications characterized by the following features:

1. There is some "local" data available. For example, for Expedia, this local data could be variables constructed from its clickstream (logfile) data. For credit card companies, all customer transactions conducted with their card create local data. Essentially, by local data we mean readily available data that is collected automatically.
2. There is also a specific objective and the target variable (e.g. "purchase prediction", "customer value") related to this is also readily available and is known for all the data records. Expedia needs to understand purchase behavior of customers and Expedia clearly knows which user sessions resulted in purchases and which did not. Credit card companies know which of their customers are profitable for them and which are not. Online media companies know who clicked on an advertisement and who did not.
3. There is additional information representing useful variables that are not usually available, but it is known *what* these variables are. For example, Expedia does not know customer information representing browsing behavior across sites, online media companies know that there are customer characteristics that affect what advertisement is likely to be clicked on but this information is not readily available, credit card companies know that customers transact with other cards but have no information on features of such transactions. In all these cases, even though the data collected is only a snapshot of the true picture, it is easy to identify what the relevant unknown (not collected) variables are.

Note that the first two conditions hold for any data mining application - these indicate the availability of data and a target variable being modeled. The third condition is particularly relevant for the ideas presented in this paper. Note that it can be empirically tested if the 'additional' data is useful. Indeed for personalization, in prior work [15], we show that the magnitudes of the gains obtained from complete data are striking. For example, with complete user browsing behavior, the purchase prediction accuracies in many cases increase by more than 100%.

What can be done in such situations? If it is not possible to acquire this additional data by any means, then there is no fix. In reality for most situations additional data *can* be acquired, but at a cost. Given that it may be possible to acquire additional data, the naïve solution to this problem – acquiring complete data for *all* customers – is impractical in many cases due to the costs of doing so. It may just not be feasible to acquire all the unknown data from all customers (for all the data records). In this paper, we investigate an alternative – whether, and if so how, to acquire complete data for 'some' customers and to use this to improve the models built. We use the term *active data acquisition strategies* to refer to such methods.

Data acquisition by itself is a well-studied problem. Literature in survey sampling [5], experimental design [2, 4] and active learning [6,9,11,13,17] have developed extensive methods that are applicable for different problems. The main goal in survey sampling and experimental design is to have a sample such that inferences from the sample will be applicable to the entire population. Given constraints, non-random sampling strategies can be useful in order to obtain points in parts of the search space that are currently not present in the sample. There are two characteristics here that are different from the data acquisition scenario considered in this paper. First, these problems normally do not know the target values for the points that they acquire and indeed, the main reason for acquiring the points in the first place is to determine what the target value for that point is. Second, the strategies are not goal-directed. They do not acquire points with the specific goal of improving the performance of a *model* – the process of data acquisition and model building are usually independent.

Active learning [7,11,13], on the other hand, represents goal-directed data acquisition. The usual scenario considered in active learning is that all explanatory variables are known, a current model of the target exists but the *target* values are often unknown and expensive to acquire. The problem is to determine which points to acquire this target value from with the specific goal of improving model performance at manageable cost. It is important to note that for the data acquisition scenario considered in this paper, it is *not* the target variables that are unknown, but rather some explanatory variables which are not known and traditional active learning approaches, therefore, cannot be directly applied. However the goal-oriented ideas of active learning could be effective for this problem, though research is needed to study how this can be done. In this paper we present initial approaches and show that active learning ideas can be applied for data acquisition strategies of the type considered in this paper.

We present two active learning based algorithms for data acquisition. The algorithms are based on two different active learning heuristics and show that using active learning ideas for data acquisition can be effective. We present results and discussion based on extensive experimentation using real web usage data as well as UCI datasets [3]. The results demonstrate that the methods perform well and indicate that active learning based methods for data acquisition can be effective and suggest that this may be a promising area for data mining research.

2. Preliminaries

Assume that in the domain, there exists a specific target variable, Y, that is being modeled. For example, Y could be whether or not a user transacts at a web site during a visit.Let N be the number of total data points. Let X_1, X_2,…,X_M, Y be attributes whose values are known initially for all points. We use the term "local data" to refer to data records consisting of X_1, X_2,…,X_M, Y. Let X_{M+1},…, X_P be the attributes whose values are all unknown initially. We use the term "global data" to refer to the complete data X_1, X_2,…, X_P, Y.

In this paper we assume that initially only local data is available for all N records and subsequently global data is acquired for K of these records where $K < N$. As currently structured, the problem of deciding which K points to acquire global data for is still under-specified.

The choice clearly depends on the modeling method used. After acquiring these additional data, there are three scenarios involving how to model Y that can be visualized as shown in Figure 1. In Scenario 1 a local model is built involving the local variables only. This is the default model that exists before any additional data is acquired. In Scenario 2 a global model is built using global data for the K data points. The tradeoff between scenarios 1 and 2 is that the model built in scenario 2 uses more global information but less local information. In Scenario 3, Y is modeled using all available data, but this scenario involves dealing with some complete and some incomplete data in the process of modeling Y. The choice of which K points to acquire complete data from clearly depends on how the final model is built – whether as scenario 2 or 3. In this paper we focus on scenario 2, i.e. when K points are acquired based on active learning, we build a global model using the K points and compare that to the default local model (scenario 1). In order to make this comparison we essentially test the performance of the models on out of sample data where all the variables are known.

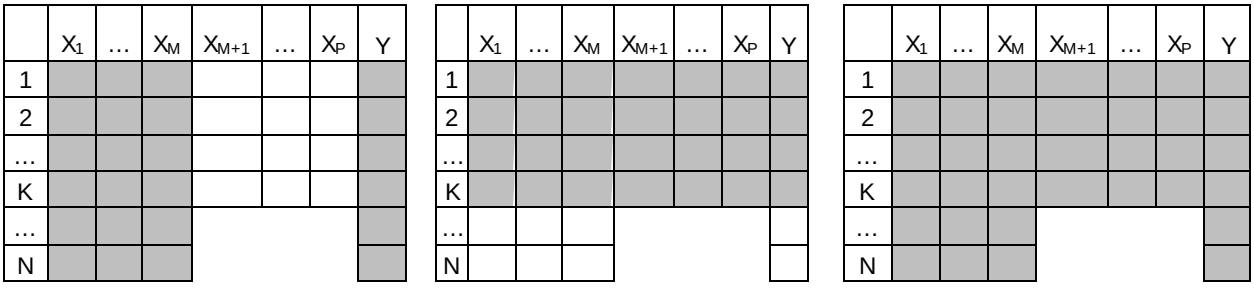

Figure 1. Three scenarios of model building

3. Algorithms for Active Data Acquisition

Let D_L denote the set of all known local data. Each record in D_L therefore consists of ID, X_1, X_2,…, X_M, Y where ID is an index ranging from 1 to N and $| D_L | = N$. D_L remains constant throughout the procedure. D_G is the set of all known global data. Each record in D_G therefore consists of ID, X_1, X_2, …, X_M, X_{M+1},…, X_P,Y. $| D_G | = 0$ initially and $| D_G | = K$ at the end of the data acquisition process. There are two active data acquisition strategies that we present in this section, both based on applying ideas from traditional active learning.

3.1 Algorithm AVID

In this section we present AVID (Acquisition based on Variance of Imputed Data), an algorithm for choosing K data points to acquire global data about. The heuristic used here is determining how useful global data may be based on trying to estimate it from known local data. For instance, for any candidate point, if the global information *can be guessed* from the local information, then global data about this point is less likely to be informative. The literature on missing data [1,12,18] provides several methods for data imputation that can be used for this purpose. The problem here is determining how good the imputation model is for a candidate point, when the true global values for this point are not known. AVID uses an approach which is based on estimating the uncertainties in imputation by using several bootstrap samples to build different imputation models and determining the variance of the imputed values. Points for which the imputed global data has higher variances are points for which the global data can be guessed with less certainty from the local data. Hence these may be good candidates from whom true global data can be acquired.

The idea of estimating variance of an unknown value based on multiple bootstrap samples has been developed independently in both the missing data and the traditional active learning literature. The multiple imputation mechanism proposed in [12,18] works as follows. Each missing value is replaced with a set of plausible values that represent the uncertainty about the right value to impute. Statistical procedures can be applied to each data set and the results are combined according to methods proposed in [12,18]. Bootstrapping is also proposed as a method to generate the set of plausible values [18, 20]. In the active learning literature, [17] proposes a method that determines the variances of class probability estimates empirically based on using several bootstrap samples.

AVID is presented in Figure 2. In addition to the local data, the inputs to AVID are an imputation model, a desirable number of points to be acquired K, a minimum step size (representing the number of points for which global data is acquired at each acquisition stage) and the number of bootstrap samples B.

The set I_G incrementally maintains the list of indexes in the local data for which global data is acquired. Initially this set consists of *sz* (the step size) random points for which global data is acquired (steps 1-10). Once an initial set of global data is acquired, step 12 builds several (*B*) imputation models based on bootstrap samples from the known global data. Steps 14 through 18 applies the imputation models to all the unknown global data in order to determine the points for which the imputation uncertainty is the most. The actual uncertainty score is computed in step 17 and this represents a measure of the variance of the imputed values for all the unknown variables. Step 19 selects the next best *sz* points to acquire. This entire process is continued until the desired number of points, K, are acquired.

Note that AVID does not depend on the classifier and also does not take the actual target values for Y into account in the data acquisition strategy. In this sense, it is a naïve approach for data acquisition. In the next section we present GODA, a *goal-oriented data acquisition* approach, which depends on the classifier and the target values during the course of data acquisition.

Input: Local data D_L, Desired number of points to be acquired K, Step size *sz*, Imputation Method IM, Number of bootstrap samples B

Output: K points for which global data is acquired

```
1  N = |D_L|
2  I_L={1, 2, .., 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 ← randomly select sz integers from I_L-I_G
6  do {
7      Forall (j ∈ S) {
8              Acquire d_G = {ID, X_1,..,X_P, Y} for the element in D_L where ID=j
9              D_G = D_G U d_G     /* add-in this newly acquired global data point */
10             I_G = I_G U {j}
11            }
12     Build B imputation models IM_1, IM_2, … IM_B by applying IM to B bootstrap samples of D_G
13     UID = I_L-I_G /* current set of IDs for which global data is unknown */
14     U_G = {t | t ∈ D_L and t.ID ∈UID}
15     Forall (t ∈ U_G) {
16     x_ij ← the imputed value for variable i of record t using imputation
              model IM_j (M+1≤ i ≤ P,  1≤ j≤ B).
17             Score(t) = Σ_{i=M+1}^{P} sqrt( Σ_{j=1}^{B} (x_ij − m_i)^2 / B )  where μ_i = ( Σ_{j=1}^{B} x_ij ) / B
18            }
19     S ← select min(sz, K-|I_G|) IDs in U_G with the highest scores
       /* alternately, one can sample according to the distribution implied by the scores */
20 }
21 while ( |S| ≤ K)
22 Output: D_G
```

Figure 2. Algorithm AVID (Data Acquisition based on Variance of Imputed Data)

3.2 Algorithm GODA

GODA (goal-oriented data acquisition) chooses K data points to maximize the performance of the model of the target using a given classifier. Assume that at some point in the process, GODA has a subset of known global data. The idea is to choose the next point to maximize the expected improvement of the model built on the global data. The heuristic used here is that for each candidate point, GODA guesses its global data first and then adds this record to the known global data, builds a model based on all available global data and then considers the goodness of the model. GODA chooses the candidate point to acquire additional data from based on the one with the most expected improvement. This is a greedy heuristic.

Rather than fixing what model goodness criterion to use, in GODA we allow various measures by treating this as an input. A number of measures have been proposed in literature [6,17] including prediction accuracy, MSE, lift curve and AIC. In particular we use AIC (Akaike Information Criteria) in our implementation of GODA. AIC measures model performance in terms of the likelihood and complexity of the model. The common AIC is in the form of AIC = -2 $L(\theta)$ + 2$|\theta|$, where θ is the set of the parameters of the classifier and $|\theta|$ is the number of parameters, and $L(\theta)$ is the likelihood of the classifier. AIC can be easily computed for probabilistic classifiers such as Logit model as implemented in our experiments.

Algorithm GODA is presented in Figure 3. The inputs are the local data, a classifier, an imputation model, a desirable number of points to be acquired K and a minimum step size.

The initial steps (1-10) are similar to that of AVID – a random *sz* points are acquired to begin. Once an initial set of global data is acquired, step 12 builds an imputation model from the known global data. Steps 13 and 14 compute 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 a model is built by adding this imputed point to the current known global data (step 17 and 18) and the goodness of this model is computed in Step 19. Step 22 selects the next best *sz* points to acquire. This entire process is continued until the desired number of points, K, are acquired.

In this section we presented AVID and GODA, two algorithms for active data acquisition. In the next section we present experimental results by applying the two algorithms to 15 datasets.

Input: Local data D_L, Desired number of points to be acquired K, Step size *sz*, Classifier C, Goodness Criterion GC, Imputation Method IM
Output: K points for which global data is acquired

```
1  N = |D_L|
2  I_L={1, 2, .., 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 ← randomly select sz integers from I_L-I_G
6  do {
7   Forall (j ∈ S) {
8          Get d_G = {ID, X_1,..,X_P, Y} for the element in D_L where ID=j
9          D_G = D_G U d_G
10         I_G = I_G U {j}
11         }
12  Build an imputation model IM_1 from D_G
13  UID = I_L-I_G /* current set of IDs for which global data is unknown */
14  U_G = {t | t ∈ D_L and t.ID ∈UID}
15  Forall (t ∈ U_G) {
16         d_g = IM_1 (record t)   /* impute record t using IM_1 to get d_g */
17         D_G = D_G U d_g
18         M_G ← Apply classifier C on D_G to get model M_G
19         Score(t) ← Compute goodness score of M_G using GC /*e.g. MSE, AIC*/
20         D_G = D_G - d_g  /* reset D_G */
21     }
22     S ← select min(sz, K-|I_G|) IDs in U_G with the best scores
23    } /* end do */
24 while ( |S| ≤ K)
25 Output: D_G
```

Figure 3: Algorithm GODA (Goal Oriented Data Acquisition)

4. Experiments

We test our results on 15 datasets each of which has a binary target variable. Five of these datasets are UCI datasets [3] and the remaining ten datasets are real user-centric browsing data (described in [15]) for ten popular web sites. For the real 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. The additional information about the users' activities across sites during a browsing session is "global" information. In our prior work [15] we describe how this data is generated from a panel of users whose browsing behavior is tracked. In each of these datasets, 15 of the 40 explanatory variables are local variables. Each UCI dataset as a whole is treated as the global data and the local data is generated by randomly 'hiding' 50% of the variables. The number of total variables in the 5 UCI datasets considered ranges from 4 to 16 (and half are local for each as explained above).

For each dataset, we apply data acquisition algorithms AVID and GODA to acquire global data and subsequently build models on the acquired data. In the absence of data acquisition, the only available data are all the local variables for the entire datasets (scenario 1 in figure 1). Based on this local data, we build local models and treat this as the benchmark against which global models (built from the data points selected by AVID and GODA) are compared. In addition, we also consider random data acquisition as a naïve alternative and use this as an additional benchmark to compare the local and global models. We use mean square error on a random 50% out of sample data to make the comparisons.

For each data acquisition procedure, a final global model is built based on only the data points acquired from the learning sample. This global model's performance is then tested on out of sample data, in which we assume the data points are points for which we know all the global variables. In essence what is being tested here is theoretical model performance, i.e. how good the data acquisition procedures are with respect to building a good final model.

The classifier we use is the Logit model [10] since it is commonly used for binary classification and moreover is relatively fast as compared to other classifiers. The imputation method used in the algorithms use multiple imputation as implemented by the *Proc MI* procedure in SAS 8.2.

4.1 Sample Graphs

We vary the desirable size of global data (*K*) from 0% to 100% of the training 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 15 datasets. Figure 4 and 5 present two examples. The x-axis represents the percentage of acquired global training data and the y-axis represents the MSE (mean square error) of the models on the out of sample data. Each learning curve shows how MSE decreases as more global data are acquired for training. 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 converging point (when all methods acquire 100% of the global data) represents the MSE of a global model built using global variables in the entire training data.

Consider the performance on the Penndigits data (Figure 4). Observe that in general, GODA > AVID > RANDOM. We use the term *critical mass* to refer to the percentage of data at which a model based on acquiring additional global variables beats the performance of the local model. Observe that from just 14% of acquired data based on GODA, a better model can be built than from using the entire local data. It hence represents the point at which additional local data can be traded off for more complete data.

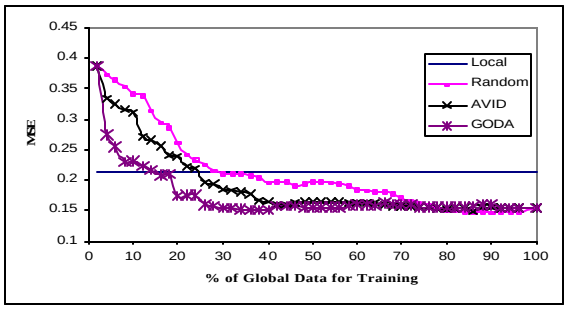

Figure 4. Performance on pendigits dataset

For some other datasets, the results are not as striking. For example, performance on the Amazon.com dataset is shown in Figure 5. The benefit of using GODA and AVID over random acquisition is lower. In this case the critical mass is closer to 20% for all the three methods. To make more general conclusions we studied the performance over a range of datasets and present the results below.

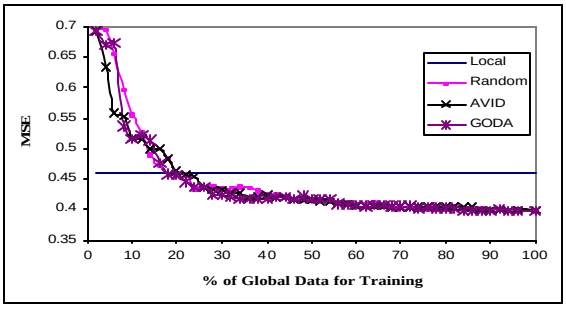

Figure 5. Performance on Amazon dataset

4.2 Comparative Results

In order to draw more general conclusions, we compared metrics across several datasets. From each chart (dataset) we construct the following metrics for each of the three methods (random, AVID, GODA):

Table 1: Experimental results summary

	Global	**Random**			**AVID**			**GODA**		
DataSet	Gain % over local	Critical Mass %	Avg gain %	Avg gain after CM	Critical Mass	Avg gain %	Avg gain after CM	Crit.Mass (CM)	Avg gain %	Avg gain after CM
Amazon	13.3	20.0	3.1	11.1	18.0	4.1	11.6	20.0	5.3	11.6
B&N	10.5	30.0	-1.8	7.1	34.0	2.0	8.3	12.0	6.0	8.0
CDNow	6.5	50.0	-11.0	11.4	60.0	-11.0	10.4	54.0	-6.7	12.8
Expedia	30.2	22.0	19.1	25.8	18.0	18.9	24.4	12.0	19.9	26.8
Travelocity	10.2	32.0	-3.3	8.2	40.0	-3.0	9.0	22.0	1.9	9.6
BMG	3.0	30.0	-4.4	6.2	32.0	-2.0	6.0	46.0	-5.0	6.6
Buy	6.1	28.0	-4.7	2.9	30.0	-2.0	1.3	38.0	-2.8	4.3
QVC	12.2	32.0	-2.9	8.6	32.0	0.0	9.0	24.0	2.5	9.8
Priceline	25.3	28.0	4.4	17.1	24.0	9.3	9.5	8.0	19.3	20.2
Etoys	12.4	44.0	-12.5	7.7	12.0	6.1	9.6	22.0	0.6	11.1
Iris	38.0	16.0	34.3	33.8	8.0	33.0	34.7	8.0	33.0	34.0
Cancer	13.9	62.0	-10.4	9.9	34.0	-4.3	8.7	24.0	-0.5	8.8
Liver	9.0	24.0	2.5	5.8	16.0	4.1	6.2	16.0	6.6	7.8
Pima	10.2	22.0	2.2	8.8	18.0	4.7	7.8	10.0	7.8	9.2
Pendigits	28.6	30.0	1.4	17.6	26.0	10.0	24.0	14.0	17.2	23.7
Average	**15.3**	**31.3**	**1.1**	**12.1**	**26.8**	**4.7**	**12.0**	**22.0**	**7.0**	**13.6**
Avg_web	13.0	31.6	-1.4	10.6	30.0	2.2	9.9	25.8	4.1	12.1
Avg_uci	19.94	30.8	6	15.18	20.4	9.5	16.28	14.4	12.8	16.7

1. Critical Mass - percentage of data to be acquired to outperform the local model
2. Average gain in MSE over the local model computed as average of percentage gains across the entire training data range (0-100%). This value can be highly affected by the initial points when very little acquired data is significantly worse than the local model, hence we also compute the next metric.
3. Average gain in MSE over the local model *after* the critical mass.

For each metric, we seek to get a better handle on the following questions:

- On average, what are the values for these metrics for the methods?
- Is the metric for GODA/AVID significantly better than that for random?

Table 1 summarizes results from each dataset for each method. As a reference, we also report the gain of the global model over the local model (column 2) as the upper bound that each method can attain.

In terms of critical mass, the above results show that a relatively small portion of global data is needed to outperform local models. The average critical masses, across the 15 datasets considered, are 31.3%, 26.8%, and 22% for Random, AVID and GODA respectively. For the 5 UCI datasets selected, GODA only needs 14.4% of global data to beat local models. In terms of average gain and critical mass, GODA > AVID > RANDOM and the pairwise differences are significant.

These results indicate that the methods work well and that active learning based approaches can be useful approaches for data acquisition strategies. There is much opportunity for future research work and better heuristics. Some of the opportunities arise from the various different fields that have studied related questions and in the next section we briefly review them. Subsequently in a discussion in Section 6 below, we go beyond the initial approaches presented thus far and raise several issues that need to be addressed in future research.

5. Related Work

As mentioned in Section 1, data acquisition methods have been considered in survey sampling, experimental design and active learning but in different contexts.

In survey sampling [5], the focus is on drawing inferences about the population through interview, email, telephone, questionnaires, etc. Survey sampling primarily uses simple random sampling. When the population forms into homogeneous groups, the *stratified sampling* or *cluster sampling* [5] method is often used where the population is divided into subgroups and then each group is randomly sampled. The missing data problem encountered in survey sampling is when there is non-response from some of those surveyed [17]. Note that survey sampling is usually goal-independent - the sampling procedure does not depend on how the data is to be used in model building.

Experimental design deals with acquiring data through experiments when the data is not available in natural settings [4]. In order to observe unbiased effect of treatment, randomization is key [4]. When subjects fall into homogenous groups, randomized block design is often used. *Optimal experimental design* (OED) aims to generate a smaller sample in experimental design than regular randomized experimental design [2,8]. OED proposes an incremental method during the course of acquiring data. At each phase, OED decides which subject to be experimented, rather than randomly select from the pool. In doing so, OED uses optimization techniques to decide which subject to go after. A variety of optimization techniques have been developed and further reviews could be found in [2,8].

Active learning is a relatively more recent approach where learning models have control on what data to feed into the model for training [7] and a good summary of active learning is provided in [11,17]. Active learning assumes that the utility of a data point to the model could be discerned by some measure during the course of learning. By selecting only those data with high utility to the model, active learning aims to minimize the number of data needed for training without compromising model performance. Ideas in active learning have similarities with those in optimal experimental design. Active learning methods broadly fall into two categories: heuristic based and optimization based approaches [11].

Query by committee (QBC) [9] is a well-known heuristic-based approach. QBC employs several committee members (each of them is a model) and each member makes its own predictions on unseen data. The data points chosen are those in which there is maximum disagreement among the committee. The rationale here is that those data points that the committee mostly disagrees with are most uncertain to the principle model and thus they are more informative. Another type of heuristic was proposed in [11,13] 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 [13,19]. In [6] data points are selected to minimize the overall prediction variance of a model and [17] selects those data according to the variance of bootstrap predictions of class probability estimates. As mentioned in Section 1, for the data

acquisition scenario considered in this paper, it is not the target variables that are missing, but rather explanatory variables and traditional active learning approaches cannot be directly applied.

The methods proposed in this paper use data imputation as a component. Some commonly used missing data approaches are mean substitution, nearest neighbor substitution, imputation using regression, EM [14], and multiple imputation [12,18]. A good review of these approaches are presented in [1].

6. Discussion

In Section 2, we suggested that data acquisition strategies could depend on how the final model from the acquired data is built. In Figure 1 we presented three scenarios and in this paper only compared scenarios 1 and 2. The third scenario has the potential of doing even better since all the available data will be used. In this scenario, local and global models can be weighted and combined. In future work this will need to be studied.

Less obvious is the fact that good data acquisition strategies could also depend on how the model is *applied* in practice (i.e. how it is used after all the points are acquired and a final model is built). Assume that we have acquired K points and have built a final model which will be used to make predictions for new customers. Now, there are three types of customers (data points) that may be encountered in practice. First, there are *friends*, customers for whom X_1, X_2,…, X_P are all known and the task is to predict Y as well as possible. This is the scenario used in the experiments in this paper where we assumed that in the out of sample data, all the global variables are known. Second, there are *strangers*, customers for whom only local data (X_1, X_2,…,X_M) will be available. In this case, predicting Y better may involve making good guesses on X_{M+1},…, X_P and it would help if the acquired points help in making good guesses. Finally there are *mercenaries*, customers from whom the additional data does not have to be guessed, but can be acquired at a cost.

These represent several opportunities for new data acquisition strategies. In this paper we focused on one such situation - building good global models by data acquisition and show experimentally how the methods perform for friends. In future work we plan to develop data acquisition procedures geared towards strangers and mercenaries and to also develop approaches to combine local and global models as laid out in Figure 1.

In this paper we introduced the idea of using active learning based procedures for data acquisition, presented initial approaches and results from extensive experimentation using proprietary as well as UCI datasets. The initial results indicate that the methods perform well and that data acquisition strategies can be a promising application of active-learning based approaches.

References

[1] Acock, A.C. (1997). Working with missing data. Family Science Review. 10(1):76-102

[2] Atkinson, A. and Donev A., 1992, Optimum Experimental Designs, Oxford Science Publications.

[3] Blake, C.L. & Merz, C.J.,1998, UCI Repository of machine learning databases Irvine, CA: University of California, Department of Information and Computer Science.

[4] Box, G E P, Hunter, W G , and Hunter, J S ,1978, *Statistics for Experimenters*, John Wiley & Sons.

[5] Chaudhuri, Arijit and Stenger, Horst, 1992, Survey Sampling: Theory and Methods, Marcel Dekker, INC.

[6] Cohn, D., Ghahramani, Z., & Jordan, M. (1996). Active learning with statistical models. Journal of Artificial Intelligence Research, 4, 129.

[7] Cohn, D., Minimizing Statistical Bias with Queries, 1995, CBCL Proceedings.

[8] Cohn, D. (1996) , Neural network exploration using optimal experiment design, Journal of Econometrics, 37, 87--114.

[9] Freund, Y., Seung, H., Shamir, E., Tishby, N., 1997, Selective Sampling Using Query by Committee Algorithm, Machine Learning, 28, 133-168.

[10] Friedman J. , H., Hastie, T. and Tibshirani, R. ,1998, Additive Logistic Regression: A statistical view of Boosting. Dept. of Statistics, Stanford University Tech. Report.

[11] Hasenjäger,M.; H. Ritter, Active Learning in Neural Networks, working paper in the university of Bielefeld, 1999, available at http://citeseer.nj.nec.com/404108.html

[12] Little, R. J. A. and D. B. Rubin. 1987, Statistical Analysis with Missing Data., New York: John Wiley & Sons.

[13] MacKay, D. J. C. 1992, Information-based objective functions for active data selection,. Neural Computation, vol. 4 (4), pp. 590-604.

[14] McLachlan G., and Krishnan, T., 1997, The EM Algorithms and Extensions, John Wiley & Sons, Inc.

[15] Padmanabhan, B.; Zheng Z. and Kimbrough, S., 2001, Personalization from Incomplete Data: What You Don't Know Can Hurt, In Proceeding of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD01.

[16] Plutowski, M., & White, H. (1993). Selecting concise training sets from clean data. IEEE Transactions on Neural Networks, 4, 305-318.

[17] Saar-Tsechansky Maytal; Foster J. Provost, 2001, Active Learning for Class Probability Estimation and Ranking, IJCAI 2001.

[18] Schafer, M. K., J. L. & Olsen. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective. Multivariate Behavioral Research , 33 (4), pp. 545-571.

[19] Tong, S., Koller, D., 2001, Active Learning for Structure in Bayesian Networks, In Proceedings of the International Joint Conference on Artificial Intelligence 2001.

[20] Yuan, Y., 2000, Multiple Imputation for Missing Data: Concepts and New Developments, SAS Institute Inc. http://www.sas.com/rnd/app/papers/multipleimputation.pdf.