Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Knowing What to Ask: A Bayesian Active Learning Approach to the Surveying Problem

Yoad Lewenberg
The Hebrew University of Jerusalem, Israel
yoadlew@cs.huji.ac.il

Yoram Bachrach*
Digital Genius Ltd., London, UK
yorambac@gmail.com

Ulrich Paquet†
Microsoft Research, Cambridge, UK
ulripa@microsoft.com

Jeffrey S. Rosenschein
The Hebrew University of Jerusalem, Israel
jeff@cs.huji.ac.il

Abstract

We examine the surveying problem, where we attempt to predict how a target user is likely to respond to questions by iteratively querying that user, collaboratively based on the responses of a sample set of users. We focus on an active learning approach, where the next question we select to ask the user depends on their responses to the previous questions. We propose a method for solving the problem based on a Bayesian dimensionality reduction technique. We empirically evaluate our method, contrasting it to benchmark approaches based on augmented linear regression, and show that it achieves much better predictive performance, and is much more robust when there is missing data.

Introduction

Consider the task of identifying people's political opinions through a poll, such as those conducted by political parties trying to determine whether their agenda is likely to be favored by potential voters. A question in such surveys typically asks people to indicate their agreement with a certain political stance, such as: "we should improve the level of free medical care, even at the expense of raising taxes", "we should have tighter gun control legislation", "abortions should be illegal". Some surveys simply ask participants to indicate whether they agree or disagree with items, whereas others use a scale for the degree of agreement.

Asking each participant all the questions in the survey achieves complete knowledge of the opinions of the participants. However, this is very costly, especially when the bank of questions is large. Given a limited budget, one possibility for lowering costs is exploiting correlations between responses to different questions. For example, if for a large sample of participants we observe that the response to one question is strongly correlated with the response to another question, we can ask further survey participants only the first question, and use their response to predict the response to the other question. More generally, one may choose a subset of the questions in the bank to ask, and predict the responses to the remaining questions based on the responses to the asked questions.

Asking all the participants the same set of questions may be suboptimal. Instead, we may tailor the next question to ask a participant based on their responses to the previous questions. Such a design is based on the active learning paradigm, where the learning algorithm can interactively query users so as to best predict the responses to all questions in the bank.

The Active Surveying Problem We consider a set of d survey questions, which can be answered by a sample of n users. The responses form an $n \times d$ matrix. We denote the set of possible responses to each question as $\mathcal{L}$. In the case where participants may only agree or disagree with a query we have $\mathcal{L} = \{0, 1\}$, but in the general case we may have L ordinal labels: $\mathcal{L} = \{0, \ldots, L-1\}$ (representing for example, a scale for the degree of agreement ranging from completely disagreeing to completely agreeing with the item). Our data has the form of a matrix $\mathcal{D} \in \mathcal{L}^{n \times d}$. However, the data may be *partial*, with some entries in $\mathcal{D}$ being missing (unobserved).

Our goal in active surveying is to minimize the prediction error given a *limited budget* of questions we are allowed to pose to a *new survey participant*. We consider the case where the budget of questions to ask is smaller than the total number of questions in the survey, so we need to generate a prediction regarding *multiple* questions which cannot be posed to the survey participant due to the limited question budget. As we know the responses to the questions we actually ask the participant, we can clearly achieve zero error on these questions, so we wish to minimize the error on the unasked questions.

As opposed to a standard supervised learning problem, we are allowed to decide on the next question to ask *after* we have observed the participant's response to the previous questions. Furthermore, we are allowed to leverage knowledge regarding correlations in the responses to several questions, based on the information obtained from past survey participants. Similarly to many dimensionality-reduction-based approaches, we represent the correlations between the users' responses to the various questions using a low dimensional structure. We assume that a user u and a question q can be represented as an h-dimensional vector (where h is

*Part of the work was done while at Microsoft Research, Cambridge, UK

†Now at DeepMind

Algorithm 1 Active Surveying Framework

Predict: $\mathcal{D}$,x, j
Input: Dataset $\mathcal{D} \in \mathcal{L}^{n \times d}$ ($\mathcal{D}$ may be partially observed), a partially observed vector $x \in \mathcal{L}^d$ for the target user and an index j
Output: A prediction of the value of x_j
Acquire: $\mathcal{D}$,x
Input: $\mathcal{D}$ and x as for Predict
Output: An index of the next element of x that should be acquired

some low dimension), $v_u \in \mathbb{R}^h$ for the user and $v_q \in \mathbb{R}^h$ for the question, so that the user's response to the question, $\mathcal{D}_{u,q}$, is determined mostly by the dot product between these vectors, $v_u^T v_q$. Thus we assume there exist latent matrices $U \in \mathbb{R}^{h \times d}$ and $V \in \mathbb{R}^{h \times n}$, so that $\mathcal{D}_i = U^T V_i + \mu + \epsilon$, where $\mathcal{D}_i$ and V_i are the i-th columns of $\mathcal{D}$ and V, respectively, representing the i-th user; $\mu \in \mathbb{R}^d$ is the (latent) model bias and $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$ is the (latent) model noise.

We focus on the task of actively determining the responses of a target user. We examine algorithms that query the target user for their responses to questions, one at a time, as long as the budget allows. After observing a response, and based on the responses of other users, the algorithm must *predict* the responses of the target user to each of the as yet unobserved questions, and decide which response to *acquire* next (i.e., choose the next question to pose to the target). Formally, given a new partially-observed response vector for the target user, $x = (x_1, \ldots, x_d) \in \mathcal{L}^d$, where x_j's value is observed only if $j \in O$ for some $O \subset [d]$, the algorithm must predict the values of x_j for $j \notin O$, and choose the next element of x to acquire.

Our active surveying framework is given in Algorithm 1. Every method for our active surveying problem must implement the *predict* and *acquire* procedures.

Differences Between Active Surveying and Collaborative Filtering We note that the active surveying problem we pose here has some resemblance to collaborative filtering-based recommender systems. Such systems attempt to recommend items to a user. In the collaborative filtering approach, the system identifies users who are "similar" to the target user in the sense that they have bought similar items to theirs in the past; the system then recommends items that the target user has not yet examined, but that many users similar to the target have bought.

A key difference between active surveying and collaborative filtering-based recommender systems is that in active surveying we are interested in how the target is likely to respond to *all* the questions, and not just in locating a few items that the target user is very likely to rate highly. More specifically, in a collaborative filtering scenario, the system does not need to obtain information regarding items that are likely to have low ratings—if the target user is not likely to give high ratings to items a and b, there is no point in determining whether a is better than b or vice versa. Thus, collaborative filtering can be designed to focus its attention and obtain very accurate ratings for items that highly match the user's preferences, and not waste effort or questions on items that are already known not to match the user's preferences (Mitzenmacher, Pagh, and Pham 2014).

In contrast, in active surveying our goal is to predict the user's responses to *all* the unobserved questions. It is thus not sufficient to simply locate a few items with which the user is very likely to agree—we need to have as reliable an estimate as we can to how the user is likely to rate each of the remaining questions.

Our Contribution We propose an approach for solving the active surveying problem using a Bayesian active learning framework. We use a probabilistic graphical model, called *DRAL* (Dimensionality Reduction Active Learning) akin to various forms of Bayesian matrix factorization, representing each participant and question as a vector in a low-dimensional space, so that the prediction regarding the response of a participant to a question depends on the inner product between the user and question vectors. Given the responses of users to questions, and given the partial responses of the target user to some questions, we use DRAL to obtain posterior distributions for the low-dimensional user and question vectors, and can thus predict the target's responses to the remaining questions. Given these distributions we can also estimate the reduction in uncertainty we expect to achieve by any possible next query, allowing us to select the next best question to ask.

We empirically evaluate our model on a dataset consisting of the responses of participants to *political questions*, contrasting it with alternative approaches based on linear regression and PCA. We show that DRAL achieves a *better tradeoff between prediction quality and the number of user queries*, allowing us to significantly reduce the number of questions we ask survey participants. Furthermore, we show that as opposed to these alternative approaches, DRAL is a robust method that is *resistant to a loss in data*: while the performance of the alternatives drops quickly as some responses are missing, DRAL still achieves high-quality predictions even when large proportions of the data are missing.

Solving the Active Surveying Problem

We now discuss several methods for active surveying. We begin with baseline approaches based on linear regression, considering both the setting where there is complete data, and the setting where there is missing data. We then present our Bayesian approach, based on the probabilistic graphical model DRAL, and discuss several alternatives for the Acquire step, based on the model's inferred posterior distributions.

Linear Regression-Based Methods

Given the data $\mathcal{D}$ for a sample of users, one can use linear regression to predict the responses of a target user. First, consider the case where $\mathcal{D}$ is fully observed, and a target user x. Denote by O the entries that we have observed for the target user (i.e., O contains the indices of the questions for which we have already queried the target's responses).

Algorithm 2 Linear Regression— Fully Observed Dataset

Predict: $\mathcal{D}$,x, j
$O =$ the set of observed indexes of x
Solve $\arg\min_{b_j, w_{k,j}}$
$\sum_{i=1}^{n} \left(b_j + \sum_{k \in O} w_{k,j} \cdot \mathcal{D}_{i,k} - \mathcal{D}_{i,j}\right)^2$
return $b_j + \sum_{k \in O} w_{k,j} \cdot x_k$
Acquire: $\mathcal{D}$,x
$O =$ the set of observed indexes of x
for $j \notin O$ in **do**
  $e_j =$ training error for predicting x_j
end for
return $\arg\max e_j$

As the questions are believed to be correlated with one another, using $\mathcal{D}$ we can train a linear regression model for each of the variables that are not in O, and predict x's responses using these models. More precisely, for an unobserved entry of x we apply linear regression and approximate $x_j = b_j + \sum_{k \in O} w_{k,j} \cdot x_k$ for $j \notin O$, when the weights b_j and $w_{k,j}$ are the least-square estimators learned from $\mathcal{D}$.

A plausible way to choose the next question to ask the target is acquiring the element about which the model is most uncertain, with the *greatest training error* on the data from $\mathcal{D}$. We note that this method for selecting the next query (the Acquire procedure) works for any prediction method (linear regression or otherwise). The pseudocode for this simple linear regression-based active surveying is given in Algorithm 2.

Handling Missing Data So far we only considered the case where we observe all the entries for the sample users, and only have missing data for the target user. How should we augment the linear regression-based method to handle missing entries in $\mathcal{D}$? When predicting the unobserved entries for x, we might assume that the linear relations between x's entries still resemble the relations observed for the sample users of $\mathcal{D}$, so a reasonable solution is to apply adaptations of linear regression for missing variables. We briefly review two existing methods for linear regression with missing data: the Complete Case Method (CCM) and the Missing Indicator Method (MIM) (Jones 1996).

The Complete Case Method (CCM) CCM simply excludes from $\mathcal{D}$ any column that is not completely observed (Jones 1996). However, in many settings there may not even exist a single sample user for which we observe all the responses. Given a partially observed vector for the target x, we wish to predict the x_j's for $j \notin O$, trying to linearly approximate $x_j = b_j + \sum_{k \in O} w_{k,j} \cdot x_k$. As we train models based only on the features (questions) O (observed for the target x), we can include a sample user s if all the responses in $O \cup \{j\}$ are observed for s (even if there are missing entries for some question $q \notin O$). Still, even this relaxed criterion may leave us with no users.

Given a parameter K, we seek for a subset of questions $\tilde{O} \subset O$ such that there are at least K sample users for which we have observed the responses of *all* of these $|\tilde{O}|$ questions (i.e., none are missing). After locating such sample users, we simply apply the linear regression active surveying method for full data. The pseudocode for the CCM active surveying method is given in Algorithm 3 (Predict), while the procedure for Acquire remains as in the full data case (Algorithm 2).

Algorithm 3 The Complete Case Method

Predict: $\mathcal{D}$,x, j, K
$O =$ the set of observed indexes of x
Find a maximal $\tilde{O} \subset O$, such that there is $I \subset [n]$, $|I| \geq K$ and for every $i \in I$, $k \in \tilde{O} \cup \{j\}$: $\mathcal{D}_{i,k}$ is observed.
Solve $\arg\min_{b_j, w_{k,j}}$
$\sum_{i \in I} \left(b_j + \sum_{k \in \tilde{O}} w_{k,j} \cdot \mathcal{D}_{i,k} - \mathcal{D}_{i,j}\right)^2$
return $b_j + \sum_{k \in \tilde{O}} w_{k,j} \cdot x_k$

Algorithm 4 Missing Indicator Method

Predict: $\mathcal{D}$,x, j
$q_{i,k}$ equals to 1 iff $\mathcal{D}_{i,k}$ is observed
Solve $\arg\min_{b_j, w_{k,j}}$
$\sum_{i:q_{i,k}=1} \Big(b_j + \sum_{k \neq j} w_{k,j} \cdot \mathcal{D}_{i,k} \cdot q_k^i$
$+ \sum_{k \neq j} \tilde{w}_{k,j} \cdot \mathcal{D}_{i,k} \cdot (1 - q_k^i) - \mathcal{D}_{i,j}\Big)^2$
return $x_j = b_j + \sum_{k \in O} w_{k,j} \cdot x_k + \sum_{k \notin O} \tilde{w}_{k,j}$

The Missing Indicator Method (MIM) The MIM method augments the original data with a "survival" indicator variable per each original variable (Jones 1996). We denote by $q_{i,k}$ the survival indicator of $\mathcal{D}_{i,k}$, defined as $q_{i,k} = 1$ if $\mathcal{D}_{i,k}$ is observed, and 0 otherwise. We now approximate $x_j = b_j + \sum_{k \in O} w_{k,j} \cdot x_k + \sum_{k \notin O} \tilde{w}_{k,j}$. The weights b_j, $w_{k,j}$ and $\tilde{w}_{k,j}$ are the least-square estimators learned from $\tilde{\mathcal{D}} = \{z^i \in \mathcal{D} : q_j^i = 1\}$.

Using MIM for active surveying simply requires making predictions using MIM linear regression (Algorithm 4), and using the Acquire procedure for full data (of Algorithm 2).

We also compared our approach (DRAL) to PCA-based methods. As our approach outperforms PCA even when comparing PCA on the full data against our method in the 75% missing data case, and due to lack of space, we elaborate on that in the full version of the paper.

The Dimensionality Reduction Active Learning (DRAL) Model

The model we use is called Dimensionality Reduction for Active Learning (DRAL). DRAL is a probabilistic graphical model resembling other Bayesian matrix factorization models (Agarwal and Chen 2010; Porteous, Asuncion, and Welling 2010; Stern, Herbrich, and Graepel 2009).

Graphical models were introduced by Pearl (2014), and we use the more general framework of factor graphs (see, e.g., (Koller and Friedman 2009)) in order to describe the factored structure of the assumed joint probability distribution among the variables. Once the graphical model is defined and the values of the observed variables are set, in-

ference algorithms (such as approximate message-passing methods) can be used in order to infer the marginal probability distribution of the unknown variables (Koller and Friedman 2009).

Our model assumes that the users and questions can be characterized by h underlying "traits". The number of dimensions h of the model, i.e., the size of the user and question trait vectors, is determined prior to the construction of the model. We model the process by which a user produces a response r to a question using the inner product between two h dimensional vectors of unobserved (latent) variables, one for the user and one for the question. Thus we assume that: a) every user has a latent trait vector and a latent bias (this bias captures the fact that some participants, on average, give higher response labels); and b) each question has a latent trait vector and a latent bias. Information such as the latent vector and bias of the users and questions are modeled as unobserved variables, whereas the given responses to a question by a user is modeled as an observed variable.

The data for DRAL is a (partially observed) matrix $\mathcal{D} \in \mathcal{L}^{n,d}$, and the partially observed response vector for the target user $x \in \mathcal{L}^n$. We denote by $\mathcal{D}_{i,j}$ the response that user i gives to question j, so the model identifies each row of $\mathcal{D}$ and x as users, and the columns of $\mathcal{D}$ are the questions.

DRAL associates with every user i a latent user trait vector $s_i \in \mathbb{R}^h$ and a latent bias parameter $b_{s_i} \in \mathbb{R}$. Similarly, it associates with every question j a latent question trait vector $t_j \in \mathbb{R}^h$ and a latent bias parameter $b_{t_j} \in \mathbb{R}$. In addition, for each user DRAL maintains user-specific threshold vector $\theta_i \in \mathbb{R}^{L+1}$, which divides the latent rating axis into $\mathcal{L}$ consecutive intervals representing an ordinal scale ($\theta_{i,0}$ and $\theta_{i,L}$ are fixed to $-\infty$ and $+\infty$, respectively). For Boolean responses, this is simply a single threshold (reflecting the boundary between negative and positive prediction).

The response that the i-th user gives to the j-th question is modeled as: $\mathcal{D}_{i,j} = r \iff \mathcal{N}\left(s_i^T t_j + b, \beta^2\right) \in [\theta_{i,r}, \theta_{i,r+1}]$, where $\mathcal{N}$ denotes a Gaussian distribution and $b = b_{s_i} + b_{t_j}$ is the bias of the user and the question and β is the standard deviation of the observation noise.

To do inference on the model, we first define prior distributions for variables of interest. We assume independent Gaussian distributions for the user vector traits $p\left(s_{i,k}\right) = \mathcal{N}(s_{i,k}; \mu_s, \sigma_s^2)$ and bias $p\left(b_{s_i}\right) = \mathcal{N}(b_{s_i}; \mu_s, \sigma_s^2)$, and question vector traits $p\left(t_{j,k}\right) = \mathcal{N}(t_{j,k}; \mu_t, \sigma_t^2)$, and bias $p\left(b_{t_j}\right) = \mathcal{N}(b_{t_j}; \mu_s, \sigma_t^2)$. The Gaussian prior allows us to specify a range of values using two parameters, and to admit simple approximate inference.

The joint distribution of all the variables factorizes as:

$$p\left(s_i, t_j, b_s, b_t, \tilde{r}, r\right) = p\left(r|\tilde{r}\right) \cdot p\left(\tilde{r}|s_i, t_j, b_s, b_t\right) \cdot \\ \cdot p\left(b_{s_i}\right) \cdot p\left(b_{t_j}\right) \cdot \prod_{k=1}^{h} p\left(s_{i,k}\right) \cdot \prod_{k=1}^{h} p\left(t_{j,k}\right),$$

where $\tilde{r}$ it the latent rating before adding noise, and is given by $p\left(\tilde{r}|s_i, t_j, b_{s_i}, b_{t_j}\right) = \mathbb{I}\left(\tilde{r} = s_i^T t_j + b_{s_i} + b_{t_j}\right)$, and $p\left(r|\tilde{r}\right) = \mathcal{N}\left(r, \beta^2\right)$. The indicator function $\mathbb{I}\left(\cdot\right)$ is equal to 1 if the proposition in the argument is true and 0 if it is false. The posterior distribution over s_i, t_j, b_{s_i} and b_{t_j}, given the observed rating $r_{i,j}$, is given by summing out the latent variables:

$$p\left(s_i, t_j, b_{s_i}, b_{t_j}|r_{i,j}\right) \propto \int_{\tilde{r}} p\left(s_i, t_j, b_{s_i}, b_{t_j}, \tilde{r}, r_{i,j}\right) \mathrm{d}\tilde{r}. \quad (1)$$

A factor graph representation of DRAL is given in Figure 1.

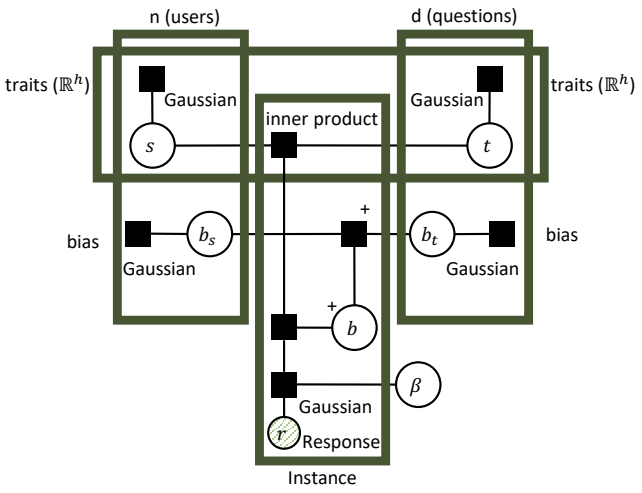


Figure 1: Factor graph for the DRAL model

We performed the inference in the DRAL model using message passage algorithms, implemented in Infer.NET (Minka et al. 2014). Specifically, we used Expectation Propagation (EP) (Minka 2001), so inference was approximate. EP calculates marginal distributions on a given factor graph by iteratively computing messages along edges that propagate information across the factor graph. As EP runs iteratively until convergence, the runtime is linear in the model's size, which in the case of DRAL is $\mathcal{O}(n \cdot d)$. We note that DRAL can handle fully-observed datasets as well as partially-observed datasets without any modification.

Active Surveying Using DRAL We now show how to use DRAL for active surveying. Consider a target user with a set of partially observed responses x (responses to the previous queries performed by the algorithm, denoted by O). Denote by s the latent vector representing this target user.

Prediction For every unanswered question $j \notin O$ we have a posterior probability distribution: $p_j\left(r\right) = \Pr\left[x_j = r\right] = \Pr\left[\theta_r \leq \mathcal{N}\left(s^T t_j + b, \beta^2\right) \leq \theta_{r+1}\right]$. We can predict the value of any unobserved entry by the expected value of r: $\mathbb{E}[r] = \sum_{r=0}^{L-1} r \cdot p_j(r)$.

Acquisition For every latent variable we also obtain a posterior distribution, which in our approximate inference procedure is captured as a Gaussian posterior distribution. The uncertainty of the model is captured by the entropy. The entropy of the j-th question is:

$$h\left(p_j\right) = -\sum_{r=0}^{L-1} p_j\left(r\right) \log\left(p_j\left(r\right)\right) \quad (2)$$

and the average uncertainty of all unobserved questions is: $h\left(x\right) = \frac{1}{N-|O|} \sum_{j \notin O} h\left(p_j\right)$.

- All couples should be given equal status whether they are heterosexual or homosexual.
- Homosexual couples should not be allowed to adopt or raise children.
- The government should subsidize religious education.
- The government should increase unemployment benefits.

Figure 2: An example of a few questions from the questionnaire. The users had been asked to indicate whether they agree or disagree with each such statement.

We propose three methods for choosing the next question to acquire for the target user.

Information Gain Selection Given the posterior distributions, we can choose the next question to ask the target so as to maximize the expected information gain (IG). For an unobserved question $j^* \notin O$ and $r \in \mathcal{L}$, the expected entropy of the model given that $x_j^* = r$ is $h\left(x|x_j^*=r\right) = \frac{1}{N-|O|+1}\sum_{j\notin O, j\neq j^*} h\left(p_j\right)_{|x_{j^*}=r}$, where $h\left(p_j\right)_{|x_{j^*}=r}$ is the posterior entropy of the j-th element given $x_{j^*} = r$. Thus, the expected information gain from acquiring element j^* is $ig(x,j^*) = \sum_{r=0}^{L-1} p_{j^*}(r) \cdot h\left(x|x_j^*=r\right)$. We can maximize the information gain by selecting the unobserved question that maximally reduces the posterior entropy: $\arg\min_{j^*\notin O} ig(x,j^*)$.

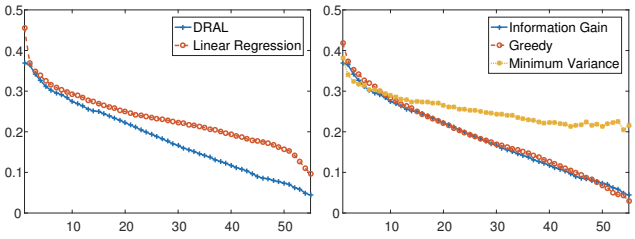
The Greedy Heuristic A drawback of information gain selection is its runtime. With b possible choices for the next query (i.e., questions where the target's response was not yet observed), and with L possible labels for each question, the information gain method builds $b \cdot L$ models and recomputes approximate posterior distributions in each. As this is very costly, a simple alternative is to use the data instance with the greatest model uncertainty (Huang 2007). Thus, the greedy heuristic acquires the maximal entropy element: $\arg\max_{j^*\notin O} h(p_{j^*})$ where $h(p_{j^*})$ is as defined above in Equation 2.

The Minimum Variance Heuristic The goal of information gain selection is to minimize the uncertainty regarding the *unobserved questions*. An alternative is to try and best place the target user in the *latent trait space*. A low variance of the posterior distribution of the user trait vector suggests that the model is more certain about the responses that the user would give to the questions, while a high variance suggests that the model is uncertain about the user's responses. Thus, the *Minimum Variance* heuristic chooses the next element of x to acquire as the element resulting in a minimal posterior total variance of the latent user variables.

Empirical Analysis

We now describe our empirical evaluation of DRAL and linear regression approaches. We first describe the dataset, then discuss our algorithm evaluation experiments.

Dataset Our dataset is based on a political stance questionnaire, posed to 1,500 users from the United States, sourced from Amazon Mechanical Turk, a prominent crowdsourcing platform. The questionnaire consists of 56 political statements, and users had to indicate whether they agree or disagree with each such statement. The statements were about various political topics, including religion, abortion, gay/lesbian rights, public health care, and immigration. Several example items are given in Figure 2, and the complete questionnaire is available in the full version of the paper.



(a) Linear Regression and DRAL with IG (b) Different heuristics of DRAL model

Figure 3: The MSE as a function of the number of exposed entries, where $\mathcal{D}$ is fully observed

Experiment Settings In each of our experiments, we randomly selected 500 of our 1,500 users as a sample population. Out of a sample of 500 users, we randomly chose a single one to be the *target user*. Thus, in each experiment the sample user data, $\mathcal{D}$, is a 499×56 binary matrix. Each entry in the matrix is either 1 (the user agreed with the statement), or 0 (the user disagreed with the statement).

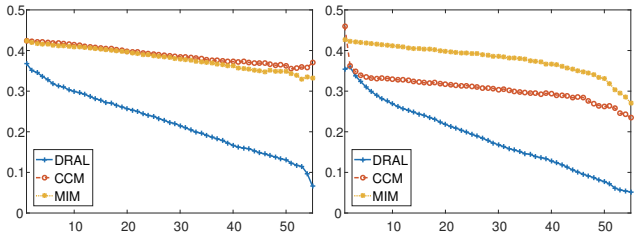
For all of the Gaussian distributions we use the standard Gaussian distribution, with zero mean and unit variance.

The size of the trait vectors can be determined using Bayesian model selection techniques (see, e.g., (Lewenberg et al. 2016)). In practice, DRAL with different h values shows similar results, and therefore we set $h = 5$; the results with different h values can be found in the full version of the paper.

In each trial, we allow an evaluated algorithm to examine the responses of the target one at a time. Once the algorithm has observed k entries, we ask it to predict the remaining $d-k$ entries, and to select the next entry to examine. We measure the performance of the algorithms, for a given number k of observed responses by their MSE on the remaining $d-k$ entries (as our data contains Boolean responses, this is simply the number of mispredicted entries).

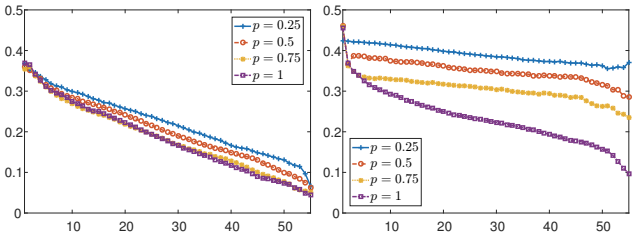
We first examine the case of complete data for the training users, i.e., all responses for the non-target users are given in advance, so $\mathcal{D}$ is fully observed. We then discuss missing data for the training users. Our analysis is done by making a certain proportion p of the entries unobserved: given a parameter p, we take each entry in $\mathcal{D}$ and eliminate it with probability $1-p$. We used the following values for p: $\{0.25, 0.5, 0.75, 0.95\}$.

Our results are presented as MSE plots, where the x-axis corresponds to the number of entries k an algorithm is allowed to observe, and the y-axis is the MSE achieved on the



(a) $p = 25\%$ (b) $p = 75\%$

Figure 4: The MSE of Linear Regression (CCM and MIM) and DRAL model (with IG selection), where every entry in $\mathcal{D}$ is observed with probability p



(a) DRAL (b) Linear Regression (CCM)

Figure 5: The MSE for different p values

remaining $d-k$ entries (so a lower value indicates better performance). Each point in the plots is generated by averaging the MSE over 500 trials (in each of which we selected a different sample of training users and a different target user).

Thus the algorithm is represented as a curve, indicating the possible tradeoff between the amount of examined entries and the performance achieved.

Full Data We compared the performance of DRAL and linear regression (with training error-based Aquire) for the complete data case ($\mathcal{D}$ is fully observed). Figure 3(a) contrasts the performance of DRAL (with information gain selection) and linear regression. It clearly shows that DRAL achieves a much better tradeoff between prediction quality and the number of entries examined. Figure 3(b) compares DRAL under different methods for acquiring the next entry to be examined. It shows that the information gain method for selecting the next entry outperforms the Minimum Variance heuristic. Interestingly, the Greedy heuristic achieves a very similar performance to the information gain method, despite having much lower computational overhead.

Missing Data We now turn to the missing data case, where entries for the training users only survive (remain observed) with probability p. Figure 4 (analogous to Figure 3(a)) contrasts the performance of DRAL and the Linear Regression method (with either the Complete Case Method, CCM, or the Missing Indicator Method, MIM, for handling missing entries), for various survival probabilities p: 0.25 and 0.75 in Figures 4(a) and 4(b) respectively. (Plots for p: 0.5 and 0.95 can be found in the full version of the paper). When we ran linear regression with CCM, we set $K = 50$, so we learn from a dataset with at least 50 complete case users.

The results show that the performance of the linear regression-based methods degrades very quickly as a larger proportion of data entries become unobserved. In contrast, DRAL is much more robust to the elimination of training user entries. Although the performance does degrade slightly, DRAL still achieves very good performance even when a very large proportion of the entries are eliminated.

As in the full data case, under all entry survival probabilities, the best performance is achieved by information gain selection; however, in most cases the performance of the greedy heuristic is still very close.

Figure 5 depicts the robustness of DRAL (information gain selection) (5(a)) and Liner Regression with the Complete Case Method (5(b)) against the loss of data, under different survival probabilities. Though decreasing the survival probability slightly increases the error of DRAL (for any number of target user queries), the prediction degradation is slight, as opposed to Linear Regression methods, where performance quickly drops as entries become unobserved.

Related Work

Active learning relates to many problems where a learning algorithm is given control of the data acquisition process so as to require less training to achieve good performance (Settles 2010). Some scenarios allow the learner to request labels for any unlabeled instance, or even synthesize queries (Angluin 1988; 2004): Namata et al. (2012) dealt with surveying strategies where the goal is to obtain the labels of nodes in a network structure; Sharara, Getoor, and Norton (2011) applied active surveying methods for identifying opinion leaders. Garnett et al. (2012) proposed active surveying methods for identifying users belonging to a certain class and for predicting the portion of the dataset belonging to a certain class.

In other scenarios, unlabeled instances arrive in a stream and the learner has to decide whether to obtain a costly label (Cohn 1994): Yu (2005) proposed active learning with an SVM algorithm that receives at each round a set of unlabeled samples, and selects the most ambiguous ones, and various adaptive sourcing approaches have been proposed for skill-based domains (Kosinski et al. 2012; Bachrach et al. 2012b; 2012a; Salek, Bachrach, and Key 2013).

Active learning was studied in collaborative filtering (Boutilier, Zemel, and Marlin 2002; Harpale and Yang 2008); however, as we discussed, active surveying is very different from collaborative filtering, as we are required to achieve a low prediction error in many unasked questions (as opposed to collaborative filtering, where we are only required to find *one* item which is likely to obtain a *high* rating).

There are several common methods for query selection. In uncertainty sampling, the instance that the learner is least certain how to label is the next one to be queried (Lewis and Catlett 1994; Settles and Craven 2008). Another strategy is to identify the instance that would impart the greatest change to the current model if we knew its label (Settles, Craven, and Ray 2008). Other approaches select the sample

generating the lowest expected error on other examples (Roy and McCallum 2001), and selecting the sample reducing the model variance (MacKay 1992; Schein and Ungar 2007).

Our DRAL model assumes the full response matrix can be decomposed as a product of two latent matrices. Probabilistic Matrix Factorization has been studied and several approximate inference methods were suggested, including Markov Chain Monte Carlo (Salakhutdinov and Mnih 2008), Latent Dirichlet Allocation (Agarwal and Chen 2010) and approximate message passing (Stern, Herbrich, and Graepel 2009).

Conclusions

We studied the active surveying problem, and proposed solving it using the DRAL model. We contrasted our approach with alternatives based on augmented linear regression, showing that our method achieves better predictive performance, and is more robust to missing data.

Our problem is reminiscent of other active learning scenarios examined in the past, but in contrast to the papers discussed in the related work section, we examine the active surveying setting, where the learner issues queries regarding the same target user, whose responses it tries to predict. Furthermore, our approach is based on maximizing expected information gain, computed from the posterior distributions in our probabilistic graphical model.

Several questions remain open for further research. First, does our model achieve good performance and robustness to data loss in other domains, and in particular in making predictions for recommender systems? Second, could non-Bayesian approaches, perhaps not based on linear regression, achieve comparable performance? Finally, could an alternative Bayesian model or an alternative graphical model outperform our model for active surveying?

Acknowledgments

This research has been partly funded by Microsoft Research through its PhD Scholarship Program, and by Israel Science Foundation grant #1227/12.

References

Agarwal, D., and Chen, B.-C. 2010. fLDA: matrix factorization through latent Dirichlet allocation. In *WSDM*.

Angluin, D. 1988. Queries and concept learning. *Machine learning* 2(4):319–342.

Angluin, D. 2004. Queries revisited. *Theoretical Computer Science* 313(2):175–194.

Bachrach, Y.; Graepel, T.; Kasneci, G.; Kosinski, M.; and Van Gael, J. 2012a. Crowd IQ: aggregating opinions to boost performance. In *AAMAS*.

Bachrach, Y.; Graepel, T.; Minka, T.; and Guiver, J. 2012b. How to grade a test without knowing the answers—a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *ICML*.

Boutilier, C.; Zemel, R. S.; and Marlin, B. 2002. Active collaborative filtering. In *UAI*.

Cohn, D. A. 1994. Neural network exploration using optimal experiment design. In *NIPS*.

Garnett, R.; Krishnamurthy, Y.; Xiong, X.; Schneider, J. G.; and Mann, R. P. 2012. Bayesian optimal active search and surveying. In *ICML*.

Harpale, A. S., and Yang, Y. 2008. Personalized active learning for collaborative filtering. In *SIGIR*.

Huang, Z. 2007. Selectively acquiring ratings for product recommendation. In *EC*.

Jones, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association* 91(433):222–230.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Kosinski, M.; Bachrach, Y.; Kasneci, G.; Van-Gael, J.; and Graepel, T. 2012. Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *WebSci*. ACM.

Lewenberg, Y.; Bachrach, Y.; Bordeaux, L.; and Kohli, P. 2016. Political dimensionality estimation using a probabilistic graphical model. In *UAI*.

Lewis, D. D., and Catlett, J. 1994. Heterogeneous uncertainty sampling for supervised learning. In *ICML*.

MacKay, D. 1992. Information-based objective functions for active data selection. *Neural computation* 4(4):590–604.

Minka, T.; Winn, J. M.; Guiver, J. P.; Webster, S.; Zaykov, Y.; Yangel, B.; Spengler, A.; and Bronskill, J. 2014. Infer.NET 2.6. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Minka, T. P. 2001. Expectation propagation for approximate Bayesian inference. In *UAI*.

Mitzenmacher, M.; Pagh, R.; and Pham, N. 2014. Efficient estimation for high similarities using odd sketches. In *WWW*.

Namata, G.; London, B.; Getoor, L.; and Huang, B. 2012. Query-driven active surveying for collective classification. In *MLG*.

Pearl, J. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Porteous, I.; Asuncion, A. U.; and Welling, M. 2010. Bayesian matrix factorization with side information and Dirichlet process mixtures. In *AAAI*.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through Monte Carlo estimation of error reduction. *ICML*.

Salakhutdinov, R., and Mnih, A. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *ICML*.

Salek, M.; Bachrach, Y.; and Key, P. 2013. Hotspotting—a probabilistic graphical model for image object localization through crowdsourcing. In *AAAI*.

Schein, A. I., and Ungar, L. H. 2007. Active learning for logistic regression: an evaluation. *Machine Learning* 68(3):235–265.

Settles, B., and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-instance active learning. In *NIPS*.

Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52(55-66):11.

Sharara, H.; Getoor, L.; and Norton, M. 2011. Active surveying: A probabilistic approach for identifying key opinion leaders. In *IJCAI*.

Stern, D. H.; Herbrich, R.; and Graepel, T. 2009. Matchbox: large scale online Bayesian recommendations. In *WWW*.

Yu, H. 2005. SVM selective sampling for ranking with application to data retrieval. In *KDD*.