Whom Should I Perform the Lab Test on Next? An Active Feature Elicitation Approach

Sriraam Natarajan[1,2], **Srijita Das**[2], **Nandini Ramanan**[2], **Gautam Kunapuli**[1], **Predrag Radivojac**[2]
[1] University of Texas at Dallas, [2] Indiana University Bloomington
{sriraam.natarajan,gautam.kunapuli}@utdallas.edu,{sridas,nramanan,predrag}@indiana.edu

Abstract

We consider the problem of active feature elicitation in which, given some examples with all the features (say, the full Electronic Health Record), and many examples with some of the features (say, demographics), the goal is to identify the set of examples on which more information (say, lab tests) need to be collected. The observation is that some set of features may be more expensive, personal or cumbersome to collect. We propose a classifier-independent, similarity metric-independent, general active learning approach which identifies examples that are dissimilar to the ones with the full set of data and acquire the complete set of features for these examples. Motivated by four real clinical tasks, our extensive evaluation demonstrates the effectiveness of this approach.

1 Introduction

Acquiring meaningful data for learning has long been a cherished goal of artificial intelligence, and is especially relevant in data scarce domains such as medicine. While there are plethora of data regarding several diseases, in some cases, it is crucial to obtain information that is particularly relevant to the learning task. The problem of choosing an example to obtain its class label has been addressed as active learning [Settles, 2012]. There have been several extensions of active learning that included presenting a set of features [Raghavan *et al.*, 2006; Druck *et al.*, 2009], or getting labels over clusters [Hofmann and Buhmann, 1998], or preferences [Odom and Natarajan, 2016] or in sequential decision making [Lopes *et al.*, 2009], to name a few.

Our problem is motivated by a different set of medical problems – that of recruiting patients for a clinical study. Consider the following scenario of collecting data (cognitive score and fMRI, both structural and functional) for an Alzheimer's study. Given a potentially large cohort size, the first step could be to simply collect the demographic information on everyone. Now, given a small amount of complete data from a related study, say the Alzheimer's Disease NeuroInitiative (ADNI), our goal is to recruit subjects who would provide the most information for learning a robust, generalized model. This scenario is highlighted in Figure 1. The top part shows the part of the data that is fully observed (potentially from a related study). The bottom left quadrant shows the observed features of the potential cohorts and the right quadrant is the data that needs to be collected for the most useful potential recruits. Given, the labels of the potential recruits, the goal is to identify the most informative cohorts that would aid the study.

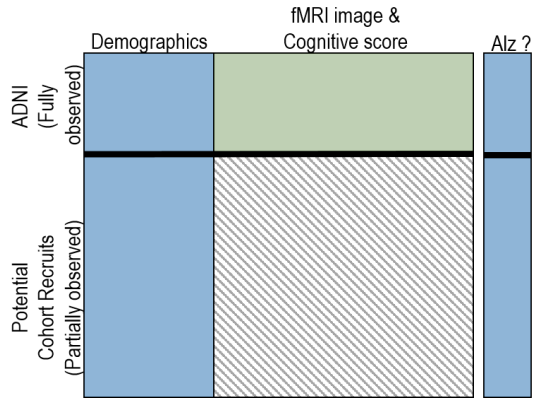

Figure 1: The active feature elicitation setting. The top part is the fully observed data and the bottom right (grey shaded area) is the unobserved feature set.

Inspired by the success of active learning methods, we define the problem of *active feature elicitation* (AFE) where the goal is to select the best set of examples on whom the missing features can be queried on to best improve the classifier performance. At a high level, our algorithm (alo called AFE) at each iteration, identifies examples that are most different from the current set of fully observed examples. These are then queried for the missing features, their feature-values are obtained and added to the training set. Then, the models are updated and the process is repeated until convergence. This is a *general-purpose* framework. Any distance metric that works well with the data and the model can be employed. So can a classifier that is capable of handling the specific intricacies of the data. Finally, the convergence criteria can be decided based on the domain.

We make a few key contributions. First, we identify and formally define the problem of actively acquiring features from a selected set of examples. Second, we show the potential of this approach in four real medical prediction tasks:

Alzheimer's from fMRI and cognitive score, Parkinson's from potential risk factors, rare diseases based on a survey questionnaire and predicting post-partum depression (PPD) in a non-clinical setting. Finally, we empirically demonstrate that AFE is particularly effective in recall while not sacrificing the overall performance in these four real tasks.

2 Related Work

Our work is closely related to active learning where the most informative examples are chosen based on a scoring metric to be labeled when learning with semi-supervised data. Active learning [Settles, 2012] relies on the fact that if an algorithm can only solicit labels of a limited number of examples, then it should choose them judiciously, since not all examples provide the same amount of information. Active learning has a long history of being successfully employed with a variety of classifiers such as logistic regression [Lewis and Gale, 1994; Lewis and Catlett, 1994], support vector machines [Tong and Koller, 2001b], Bayesian network learning [Tong and Koller, 2000; 2001a] and in sequential decision making tasks such as imitation learning [Judah *et al.*, 2014] and inverse reinforcement learning [Odom and Natarajan, 2016].

Active learning has also been used with missing features. Zheng et al. [2002], for instance, considered a setting where an imputation method was used to fill the incomplete feature subset and used scoring methods to acquire the most informative example for labeling. Melville et al. [2004] used uncertainty sampling to acquire the maximally informative unlabeled examples from the partially observed set of features during training time. There is also a different body of work where instances are chosen based on individual feature utilities [Melville *et al.*, 2005; Lizotte *et al.*, 2003].

Our work is heavily inspired by the work of Kanani and Melville [2008] on *active feature acquisition* that addressed a similar problem where a few examples with the full set of features are present while others are incomplete examples. Their work also scored these examples based on uncertainty sampling and then updated the model at prediction time. There are a few key differences between this work and ours. First, our model updates occur *during training* and *not during the test time*. Second, we return a single model on the best set of training data while their approach had two different models for testing. Our approach explicitly grows the set of training examples for a single model iteratively. Finally, they employed uncertainty sampling on the observed feature sets (which is a baseline in our approach), while we explicitly compute the distance between the two sets of points.

This work was later extended by Thahir et al. [2012] for protein-protein interaction prediction where an extra term was added to the utility function that explicitly computed the value of adding an example to the classifier set. While it is possible to compute the value of adding an example to the training set in our work, we will pursue this as a future research direction. The AFA framework was then later generalized and rigorously analyzed by Saar-Tsechansky et al. [2009] where even class labels can be considered to be missing and acquired. We assume that these labels are observed and that full sets of features need to be acquired. A key difference to this general direction is that both the observed set of examples and observed set of features are significantly smaller in our work compared to the general AFA setting which is clearly demonstrated in our experiments.

Bilgic and Getoor [2007] took a different approach to a similar task where they assumed different costs for misclassification and information acquisition. They proposed a probabilistic framework that explicitly modeled this dependency and developed an algorithm to identify the set of features that can be optimally identified. Using such a strategy for discovering sets of features that one could acquire for different sets of patients is an interesting direction. Feature elicitation is inspired by the preference framework of concept learning by Boutilier et al. [2009], where minimax regret is used for computing the utility of subjective features. The violated constraints are repeatedly added to the computation and can potentially make the problem harder to solve.

To summarize, our work is inspired by the contributions from several of these related works but **differs** in the motivation of collecting more features by identifying the right set of examples during training time to improve the model. As mentioned, we differ from active feature acquisition in both motivation and execution – we collect a large number of features from a small set of examples during training and use distances to calculate the most diverse set of examples.[1] The other important difference is in the number of observed features, which is assumed to be much smaller in our work. And our solution that explicitly computes the relationship between the observed and unobserved data is independent of the choices of classifiers and distance functions. One of the key assumptions that we make is that all unobserved features are collected for the selected examples and identifying the relevant set of features along the lines of Bilgic and Getoor is an exciting direction for future work.

3 Motivating Medical Tasks

We motivate active feature elicitation with four medical tasks:

1. **Parkinson's**: The Parkinson's Progression Markers Initiative (PPMI) is an observational study with the aim of identifying biomarkers that impact Parkinson's progression [Marek *et al.*, 2011]. The data can be divided broadly into four categories: imaging data, clinical data, bio-specimens and demographics. Among these data types, while other modalities are either costly or cumbersome to obtain, the total Montreal Cognitive Assessment Score (MoCA) is a standard measure that can be used to select subjects for whom information from other modalities can improve classifier performance significantly.
2. **Alzheimer's**: The Alzheimer's Disease NeuroIntiative (ADNI[2]) aims to test whether serial MRI, PET, biological markers, and clinical and neuro-psychological assessments can measure the progression of mild cognitive impairment and early Alzheimer's disease. Given a small number of subjects with all the measurements, demographic data can be used to select subjects for whom obtaining more information such as the cognitive MMScore and imaging data could maximize performance in predicting Alzheimer's progression.

[1] Genome sequencing would best exemplify such a scenario.

[2] `www.loni.ucla.edu/ADNI`

3. **Rare Diseases**: A recent work [MacLeod *et al.*, 2016] focused on predicting rare diseases from a survey questionnaire that consisted of questions in the following categories: demographics, technology use, disease information and healthcare provider information. The set of diseases in the study includes Ehlers Danlos Syndrome (23%), Wilson's Disease (21.9%), Kallmann's Syndrome (9.9%), etc. Demographics can be used to identify the future participants of the survey as this can avoid more personal questions such as technology use and the provider details along with the disease information itself.
4. **Post-Partum Depression**: This work collects demographic information along with several sensitive questions including relationship troubles, social support, economic status, infant behavior and the CDC questions to identify PPD in subjects outside the clinic [Natarajan *et al.*, 2017]. As with the earlier cases, demographics can be used to recruit the subjects on whom more sensitive information can be collected.

These varied medical tasks demonstrate the need for employing an active feature elicitation approach that allows for collecting relevant information in an effective manner. While the presented motivating tasks are medical, one could imagine the use of such approach in any domain where some features are either expensive or cumbersome to obtain.

4 Active Feature Elicitation

Let us denote the label of an example i as y^i, the set of fully observed features (i.e., the features that are observed for the entire data set) as $\mathbf{X}_o$, the set of features that are partially observed as $\mathbf{X}_u$, the set of fully observed examples set as $\mathbf{E}_o = \langle\langle \mathbf{X}_o^1; \mathbf{X}_u^1; y^1\rangle ::: \langle \mathbf{X}_o^k; \mathbf{X}_u^k; y^k\rangle\rangle$ and the set of partially observed examples as $\mathbf{E}_u = \langle\langle \mathbf{X}_o^1; y^1\rangle ::: \langle \mathbf{X}_o^{`}; y^{`}\rangle\rangle$. The learning problem in our setting can be defined as follows:

Given: A data set with $\mathbf{E}_o$ and $\mathbf{E}_u$.
To Do: Identify the best set of examples $\mathbf{E}_a \subset \mathbf{E}_u$ for which to obtain more information $\mathbf{X}_u$ such that the classifier performance improves.

In the above definitions, the notion of *best* and *improve* have been *intentionally* left vague. This definition allows for any notion of the best examples and improvement of the classifier. In our work, to be precise, we consider *best* to denote the set of the examples with maximal difference to the observed set and *performance* to be the log-likelihood of the classifier. The classifier we consider is the well understood gradient-boosting [Friedman, 2001].

Since our focus is on clinical (study) data, our hypotheses is that the best examples are chosen to obtain extra information from those that are significantly different from the remaining examples. In principle, any distance function could be used to determine the set of examples $\mathbf{E}_a$ from $\mathbf{E}_u$ that are significantly different from the ones in $\mathbf{E}_o$. We use the *mean Kullback-Leibler (KL) divergence* between every example $e_i = \langle \mathbf{X}_o^i; y^i\rangle$ in $\mathbf{E}_u$ and every example $e_j = \langle \mathbf{X}_o^j; \mathbf{X}_u^j; y^j\rangle$ in $\mathbf{E}_o$ to determine the set of examples in $\mathbf{E}_u$ that are different from the observed set $\mathbf{E}_o$. To compute this mean KL-divergence at every iteration t, we use the current models: $M_u^t = \mathsf{P}_t(y^i \mid \mathbf{X}_o^i)$ and $M_o^t = \mathsf{P}_t(y^j \mid \mathbf{X}_o^j; \mathbf{X}_u^j)$, learned on the two different sets of data. More precisely, we compute the mean distance of an example $\mathbf{X}_u^i$ from all the observed examples $\langle \mathbf{X}_o^j; \mathbf{X}_u^j\rangle$ as,

$$\mathsf{MD}_i = \frac{1}{|\mathbf{E}_o|} \sum_{j=1}^{j \in \mathbf{E}_o} D_{ij}; \qquad (1)$$

where the distance D_{ij} is the asymmetric KL divergence:

$$D_{ij} = \mathsf{KL}\ \mathsf{P}(y^i \mid \mathbf{X}_o^i) \parallel \mathsf{P}(y^j \mid \mathbf{X}_o^j; \mathbf{X}_u^j) \ : \qquad (2)$$

A natural question to ask is: what is the need for two different distributions, even if they are conditionals on the target. Note that, with the set of examples $\mathbf{E}_o$, all the features are assumed to be fully observed. Ignoring the informative features ($\mathbf{X}_u^j$) when computing the distances can lead to a loss of information and our experiments confirmed this. Hence, we employ the model learned over the full set of features for the fully observed example set $\mathbf{E}_o$, which is typically smaller than the unobserved set in the initial iterations.

Now that the distances have been computed, we next sort them to pick the n most distinct examples from $\mathbf{E}_u$. These n examples are queried for their missing features and are then added to the training set before the model is retrained. Note that at each iteration, the model $\mathsf{P}(y^j \mid \mathbf{X}_o^j; \mathbf{X}_u^j)$ is updated after the examples are appropriately chosen and queried. The model $\mathsf{P}(y^i \mid \mathbf{X}_o^i)$ remains unchanged because it is trained on $\mathbf{X}_o$ of the entire example set $\mathbf{E}_o \cup \mathbf{E}_u$. The process is repeated until convergence or a predetermined budget is realized.

This presents a generalized and unifying framework, which can be adapted in multiple ways:

1. As we discuss in Section 4.1, this formulation admits a large class of divergences and distance metrics for computing distances between examples in $\mathbf{E}_o$ and $\mathbf{E}_u$. One could imagine the use of other classifiers, kernels or learned metrics [Kunapuli and Shavlik, 2012] as well.
2. The gradient boosting classifier can be replaced with any classifier. Our framework is classifier-agnostic, allowing the user to select the best one for the task at hand.
3. Various convergence criteria can also be used. For instance, one could simply preset the number of iterations, or employ a tuning set to determine the change in performance from the previous iteration or compute the difference between scores from successive iterations. We employ this final strategy: computing the difference between log-likelihoods of the training data in successive iterations. If the difference is smaller than , we terminate the algorithm. One could also imagine reducing the number of queries at every iteration (i.e., successively reduce $n = \frac{n}{n+}$) such that the number of examples selected at each iteration naturally comes down to 0.

We present the algorithm for active feature elicitation in Algorithm 1. The AFE algorithm takes as input the set of fully labeled examples ($\mathbf{E}_o$), the set of partially labeled examples ($\mathbf{E}_u$), the number of active learning examples for each query step (n) and step size ( ). In *this algorithm*, sufficient decrease in step size ($\frac{n}{n+}$) is used as the stoppage

Algorithm 1 **A**ctive **F**eature **E**licitation

1: **function** ActiveFeatureElicitation($\mathbf{E}_o$; $\mathbf{E}_u$; n;)
2: $t = 0$. iteration counter
3: M_t = TrainInitialModel($\mathbf{E}_o$; $\mathbf{E}_u$; $\mathbf{X}_o$; $\mathbf{X}_u$)
4: **while** $n \geq 1$ **do** . while not converged
5: MD = **0** . initialize mean divergences
6: **for** $i = 1$ **to** $|\mathbf{E}_u|$ **do**
7: $\mathbf{D} = \mathbf{0}$. init divergence for unobserved ex. i
8: **for** $j = 1$ **to** $|\mathbf{E}_o|$ **do**
9: D_j = ComputeDistance(E_i; E_j; M_t)
10: **end for**
11: $\mathrm{MD}_i = \frac{\sum_{j=1}^{|\mathbf{E}_o|} D_j}{j\mathbf{E}_o j}$. average distance
12: **end for**
13: $\mathbf{E}_u^q$ = GetTopN(MD)
14: . n most divergent partially-observed examples
15: $\mathbf{E}_o^q$ = AppendNewFeature($\mathbf{E}_u^q$)
16: . actively query to elicit missing features
17: $\mathbf{E}_o = \mathbf{E}_o \cup \mathbf{E}_o^q$. add queried to observed
18: $\mathbf{E}_u = \mathbf{E}_u \setminus \mathbf{E}_u^q$. remove queried from unobs.
19: M_t = UpdateModel($\mathbf{E}_o$; $\mathbf{E}_u$; $\mathbf{X}_o$; $\mathbf{X}_u$)
20: . retrain or update classifier
21: $n = \frac{n}{n^+}$. check convergence/update budget
22: **end while**
return TrainFinalModel($\mathbf{E}_o$)
23: **end function**

criterion (lines 4, 21) as an example. This can be replaced by other task-relevant budgets or convergence criteria as discussed previously.

After initializing mean distances of each unlabeled example, AFE iterates through every partially-labeled example in $\mathbf{E}_u$, and computes the mean distance to all the fully labeled examples in $\mathbf{E}_o$ based on the divergence between the respective current models. The n-most divergent (dissimilar) examples are selected and features are actively obtained for these examples (AppendNewFeature). These examples are then added to $\mathbf{E}_o$ and removed from $\mathbf{E}_u$. A new model can be trained (or updated depending on choice of classifier), and the process is repeated. Note that M_t consists of two classifiers – one trained on $\mathbf{E}_o$, which contains new examples provided by the user after active feature elicitation with all the features, and the other trained on entire data $\mathbf{E}_o \cup \mathbf{E}_u$. After convergence, $\mathbf{E}_o$ has the full set of training examples.

In our experiments, we employ gradient boosting as the classifier, KL-divergence to identify the unobserved examples whose features we would like to acquire, and the difference in average log-likelihoods between two iterations as our convergence criterion.

4.1 Other Model Divergences

The KL divergence is a special case of the Csiszár f-divergence [Csiszár, 1967], which is a *generalized measure* for the difference between two probability distributions, in our case $\mathrm{P}(y^i \mid \mathbf{X}_o^i)$ and $\mathrm{P}(y^i \mid \mathbf{X}_o^i; \mathbf{X}_u^i)$. $D_f(P \parallel Q) = \int f(dP/dQ)\, dQ$. Generally, given two distributions P and Q over some space , for a convex function f (with $f(1) = 0$), the divergence of P from Q is defined as $D_f(P \parallel Q) = \int f(dP/dQ)\, dQ$. Such f-divergences satisfy non-negativity, monotonicity and convexity, though they are *not always symmetric*. Several well-known distribution distance measures are special cases of the f-divergence and are shown in Table 1. For example, the ^{2}-divergence might be well-suited for histogram data [Kedem *et al.*, 2012], while the Hellinger distance might benefit applications with highly-skewed data [Cieslak *et al.*, 2012].

Recently it was shown that families of divergences including the - and -divergence are also a special cases of the f-divergence [Cichocki and Amari, 2010]. The latter includes generalizations of measures such as the Euclidean distance and the Itakura-Saito distance, which are appropriate for unsupervised and semi-supervised learning problems.[3] We consider the usefulness of various divergences to different machine learning problem types and applications in future work.

Divergence	$f(x)$	$D_f(\mathbf{p} \parallel \mathbf{q})$
KL-divergence	$x \log x$	$\sum_i p_i \log \frac{p_i}{q_i}$
Hellinger distance	$(\sqrt{x} - 1)^2$	$\frac{1}{\sqrt{2}} \lVert \mathbf{p} - \mathbf{q} \rVert_2$
Total variation	$\frac{1}{2}\lvert x - 1 \rvert$	$\frac{1}{2}\lVert \mathbf{p} - \mathbf{q} \rVert_1$
^{2}-divergence	$(x-1)^2$	$\sum_i \frac{p_i^2}{q_i} - 1$

Table 1: Several well-known f-divergences for discrete distributions (where $\mathbf{p}$ and $\mathbf{q}$ are vector representations of the distributions) are shown. Their continuous extensions can be obtained by replacing the sums with integrals over the support of the distributions.

4.2 Multi-class and Other Extensions

As our approach is algorithm- and divergence-agnostic, it can be seamlessly extended to multi-class settings. As long as the underlying classification algorithm can produce (multinomial) distributions over the label space: $\mathbf{p} = \mathrm{P}(y^i \mid \mathbf{X}_o^i)$ and $\mathbf{q} = \mathrm{P}(y^i \mid \mathbf{X}_o^i; \mathbf{X}_u^i)$, we can use any model divergence discussed in Section 4.1.

There are several other possible extensions of the proposed approach. First is the necessity to move beyond active learning; while standard methods acquire a label for each example, in many situations where the goal is to understand why an event happens (such as clinical studies), it is necessary to obtain more tests/features. Also, given that the original model is learned from a small set of features, the model will not be necessarily generalizable. Second, it is possible that some specific set of features are the most informative for specific example. For instance, some subjects' predictions will benefit from some lab test while a different test is a better indicator for someone else. Extending our framework to handle these different types of examples/feature combinations is outside the scope and is an interesting future direction.

5 Empirical Evaluation

We now present evaluation results on one standard UCI data set (PIMA, [Smith *et al.*, 1988]) and four real medical tasks to demonstrate the efficacy of our approach. It must be mentioned clearly that while healthcare is one domain, the data

[3] See Deza & Deza [2013] for other useful distance functions.

sets are varied: from online behavior to images to risk factors to survey. The goal is to demonstrate the versatility of the approach with real problems. It must also be noted that while healthcare is considered in this work, the ideas are not limited to this domain and any problem where a small set of data is fully observed and the rest are partially observed can render itself as an useful domain for the proposed approach.

1. **Parkinson's prediction from clinical study:** The task is to predict the occurrence of Parkinson's disease from different modalities. We focus on a smaller set of features, primarily motor and non-motor assessments resulting in a set of 37 attributes including the class label. The observed feature is the MoCA test result, while the other 35 motor scores are treated as unobserved.
2. **Alzheimer's prediction from ADNI:** We assume that demographics are observed, while cognitive score (MMScore) and fMRI image features are unobserved. We use the AAL Atlas[4,5] to segment the image into 108 Regions of Interest (RoIs), and for each RoI, we derive their summary attributes: white matter, cerebral spinal fluid, and gray matter intensities along with regional variance, size and spread.

   While the original data set has three classes: Alzheimer's (AD), Cognitively Normal (CN) and Mildly Cognitively Impaired (MCI), we consider the binary task of predicting AD vs. the rest. The presence of MCI subjects makes this particular task challenging, yet interesting; this is because these subjects may or may not end up having Alzheimer's eventually. Identifying the right set of subjects to target for feature elicitation can considerably improve classifier performance, as we show below.
3. **Rare disease prediction from self-reported survey data:** The task is to predict if a subject has a rare disease [MacLeod *et al.*, 2016]; by definition, a rare disease is hard to diagnose and affects less than 10% of the world's population. The data for this prediction task arises from survey questionnaires and we assume that demographic data are fully observed. Other survey answers concerning technology use, disease information and healthcare details are treated as unobserved.
4. **Post-partum depression prediction from online questionnaire data:** Recently, Natarajan et al. [2017] employed online questionnaires to predict PPD from demographics, social support, relevant medical history, child birth issues and screening data. We assume that demographics are observed and are used to select subjects on whom the rest of the data can be collected to learn the model.

We also test our algorithm on the well-studied PIMA Indians Diabetes data to demonstrate generality. Table 2 shows the details of these domains; a common characteristic across all domains is *class imbalance* where it is important that the most informative subjects are added to the training set.

Evaluation Methodology: All data sets are partitioned as 80% train and 20% test. Results are averaged over 10 runs with a fixed test set. At each active learning step, we solicit 5 new data points until convergence. As mentioned earlier, Friedman's [2001] gradient-boosting was employed as the underlying classifier with the same settings across all methods and KL-divergence is our distance metric. We compare three different evaluation metrics: *recall* (to measure the clinically relevant *sensitivity*), *F1-score*, and geometric mean of sensitivity and specificity (*gmean*), that provide a reasonably robust evaluation in the presence of class imbalance. We considered AUCROC but as pointed out by Davis and Goadrich [2006], for severe class imbalance data sets, this is not ideal and hence we settled on our metrics.

[4]http://prefrontal.org/blog/2008/05/brain-art-aal-patchwork.

[5]http://www.slicer.org.

Data set	# Pos	# Neg	# Features		# Examples	
			FO	PO	FO	PO
PPMI	554	919	1	35	5	1174
ADNI	76	260	6	69	10	294
Rare Disease	87	174	6	63	10	198
PPD	38	115	8	33	6	147
PIMA	268	500	4	4	10	681

Table 2: Data set details. FO - Fully observed, PO - Partially observed. # Pos is number of positive examples, # Neg is number of negative examples, # Features (FO) is the number of features in the fully observed feature set, # Features (PO) is number of features in the partially observed feature set, # Examples (FO) is number of examples in the fully observed example set, # Examples (PO) is number of examples in the partially observed example set.

Baselines: In addition to the proposed AFE approach, we considered three other baselines: (1) Randomly choosing points to query which can potentially yield strong results when closer to convergence. This method is denoted as RND; (2) We also used uncertainty sampling on partially observed example set using only the fully observed features. The top 5 instances that have the highest entropy were then queried for unobserved features and added to the training set. This is denoted as USObs; (3) In the third approach, we imputed all missing features using mode as the feature value; uncertainty sampling is then employed by computing the entropy on the full feature set, following which the top 5 values were chosen for querying. This baseline is denoted as USAll. Other active-learning baselines can be considered (such as min-max), but these generally tend to be prohibitively expensive in large feature spaces.

Results: We aim to answer the following questions:

Q1: Does AFE perform better than choosing the examples randomly for active learning?

Q2: Does AFE outperform other active baselines that employ entropy to choose examples?

Q3: Does AFE robustly handle imbalanced data?

Q4: Can AFE be useful for a semi-supervised formalism?

Q5: Is AFE faithful to the motivation and is the proposed solution effective in modeling clinical data?

The results across the five domains and all the three metrics are presented in Figure 2. It can be observed that AFE outperforms RND on all domains across all metrics, specifically in the recall in the first few iterations in 4 of the 5 domains where the effect of choosing the most informative set of examples can have the maximal impact on the classifier performance. As expected, the variance in recall due to random selection of examples is high. This can be seen in the left column, where the performance does not increase steadily in all

domains. Similar observations can be made for the geometric mean. This allows us to answer **Q1** affirmatively.

Observe that as we add more informative examples, the performance improves significantly for the proposed AFE approach over the rest of the classifiers. This demonstrates that the gains are not necessarily in the beginning alone. Adding more useful examples can construct a more robust training set that can lead the classifier to a superior performance. Note that one of the reasons that we are averaging over 10 runs is to alleviate/minimize the effect of sampling bias in construction of training set (particularly in the initial samples). As expected, the variance is initially on the higher side indicating the effect of sampling bias on smaller training sets but it decreases as more examples are added. However, this effect is minimal for AFE that chooses good training examples compared to other methods. Understanding the effect of the initial choice of samples on the performance of the classifier is itself an interesting future research direction.

Similar results can be observed when comparing AFE to USObs and USAll in that AFE consistently outperforms the two active learning baselines across all the domains in all the metrics. In general, the use of only observed features still appear better than using imputation (mode) to fill in the missing values and then use them for entropy. We speculate that the use of better imputation techniques may improve the performance. However, the difference between AFE and USAll in recall and gmean across all domains is statistically significant in several iterations. This suggest that other imputation techniques may marginally improve the performance, but it appears that the use of imputation may not influence the final performance and a good selection of examples is necessary. This strongly answers **Q2** affirmatively.

Another natural question to ask is how the variance in performance of the different methods tend to behave across all data sets? It was generally observed that the AFE had the smallest variance both in the selection of the first few examples and in the last few iterations of the algorithms. The variance of AFE across all data sets was at least half the variance of the Random baseline (RND) on an average in the first few iterations and as low as 10% of random selection's variance in later iterations. This is also consistent across all metrics. When compared to USObs and USAll, in general, the average variance of AFE was lower across all metrics and all data sets. While AFE's variance is significantly better than the random selection, the differences to the uncertainty methods, while better, are not necessarily significant.

As shown in Table 2, all domains are imbalanced. The class prior ratio is particularly skewed in the case of ADNI and PPD data sets. In these two domains, it can be seen clearly that the proposed approach achieves a recall of over 0:8 after just a few early iterations. This demonstrates that AFE can identify the most important examples that allow for increasing the clinically relevant sensitivity effectively enabling us to answer **Q3** positively as well. Our results in all domains show that this method achieves high recall without significantly sacrificing precision making it an ideal choice for semi-supervised imbalanced data sets (**Q4**). The intuition here is that because AFE obtains high recall across all data sets, in domains where many examples are labelled, it provides an opportunity for selecting the right sets of examples that can be labelled or extended with more features. Although the proposed approach is not a semi-supervised one, it provides an opportunity for developing methods that can learn from partially labelled data.

Our original motivation was to identify the set of subjects on whom to perform specific lab tests, given some basic information about the potential recruits. Our algorithm does not use any extra information about the potential recruits beyond the observed features and their labels (whether they are cases or controls) for identifying the best set of subjects to elicit more information about. Secondly, we do not make any assumptions about the underlying distributions of the data or the classifier employed while learning. Finally, we make effective use of the fully observed data (from possibly a related study) by using them to compute the distance with the potential cohorts. This clearly demonstrates that the AFE is indeed faithful to the original goals and the results clearly show the efficacy of AFE thus answering (**Q5**) in the affirmative.

6 Conclusion

We considered the problem of eliciting new sets of features based on a small amount of fully observed data. We address this problem specifically in the context of medical domains with severe class imbalance. Our proposed active feature elicitation approach computes the similarity between the potentially interesting examples with the fully observed examples and chooses the most significantly different examples to elicit the feature information. These are then added to the full observed set and the process continues until convergence. Experiments on four real high-impact medical tasks clearly demonstrate the effectiveness and efficiency of the proposed approach. Our approach has a few salient features – it is domain, model and distance, i.e., representation agnostic in that any reasonable classifier and a compatible distance metric for a specific domain can be employed in a plug and play manner.

We currently elicit all missing features for every chosen example at every iteration. One could extend this work by identifying sets of features that are most informative for every example (i.e., the most relevant lab test for each subject) along the lines of the work of Krishnapuram et al. [2005] that addressed a similar problem using multi-view, co-training setting which could allow for the realization of the vision of personalized medicine. Another interesting future research direction could be to identify groups of examples (subpopulations) that would provide the most information at each iteration. Extending the work to handle fully relational/graph data is another possible direction. Finally, rigorous evaluation of the approach on more clinical data sets can yield more interesting insights into the proposed approach.

7 Acknowledgements

The authors acknowledge the support of Indiana University's Precision Health Initiative. SN and GK also gratefully acknowledge the National Science Foundation grant no. IIS-1343940. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the Precition Health Initiative or the United States government.

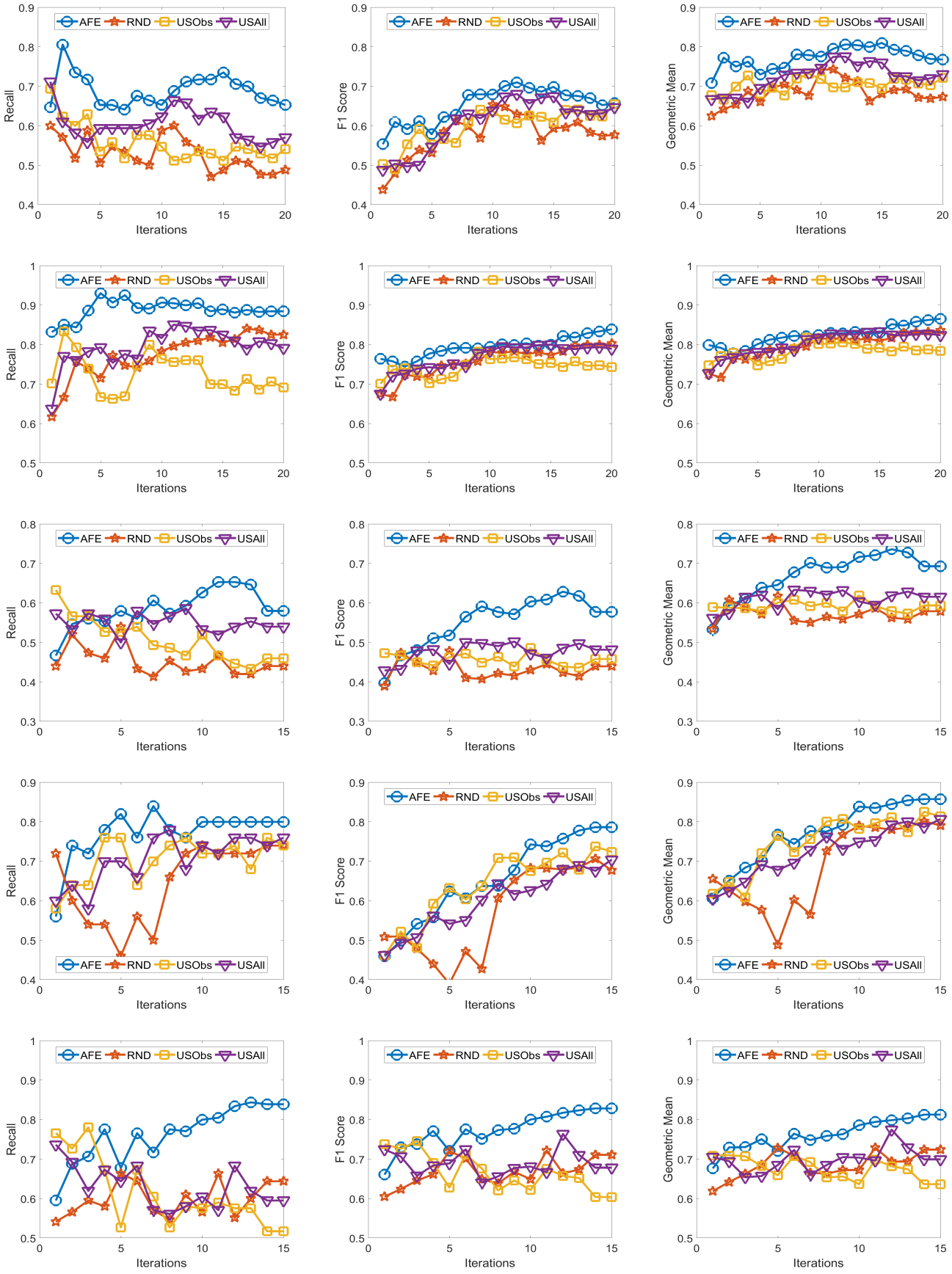

Figure 2: Recall (left), F1 (middle), g-mean (right) for (from top to bottom) ADNI, PPMI, Rare Disease, PPD and PIMA. Each iteration corresponds to acquiring the 5 best examples.

References

[Bilgic and Getoor, 2007] M. Bilgic and L. Getoor. VOILA: efficient feature-value acquisition for classification. NCAI, pages 1225–1230, 2007.

[Boutilier *et al.*, 2009] C. Boutilier, K. Regan, and P. Viappiani. Online feature elicitation in interactive optimization. ICML, pages 73–80, 2009.

[Cichocki and Amari, 2010] A. Cichocki and S. Amari. Families of alpha- beta- and gamma- divergences: flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[Cieslak *et al.*, 2012] D. A. Cieslak, T. R. Hoens, et al. Hellinger distance decision trees are robust and skew-insensitive. *Data Min Knowl Discov*, 24(1):136–158, 2012.

[Csiszár, 1967] I. Csiszár. Information measures of difference of probability distributions and indirect observations. *Studia Sci Math Hungar*, 2:299–318, 1967.

[Davis and Goadrich, 2006] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. ICML, pages 233–240, 2006.

[Deza and Deza, 2013] M. M. Deza and E. Deza. *Encyclopedia of distances*. Springer, 2013.

[Druck *et al.*, 2009] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. EMNLP, pages 81–90, 2009.

[Friedman, 2001] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann Stat*, 29(5):1189–1232, 2001.

[Hofmann and Buhmann, 1998] T. Hofmann and J. M. Buhmann. Active data clustering. NIPS, pages 528–534, 1998.

[Judah *et al.*, 2014] K. Judah, A. Fern, et al. Imitation learning with demonstrations and shaping rewards. AAAI, pages 1890–1896, 2014.

[Kanani and Melville, 2008] P. Kanani and P. Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Workshop on Cost Sensitive Learning at NIPS*, 2008.

[Kedem *et al.*, 2012] D. Kedem, S. Tyree, et al. Non-linear metric learning. NIPS, pages 2573–2581. 2012.

[Krishnapuram *et al.*, 2005] B. Krishnapuram, D. Williams, et al. Active learning of features and labels. *Workshop on Learning with Multiple Views at ICML*, 2005.

[Kunapuli and Shavlik, 2012] G. Kunapuli and J. Shavlik. Mirror descent for metric learning: a unified approach. ECML PKDD, pages 859–874, 2012.

[Lewis and Catlett, 1994] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. ICML, pages 148–156, 1994.

[Lewis and Gale, 1994] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. SIGIR, pages 3–12, 1994.

[Lizotte *et al.*, 2003] D. J. Lizotte, O. Madani, and R. Greiner. Budgeted learning of naive-bayes classifiers. UAI, pages 378–385, 2003.

[Lopes *et al.*, 2009] M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. ECML PKDD, pages 31–46, 2009.

[MacLeod *et al.*, 2016] H. MacLeod, S. Yang, et al. Identifying rare diseases from behavioural data: a machine learning approach. CHASE, pages 130–139, 2016.

[Marek *et al.*, 2011] K. Marek, D. Jennings, et al. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol*, 95(4):629–635, 2011.

[Melville *et al.*, 2004] P. Melville, M. Saar-Tsechansky, et al. Active feature-value acquisition for classifier induction. ICDM, pages 483–486, 2004.

[Melville *et al.*, 2005] P. Melville, M. Saar-Tsechansky, et al. An expected utility approach to active feature-value acquisition. ICDM, pages 745–748, 2005.

[Natarajan *et al.*, 2017] S. Natarajan, A. Prabhakar, et al. Boosting for postpartum depression prediction. CHASE, pages 232–240, 2017.

[Odom and Natarajan, 2016] P. Odom and S. Natarajan. Active advice seeking for inverse reinforcement learning. AAMAS, pages 512–520, 2016.

[Raghavan *et al.*, 2006] H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on features and instances. *J Mach Learn Res*, 7(Aug):1655–1686, 2006.

[Saar-Tsechansky *et al.*, 2009] M. Saar-Tsechansky, P. Melville, and F. Provost. Active feature-value acquisition. *Manag Sci*, 55(4), 2009.

[Settles, 2012] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1), 2012.

[Smith *et al.*, 1988] J. W. Smith, J. E. Everhart, et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc Annu Symp Comput Appl Med Care*, pages 261–265, 1988.

[Thahir *et al.*, 2012] M. Thahir, T. Sharma, and M. K. Ganapathiraju. An efficient heuristic method for active feature acquisition and its application to protein-protein interaction prediction. *BMC Proc*, 6(Suppl 7):S2, 2012.

[Tong and Koller, 2000] S. Tong and D. Koller. Active learning for parameter estimation in Bayesian networks. NIPS, pages 647–653, 2000.

[Tong and Koller, 2001a] S. Tong and D. Koller. Active learning for structure in Bayesian networks. IJCAI, pages 863–869, 2001.

[Tong and Koller, 2001b] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J Mach Learn Res*, 2(Nov):45–66, 2001.

[Zheng and Padmanabhan, 2002] Z. Zheng and B. Padmanabhan. On active learning for data acquisition. ICDM, pages 562–569, 2002.