A Clustering based Selection Framework for Cost Aware and Test-time Feature Elicitation

Srijita Das
The University of Texas at Dallas
Srijita.Das@utdallas.edu

Rishabh Iyer
The University of Texas at Dallas
Rishabh.Iyer@utdallas.edu

Sriraam Natarajan
The University of Texas at Dallas
Sriraam.Natarajan@utdallas.edu

ABSTRACT

Most learning algorithms are optimized with generalization and predictive performance as the goal. However, in most real-world machine learning applications, obtaining features at test time can incur a cost. For example, in clinical tasks, acquiring certain features such as FMRI or certain lab tests for patients can be expensive, while other features like patient demography or history are easily obtained and do not have a cost involved. Motivated by this, we address the problem of test-time elicitation of features. We formulate the problem of cost-aware feature elicitation as an optimization problem with trade-off between performance and feature acquisition cost. We assume that the cost of the features has already been paid in obtaining the training data. We propose a *Clustering based **C**ost **A**ware **T**est-time Feature **E**licitation* (CATE) algorithm, which can select the relevant feature set given the observed attributes of the test instance. Our experiments on four real-world tasks demonstrate the efficacy and effectiveness of our proposed approach in both cost and performance.

CCS CONCEPTS

• **Supervised learning** → **Budgeted learning**; *Feature selection*;
• **Applications** → Healthcare.

KEYWORDS

cost sensitive learning, supervised learning, classification

ACM Reference Format:
Srijita Das, Rishabh Iyer, and Sriraam Natarajan. 2021. A Clustering based Selection Framework for Cost Aware and Test-time Feature Elicitation. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3430984.3431008

1 INTRODUCTION

In supervised classification setting, every instance has a fixed feature vector and a discriminative function is learnt on such a fixed-length feature vector and it's corresponding class variable. However, several practical problems such as healthcare, networks, recommender systems, surveys [24, 25] etc. include an *associated feature acquisition* cost. In such domains, there is a cost budget for specific feature subsets since acquiring all the features for all the instances can become prohibitively expensive. Consequently, many cost sensitive classifier models [3, 13, 29] have been proposed to incorporate the cost of acquisition into the model objective during training and prediction.

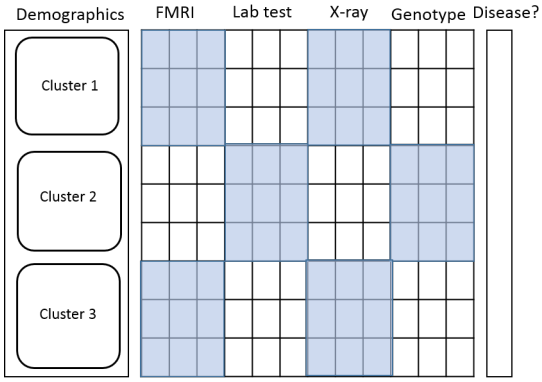


Figure 1: CATE intuition. The assumption is that certain features like demographics are easily available, whereas features like Lab tests, image features etc. are expensive and incurs cost. Blue shaded regions refers to different elicitable features for different clusters.

Our problem is motivated by such a cost-aware setting where the assumption is that prediction time features have an acquisition cost and adheres to a strict budget. As an example, consider the diagnosis of a patient. In such cases, demographic information (age, gender etc.) are easily available at zero costs. Some specific lab tests including imaging and genotyping on the other hand, can be quite expensive. In such tasks, during learning it is essential that the classifier considers such *query-time* costs. The intuition of this work (see Figure 1) is that different patients, depending on their history, ethnicity, age and gender, may require different tests for reasonably accurate prediction. We build on the intuition that given certain observed features like one's demographic details, the most important features for a patient depends on the important features for similar patients. Consequently, we identify similar data points in the observed feature space and determine the important feature subsets of these similar instances by employing a greedy information theoretic feature selector objective. Moreover, our approach is applicable in *test-time*/deployment as it mirrors the real-world deployment of ML methods.

We make a few key contributions: (1) We consider the problem of query-time cost-aware learning and We develop a clustering-based framework that identifies the closest set of examples in the training set to determine the best feature sets to query at deployment time. (2) We present our algorithm **CATE** that performs cost-aware test

Bangalore '21, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01…$15.00
https://doi.org/10.1145/3430984.3431008

time feature elicitation. (3) Finally, we demonstrate empirically the efficiency and effectiveness of our approach.

2 RELATED WORK

Prior research on cost-sensitive feature selection and learning can be categorized into the following four broad approaches.

Tree-based budgeted learning: Prediction time elicitation of features under a cost budget has been widely studied using tree-based models [8, 21, 22, 31–33] by adding a cost term to the objective function in either decision trees or ensemble methods like gradient-boosted trees. Broadly, these methods aim to build an adaptive and complex decision boundary by considering trade-off between performance and test-time feature acquisition cost. While in principle, our motivations are similar, our work is not restricted to tree-based models. Instead we find local feature subsets using an information theoretic feature selector for different clusters of training instance built in a lower dimensional space.

Adaptive classification and dynamic feature discovery: Our work draws inspiration from Nan et al.'s work [20] where they learn a high performance costly model and approximate the model's performance adaptively by building a low cost model using a gating function which decides which model to use for specific training instances. This adaptive switching between low and high cost model takes care of the trade-off between cost and performance. Our method is different from theirs as we do not maintain a high cost model which is costly to build and difficult to decide. We refine the parameters of a single low cost model by incorporating a cost penalty in the feature selector and model objective. Our work is also along the direction of Nan et al.'s work [23] where they select varying feature subsets for test instance using neighbourhood information of the training data. While calculating the neighborhood information from training data is similar to building clusters in our approach, the training neighborhood for our method is on *only* the observed zero cost feature space. In addition, we incorporate the neighborhood information in the training time. Ma et al. [15] take a different approach of dynamic discovery of features based on generative modeling.

Feature elicitation using Reinforcement learning: Sequential decision making has been used [6, 14, 27] to model the test time feature elicitation by learning an optimal policy. Action in such setting is whether to acquire a particular feature of a single instance or all the instances and optimal policy refers to the order in which these features are acquired. Our work aligns with the work of Shim et al. [30] where they jointly train a classifier and RL agent together. The key differences lie in the nature of the solution – RL vs supervised joint learning.

Active Feature Acquisition: Our problem set-up is also inspired by work along active feature acquisition [18, 19, 24, 28, 34] where certain feature subsets are observed and rest are acquired at a cost. These methods acquire new features during training time and typically use active learning to seek informative instances at every iteration, we elicit features for test instances. *All the training instances in our work are fully observed* and the assumption is that the feature acquisition cost has already been paid during training. Our problem set up is similar to Kanani et al. [10] in that they assume partial test instances, however their problem is that of instance acquisition where the acquired feature subset is fixed.

Our contributions: Although the problem of prediction time feature elicitation has been explored in literature from various directions and with various assumptions, we develop an intuitive solution by formulating the problem in a **two step optimization framework**. *We incorporate acquisition cost in both the feature selector and model objectives to balance the performance and cost trade-off.* The problem set up is naturally applicable in real world health care and other domains where the knowledge of the observed features also needs to be accounted while selecting the elicitable features. We formalize the problem as a joint optimization problem of selecting the best feature subset for similar data points and optimizing the loss function using the important feature subsets.

3 CLUSTERING BASED FRAMEWORK FOR TEST TIME FEATURE ELICITATION

Notations: An upper-case letter in bold represents a set (e.g. $\mathbf{P}$) and P_i denotes the i^{th} element of $\mathbf{P}$. $\mathbf{E}$ denotes the set of training instances $(x_1, x_2 \cdots, x_n)$ where $x_i \in \mathbf{R}^d$ is the feature vector of each instance, $\mathbf{Y}$ denotes the set of labels for the training points. The whole feature set is denoted by $\mathcal{F}$ which is partitioned into two feature subsets; $\mathcal{O}$ and $\mathcal{E}$. $\mathbf{E}$ is partitioned into two sets: $\mathbf{E}_{\mathcal{O}}$ which denotes the set of instances with the feature set $\mathcal{O}$ and $\mathbf{E}_{\mathcal{E}}$ denotes the set of instances with the feature set $\mathcal{E}$. In general, $\mathbf{E}_{\mathbf{S}}$ where $\mathbf{S} \subseteq \mathcal{F}$ refers to set of instances restricted to $\mathbf{S}$. A cluster i is denoted by c_i. $\mathbf{E}^{c_i}$ and $\mathbf{Y}^{c_i}$ denote the set of instances and labels belonging to a cluster c_i. There is an associated cost vector $\mathbf{M}$ where $\mathbf{M} \in \mathbf{R}^d$ is the feature acquisition cost of $\mathcal{F}$. The final training model is denoted by G. A budget is denoted by B and can either be a budget on features or cost.

3.1 Problem setup

Given: A dataset with instances $\mathbf{E} = (x_1, \cdots, x_n)$, labels $\mathbf{Y} = (y_1, \cdots, y_n)$, cost vector $\mathbf{M}$ and a budget B.

Objective: Learn a discriminative model $\mathbf{G}$ that is aware of the feature costs and can balance the trade-off between feature acquisition cost and model performance and make predictions on partially observed test instances.

We make an *additional assumption* – there is a subset of features which have 0 cost. These could be, for example, demographic features (e.g. age, gender, etc) in a medical domain which are easily available as compared to other features. In other words, we can partition the feature set $\mathcal{F} = \mathcal{O} \cup \mathcal{E}$ where $\mathcal{O}$ are the *zero cost observed features* and $\mathcal{E}$ are the *elicitable features* which can be acquired at a cost. We also assume that the training data is completely available with all features (i.e. the cost for all the features has already been paid while training). We will relax this assumption in future work.

Our goal is to use the observed features to find similar instances in the training set (for examples, similar age group, gender, ethnicity etc.) and identify the important feature subset based on these instances. To this effect, we first *cluster* the training data points based on observed features (e.g. age, gender, ethnicity). Next, for each cluster, we identify the most important features under a cost constraint. We then train the model by using the appropriate feature set for the specific cluster – this is in contrast to general feature

selection algorithms which select a fixed feature set for the entire dataset. At prediction time, for every instance, since only the observed feature set O is available, we seek an appropriate subset of the elicitable features and make predictions based on the trained model.

3.2 Proposed solution

As a first step, we cluster all the training instances based on only the observed zero cost feature set O. The key intuition is that instances with similar observable features will potentially have similar most informative elicitable features. For example, in a clinical task, the choice of one of blood test, CT-scan or MRI could depend on factors such as age, gender, ethnicity and whether patients with similar demographic features had requested these tests, an intuition that we formalize here.

Our model consists of a parameterized feature selector module $F_i(\mathrm{S}, \alpha)$ which is a restriction of the global function $F(\mathrm{S}, \alpha)$ to the i^{th} cluster c_i build using the feature set O. $F_i(\mathrm{S}, \alpha)$ outputs a subset S of most important features for the discriminative task. The feature selection model is based on an *information theoretic framework* (see more details below) and is augmented with the feature cost to account for the trade off between model performance and acquisition cost at test-time. The feature subset S from the feature selector module is used to update the parameters of the model. The optimization framework for our proposed approach is shown in Figure 2.

Information theoretic feature selector model: The feature selector module selects the best subset of features for each cluster of training data (build on observed feature set) based on an information theoretic objective score. Since at test time, the elicitable feature subset $\mathcal{E}$ (which will be needed by a feature selection algorithm if we were to run it on the test instances) is not known, we propose to use the closest set of instances in the training data to the current instance, and find the elicitable feature subset from that. Since we assume that the training data has already been elicited, all the features are fully observed in the training data. We compute the distance based on only the observed feature set O and cluster the training data into m clusters $c_1, c_2, \cdots c_m$. Next, we use the Minimum-Redundancy-Maximum Relevance (MRMR) feature Selection paradigm [2, 26]. We denote parameters $[\alpha^1_{c_i}, \alpha^2_{c_i}, \alpha^3_{c_i}, \alpha^4_{c_i}]$ as parameters of a particular cluster c_i. Let $X_1, X_2, \cdots, X_n$ be the elicitable features with $n = |\mathcal{E}|$. The feature selection module is a function of the parameters of the cluster to which a set of instances belong and is defined as:

$$
\begin{aligned}
F_i(\mathrm{S}, \alpha_{c_i}) \quad &= \underbrace{\sum_{p \in \mathrm{S}} \alpha^1_{c_i} I(X_p; \mathbf{Y})}_{\textbf{max. relevance}} \\
&\underbrace{- \sum_{p \in \mathrm{S}} \sum_{j \in \mathrm{S}} \left(\alpha^2_{c_i} I(X_j; X_p) - \alpha^3_{c_i} I(X_p; X_j | \mathbf{Y}) \right)}_{\textbf{min. redundancy}} \\
&\underbrace{- \alpha^4_{c_i} \sum_{p \in \mathrm{S}} c(X_p)}_{\textbf{cost penalty}}
\end{aligned} \tag{1}
$$

where $I(X; Y)$ is the mutual information between the random variables X (the feature) and Y (label). In the above equation, the feature subset S is grown greedily using a greedy optimization strategy maximizing the above objective function. In equation 1, X_p denotes a single feature from the elicitable set $\mathcal{E}$ that is considered for current evaluation based on the subset **S** grown so far. The first term in Equation 1 refers to the **mutual information** between each feature from the elicitable set $\mathcal{E}$ considered for evaluation and the class variable Y. In a discriminative task, this value should be maximized. The second term is the **pairwise mutual information** between each feature to be evaluated (X_p) and the features already added to the feature subset S. This value needs to be minimized for selecting informative features as correlated features give redundant information about the target. The third term is called the **conditional redundancy** [2] and this term needs to be maximized. The last term adds the **penalty for cost of every feature** and ensures the right trade-off between cost, relevance and redundancy. We do not learn the parameters α_{c_i} for each cluster, instead fix these parameters to 1. We leave the learning of the feature selector module to future work.

In the problem setup, since the zero cost feature subset is always present, we always consider the observed feature subset O in addition to the most important feature subset as returned by the feature selector objective. We also account for the knowledge of the observed features while growing the informative feature subset through greedy optimization. Specifically, while calculating the pairwise mutual information between the features and the conditional redundancy term (second and third term of equation 1), we also evaluate the mutual information of the candidate features with the observed features. Our method is robust to cases where the 0 cost features are not correlated with the target because in such scenario, the feature selector model will identify features from the elicitable set that are correlated with the target (max relevance).

Optimization Problem: The cost-aware feature selector $F_i(\mathrm{S}, \alpha)$ for a given set of instance $\mathbf{E}^{c_i}$ belonging to a specific cluster c_i solves the following optimization problem:

$$
\mathrm{S}^i_\alpha = \mathrm{argmax}_{\mathrm{S} \subseteq \mathcal{E}} F_i(\mathrm{S}, \alpha_{c_i}) \tag{2}
$$

For a given set of instances $(\mathbf{E}^{c_i}, \mathbf{Y}^{c_i})$ belonging to cluster c_i, we denote $L(\mathbf{E}^{c_i}, \mathbf{Y}^{c_i}, \mathrm{S}, \theta)$ as the loss function using a subset S^i_α of the features as obtained from the feature selector optimization problem. The optimization problem for learning the parameters of a classifier can be posed as:

$$
\min_\theta \sum_{i=1}^{|C|} L(\mathbf{E}^{c_i}, \mathbf{Y}^{c_i}, \mathrm{S}^i_\alpha, \theta) + \lambda_1 c(\mathrm{S}^i_\alpha) + \lambda_2 ||\theta||^2 \tag{3}
$$

where λ_1 and λ_2 are hyper-parameters and $|C|$ refers to the total number of clusters. In the above equation, θ is the parameter of the model and can be updated by standard gradient based techniques. This loss function takes into account the important feature subset for each cluster S^i_α and updates the parameter accordingly. The classifier objective also consists of a cost term denoted by $c(\mathrm{S}^i_\alpha)$ to account for the cost of the selected feature subset. For hard budget on the elicited features, the cost component in the model objective can be considered. In case of a cost budget, this component can be

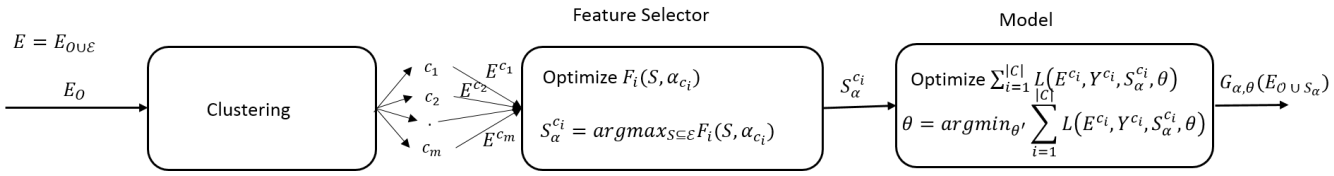


Figure 2: Optimization framework for CATE

ignored because the elicited feature subset adheres to a fixed cost and hence, this term is constant.

Submodularity of the Optimization Problem: We now argue that the optimization problem (equation (2)) is submodular under certain restricted settings. Since, we set $\alpha_{c_i} = 1$, the following is a sufficient condition for submodularity:

LEMMA 3.1. *The objective function (equation* (1)*) is submodular if for all features* $i, j \in \mathcal{E}$, $I(X_i; X_j) \geq I(X_i; X_j|Y)$.

Note that the conditioning does not reduce the mutual information, unlike entropy. Furthermore, the submodularity of this follows from the fact that the negative pairwise similarity function is submodular [9]. Unfortunately, it is not monotone. However, as shown in [12], this function is *approximately* monotone and the greedy algorithm described in the next sub-section has an approximation guarantee.

3.3 Algorithm

We present the algorithm for **Cost Aware Test-time Feature Elicitation** (CATE) in Algorithm 1. CATE takes as input set of training examples $\mathbf{E}$, labels of the example $\mathbf{Y}$, zero cost feature set O, the elicitable feature subset $\mathcal{E}$, a cost vector $\mathbf{M} \in \mathbf{R}^d$ and a budget B.

The training instances $\mathbf{E}$ are clustered based on just the observed feature set O using K-means clustering (`Cluster`). For every cluster c_i, the training instances belonging to a specific cluster is assigned to $\mathbf{E}^{c_i}$, their labels to $\mathbf{Y}^{c_i}$ and is passed to the feature Selector module (lines 5-8). The `FeatureSelector` module identifies the most important feature subset $\mathbf{S}_\alpha^{c_i}$ corresponding to a cluster c_i. Once all the important feature subsets are identified for all the $|C|$ clusters, the model objective function is optimized as mentioned in Equation 3 for all the training instances using the important feature subsets (line 10). All the remaining features are imputed by using either 0 or any other imputation model before training the model. The final training model $\mathtt{G}_{\alpha,\theta}(\mathbf{E}_{O \cup \mathtt{S}_\alpha})$ is a single model which is used to make predictions for a test-instance consisting of just the observed feature subset O. When a test instance with observed feature subset O is encountered, it is first assigned to the closest cluster and the important elicitable feature subset of that cluster is acquired at query time to make predictions.

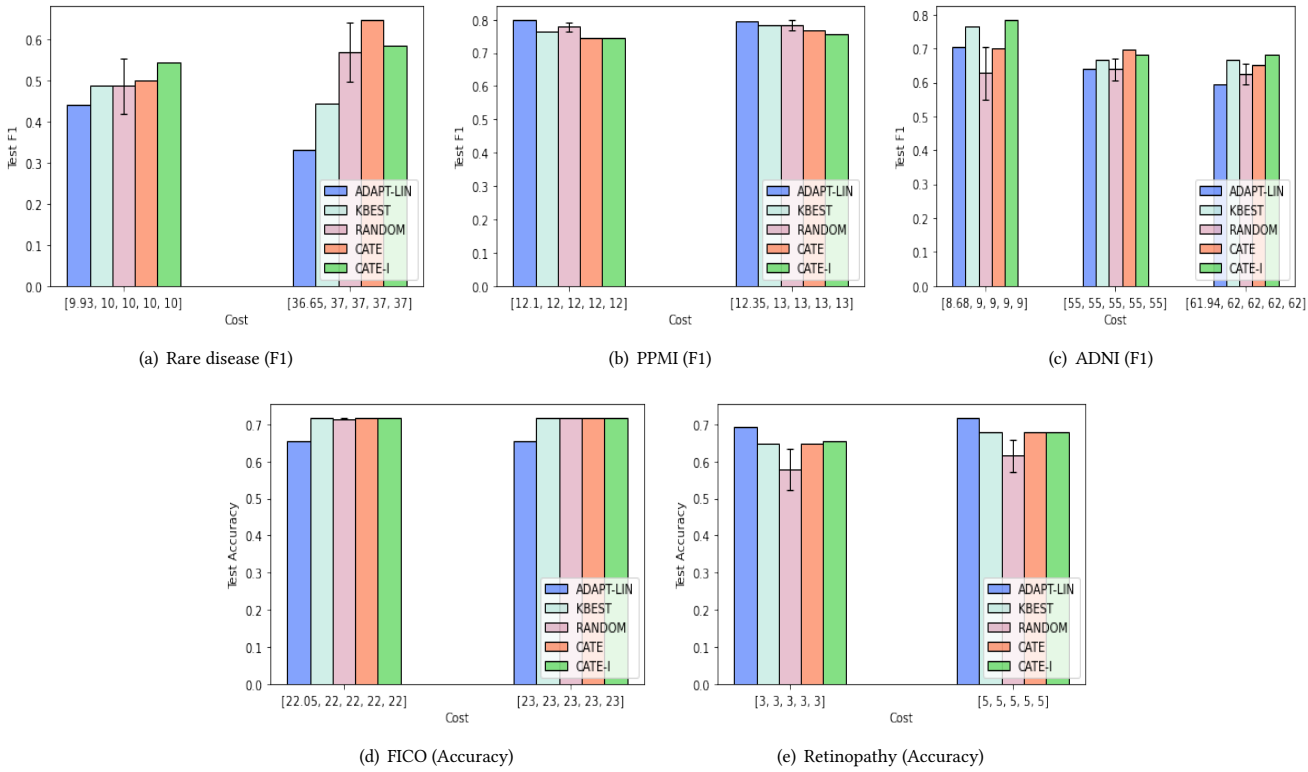
The `FeatureSelector` module in Algorithm 2 takes set of instances belonging to a cluster referred as $\mathbf{I}$, their labels $\mathbf{Y}$, feature selector parameter α, the feature subsets O and $\mathcal{E}$, cost vector $\mathbf{M}$ and a predefined budget B as input and outputs the important feature subset for that cluster. A greedy optimization technique is used to grow the feature subset $\mathbf{F}$ of every cluster based on the feature selector objective function as defined in Equation 1. For every feature in the elicitable set $\mathcal{E}$, the feature selector score defined in Equation 1 is evaluated (lines 10-11). It is to be noted here that the objective function consists of two components: the mutual information related component and the cost component. While individual mutual information is always > 0, the total mutual information score (*MI_score*) is a combination of relevance, redundancy and conditional redundancy and this joint score can be < 0. After calculating the individual feature scores, the best feature is selected and added to the best feature subset $\mathbf{F}$ (line 17). The best feature is removed from the elicitable set $\mathcal{E}$ (line 18) and the algorithm is repeated. Finally, once the entire budget B is exhausted or the mutual information score becomes negative (lines 6-8; 13-15), the `FeatureSelector` module terminates and returns the subset $\mathbf{F}$.

Algorithm 1 Cost Aware Test-time Feature Elicitation

1: **function** CATE($\mathbf{E}, \mathbf{Y}, O, \mathcal{E}, \mathbf{M}, B$)
2: $\mathbf{E} = \mathbf{E}_{O \cup \mathcal{E}}$ ▷ E consists of 0 cost features O and costly features $\mathcal{E}$
3: $C =$ `Cluster`$(\mathbf{E}_O)$ ▷ Clustering based on observed features O
4: **for** $i = 1$ **to** $|C|$ **do**
5: $\mathbf{E}^{c_i}, \mathbf{Y}^{c_i} =$ `GetClusterMember`$(\mathbf{E}, C, i)$
6: ▷ Get data points belonging to each cluster c_i
7: $\mathbf{S}_\alpha^{c_i} =$ `FeatureSelector`$(\mathbf{E}^{c_i}, Y^{c_i}, \alpha, O, \mathcal{E}, \mathbf{M}, B)$
8: ▷ Parameterized feature selector for each cluster
9: **end for**
10: `Optimize` $J(\mathbf{E}, \mathbf{Y}, \mathbf{S}_\alpha, \theta, \mathbf{M})$
11: ▷ Optimize objective function in Equation 3
12: `Update` θ
return $\mathtt{G}_{\alpha,\theta}(\mathbf{E}_{O \cup \mathtt{S}_\alpha})$ ▷ G is final training model
13: **end function**

4 EMPIRICAL EVALUATION

We performed experiments with 3 real-world **medical data sets**, HELOC data set released as part of FICO explainable machine learning challenge and a standard diabetic retinopathy [1] data set from UCI. The data collection in medical domains makes a compelling case for CATE. While we use the medical data sets as motivation we use HELOC data set to demonstrate the algorithm's wide applicability. Table 2 presents down the various features of the data sets used in our experiments. We now briefly explain the real-world data sets before presenting the results.



(a) Rare disease (F1) (b) PPMI (F1) (c) ADNI (F1)

(d) FICO (Accuracy) (e) Retinopathy (Accuracy)

Figure 3: Comparison of CATE against various baselines and Adaptlin by Nan et al. [20] for various feature budgets. Uniform feature cost is assumed for all the algorithms. F1-score is reported for the imbalanced data sets and Accuracy is reported for the balanced data sets.

1. **Parkinson's disease:** The Parkinson's Progression Marker Initiative (PPMI) [17] is an observational study where the aim is to identify Parkinson's disease progression from various types of features. The PPMI data set [5] consists of features related to various motor functions and non-motor behavioral and psychological tests. We consider certain motor assessment features like rising from chair, gait, freezing of gait, posture and postural stability as observed features and treat the other features as elicitable features that have an associated cost.
2. **Alzheimer's disease:** The Alzheimer's Disease NeuroIntiative (ADNI[1]) is a study that aims to test whether various clinical, FMRI and biomarkers can be used to predict the early onset of Alzheimer's disease. In this data set, we consider the demographics of the patients such as age, gender etc as observed with zero cost and the FMRI image features and cognitive scores as unobserved and elicitable features.
3. **Rare disease:** This data set is created from survey questionnaires [16] and the task here is to predict whether a person has a rare disease. The demographic features of subjects are observed while other sensitive questions in the survey regarding technology use, health and disease related meta information are considered elicitable.
4. **HELOC data set:** The Home Equity Line of Credit (HELOC) data set is an anonymized data set made by real home owners released as part of **FICO** explainable machine learning (xML) challenge found at community.fico.com/s/xml [7]. The prediction task is to use information about customers to predict whether they will repay their HELOC account within 2 years of purchase. Trade related information about the customers such as month since the oldest trade was opened, months since the most recent trade has been opened and number of satisfactory trades are considered as observed features and the remaining features are considered as elicitable.

Evaluation Methodology: All the data sets were partitioned randomly into a $80 : 20$ train-test split. Hyper-parameters like the number of clusters on the observed features were chosen by performing $5 - fold$ cross-validation on all the data sets and are reported in Table 1. For the results reported in Table 1, we considered a hard budget on the number of elicitable features. Our proposed algorithm has the best performance when the number of elicitable features is set to approximately half of the total number of features and we picked the number of elicitable features through a validation set. This is because CATE has a trade-off between the amount of information obtained by selecting the important features versus the number of features whose values are imputed. We use $K - means$ clustering as the underlying clustering algorithm and note that any

[1] www.loni.ucla.edu/ADNI

Data set	# elicit feat	# clusters	Algorithm	Accuracy	Recall	F1	AUC-ROC	AUC-PR
Rare disease	34		OBS	-	**0.705**	0.510	0.661	0.360
			RANDOM	-	0.582 ± 0.092	0.563± 0.070	0.705 ± 0.050	0.434 ± 0.056
			KBEST	-	0.47	0.444	0.618	0.338
		9	CATE	-	**0.705**	**0.648**	**0.767**	**0.501**
		9	CATE-I	-	**0.705**	**0.648**	**0.767**	**0.501**
PPMI	15		OBS	-	0.756	0.682	0.739	0.562
			RANDOM	-	**0.847 ± 0.023**	0.807 ± 0.008	0.847 ± 0.006	0.710 ± 0.015
			KBEST	-	0.810	0.80	0.840	0.711
		7	CATE	-	0.819	0.801	0.841	0.710
		7	CATE-I	-	0.819	**0.819**	**0.855**	**0.739**
ADNI	35		OBS	-	0.461	0.400	0.515	0.344
			RANDOM	-	0.526 ± 0.042	0.630 ± 0.047	0.726 ± 0.030	0.579 ± 0.059
			KBEST	-	0.576	0.714	0.778	0.683
		3	CATE	-	**0.615**	**0.744**	**0.797**	**0.709**
		3	CATE-I	-	**0.615**	0.665	0.739	0.557
HELOC	12		OBS	0.582	0.651	0.618	0.579	0.564
			RANDOM	0.706 ± 0.011	0.723 ± 0.020	0.719 ± 0.013	0.705 ± 0.011	0.661 ± 0.009
			KBEST	0.711	0.731	0.724	0.710	0.664
		5	CATE	**0.720**	**0.747**	**0.735**	**0.719**	**0.672**
		5	CATE-I	0.717	0.733	0.729	0.717	0.671
Retinopathy	9		OBS	0.536	**0.691**	0.613	0.526	0.545
			RANDOM	0.667 ± 0.041	0.570 ± 0.023	0.647 ± 0.031	0.674 ± 0.043	0.657 ± 0.037
			KBEST	**0.705**	0.577	**0.676**	**0.714**	**0.696**
		3	CATE	**0.705**	0.577	**0.676**	**0.714**	**0.696**
		3	CATE-I	0.701	0.577	0.672	0.709	0.690

Table 1: Comparison of CATE against other baseline methods on all the data sets. # elicit feat refers to the number of elicitable features used by the algorithms, # clusters refers to the number of clusters used for CATE and CATE-I

Algorithm 2 Feature Selector

function FeatureSelector($\mathbf{I}, \mathbf{Y}, \alpha, O, \mathcal{E}, \mathbf{M}, B$)
  $\mathbf{I} = \mathbf{I}_{O \cup \mathcal{E}}$ ▷ $\mathbf{I}$ consists of 0 cost features O and costly features $\mathcal{E}$ of a cluster
  $\mathbf{F} = \{\varnothing\}$ ▷ Stores best features
  while $True$ **do**
    $feature\_score = \{\varnothing\}$
    if Budget B is exhausted **then**
      exit
    end if
    for $i = 1$ **to** $|\mathcal{E}|$ **do** ▷ Repeat for all elicitable features
      $score = MI\_score - M_i$ ▷ Evaluate score of each $\mathcal{E}_i$ from Equation 1
      $feature\_score = feature\_score \cup score(\mathcal{E}_i)$
    end for
    if $MI\_score < 0$ **then**
      exit
    end if
    $j = argmax(feature\_score)$ ▷ Select feature with highest score
    $\mathbf{F} = \mathbf{F} \cup \mathcal{E}_j$
    $\mathcal{E} = \mathcal{E} \setminus \mathcal{E}_j$ ▷ Remove selected feature from elicitable set
  end while
  return F
end function

clustering algorithm can be employed. For all the reported results, we use an underlying SVM [4] classifier with Radial basis kernel (RBF) function except the HELOC data set where we use Logistic Regression as the underlying model. Since many of our data sets are highly imbalanced, we present metrics such as *precision*, *recall*, *F1*, and *AUC-ROC* for our reported results for imbalanced data sets. We present *accuracy* for balanced data sets. For the Feature selector module, we built upon the existing implementation of Li et al. [11]. We consider two variants of CATE : (1) **CATE** in which we replace the missing and unimportant features of every cluster with 0 and then update the classifier parameters (2) **CATE-I** where we replace the missing features by using a simple imputation model learnt from only the acquired features of training instances. We use mean to impute numeric features and mode to impute categorical features.

Baselines: We consider 4 baselines for evaluating CATE:

(1) Using only the observed and zero cost features to update the training model denoted as OBS.
(2) Using a random subset of fixed number of elicitable features along with all the observed features to update the training model denoted as RANDOM. For this baseline, the results are averaged over 10 runs.
(3) Using the information theoretic feature selector score as defined in Equation 1 to select the 'k' best elicitable features on the entire data without any cluster consideration along with the observed features. The value of 'k' was kept the same as that picked for CATE. This method is denoted as KBEST.
(4) Using an existing approach called ADAPT-LIN by Nan et al. [23] to get the feature budgets and their corresponding performance metric denoted as ADAPT-LIN.

The work by Shim et al. [30] was not chosen as a baseline as they differ significantly from the experimental setting of our approach. They use deep neural networks and related set-encoding method for imputation whereas we use SVM; also their setting considers all the features at prediction time to be unobserved unlike ours where we have a few observed features and thus not a *fair comparison.*

Dataset	# Pos	# Neg	# Observed	# Elicitable
PPMI	554	919	5	31
ADNI	94	287	6	69
Rare Disease	87	232	6	63
HELOC	18545	17017	3	20
Retinopathy	611	540	2	17

Table 2: Details of the 5 data sets used for the experiments # Pos refers to the number of positive examples and # Neg refers to the number of negative examples. # Observed refers to the number of observed features and # Elicitable refers to the maximum number of features that can be acquired.

Results: We aim to answer the following questions:

Q1: How do **CATE** and **CATE-I** with hard budget on features compare against the standard baselines?
Q2: How do the cost-sensitive version of **CATE** and **CATE-I** compare against the cost-sensitive versions of KBEST and RANDOM?
Q3: How do **CATE** and **CATE-I** compare against an existing baseline approach ADAPT-LIN?
Q4: How does **CATE** behave in the absence of cluster-specific features in the underlying data?

For the first set of experiments reported in Table 1, we consider uniform cost on all the features and employ a hard budget constraint on the number of elicitable features. The results reported in Table 1 suggests CATE significantly outperform the other standard baselines OBS, RANDOM and KBEST in almost all the metrics for Rare disease, ADNI and HELOC data set. CATE-I on the other hand significantly outperforms the standard baselines for Rare disease and PPMI data set. For ADNI, CATE-I does better than RANDOM and KBEST on the clinically relevant *recall* metric. For the diabetic retinopathy data set, CATE and CATE-I is at par with KBEST in accuracy and performs significantly better than RANDOM. This answers **Q1** affirmatively.

In Figure 4, we compare the cost version of CATE and CATE-I against KBEST and RANDOM baselines. Cost version takes into account the cost of individual features and adds it as penalty in the feature selector module. Hence, in this version of CATE, a cost budget is used as opposed to hard budget on the number of elicitable features. We generate the cost vector by sampling each feature cost uniformly from (0,1). For PPMI and Rare disease, it can be observed that cost sensitive version of CATE performs consistently better than KBEST with increasing cost budget. In the PPMI data set, the greedy optimization of the feature selector objective on the entire data set for KBEST algorithm leads to elicitation of just a single feature, beyond that the information gain was negative, hence the performance of PPMI across various cost budget remains the same for KBEST. CATE on the other hand, was able to select important feature subsets for various clusters based on the observed features related to gait and postures for the PPMI data set. For ADNI data set, CATE performs better than KBEST mainly in the middle zone of cost budget because this is where the maximum diversity in features are captured by CATE. For the HELOC and Retinopathy data set, the cost version of KBEST and CATE performs almost at par with each other. CATE also performs significantly better than RANDOM baseline for Rare disease, ADNI and Retinopathy data set and at par with other data sets. The reason for CATE performing at par with RANDOM in some data sets is because in CATE, there is a trade-off between choosing the useful features versus the feature acquisition cost. Hence, sometimes to balance this trade-off, CATE has a drop in performance to account for low acquisition cost. This helps in answering **Q2**.

We also compared CATE and CATE-I against Nan et al.'s approach of ADAPT-LIN where a costly model (SVM using RBF) is learnt on all the features and a low cost linear model and gating function is learnt to approximate the function learnt by the high cost model. Since their method assumes uniform cost on the features, we compared the version of CATE with uniform cost consideration against their approach. For fair comparison, we compared against the same number of features as reported by their method. Also, for the imbalanced data sets, we changed the RBF SVM method in the baseline to handle class-imbalance as in CATE and changed the evaluation metric to F1 instead of accuracy. For the two balanced data sets, we employed the same setting as their method. The comparison of CATE and CATE-I against ADAP-LIN, KBEST and RANDOM is shown in Figure 3. In the figures, the x-axis refers to various feature budgets used by the various methods. In 3 out of the 5 data sets, we can see that CATE and CATE-I performs better than ADAPT-LIN across all the feature budgets reported by ADAPT-LIN. Another observation is that for the HELOC data set, RANDOM, KBEST and CATE variants have the same performance because the feature budget is equal to all the features in the data set, hence the methods all perform similarly. This will answer **Q3**.

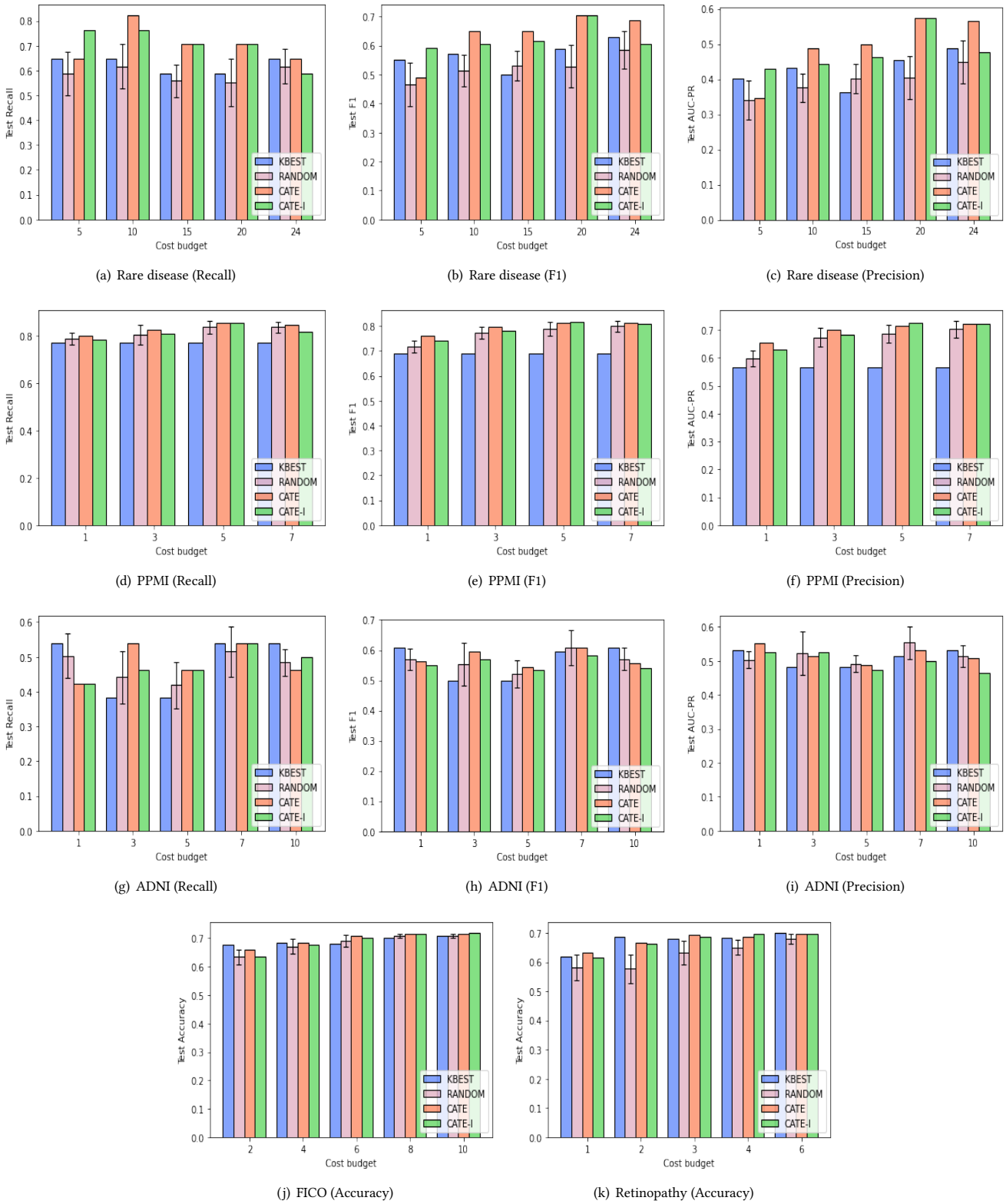
Finally, to answer **Q4**, CATE works in cases where there are cluster-specific features present in the data. In the cases where the data set has important global features, CATE reduces to one of the baselines KBEST and performs similarly to KBEST. This hypothesis can be validated for the diabetic retinopathy data set where KBEST and CATE have similar performance as can be seen in Table 1 and the cost versions of CATE for retinopathy in Figure 4.

5 CONCLUSION

We pose the prediction time feature elicitation problem as an optimization problem by employing a cluster-specific feature selector to choose the best feature subset and demonstrate the effectiveness of our approach in real data sets. We next plan to learn the parameters of the feature selector module by jointly optimizing the feature selector and model parameters. Other interesting avenues include extending our framework to an *active* setting, where, as we obtain new subsets of features for test instances, one could update the model parameters and clustering information. Another direction would be to employ richer feature selection functions and inspired by ideas from submodular function literature to theoretically analyze our algorithms.

ACKNOWLEDGMENTS

SN was supported by grant 1R01HD101246 from NICHD. SD gratefully acknowledge the support of NSF grant IIS-1836565. Any opinions, findings and conclusion or recommendations are those of the authors and do not necessarily reflect the view of the US government, NSF or NIH.



(a) Rare disease (Recall) (b) Rare disease (F1) (c) Rare disease (Precision)

(d) PPMI (Recall) (e) PPMI (F1) (f) PPMI (Precision)

(g) ADNI (Recall) (h) ADNI (F1) (i) ADNI (Precision)

(j) FICO (Accuracy) (k) Retinopathy (Accuracy)

Figure 4: Comparison of various performance metric between CATE ,CATE-I and the baselines. Recall,F1 and AUC-PR is reported for imbalanced data sets, Accuracy for balanced data sets. The x-axis refers to various cost budgets considered which lead to acquisition of different number of features for different budgets.

REFERENCES

[1] Bálint Antal and András Hajdu. 2014. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-based systems* (2014).

[2] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *JMLR* (2012).

[3] Xiaoyong Chai, Lin Deng, Qiang Yang, and Charles X Ling. 2004. Test-cost sensitive naive bayes classification. In *ICDM*.

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* (1995).

[5] Devendra Singh Dhami, Ameet Soni, David Page, and Sriraam Natarajan. 2017. Identifying Parkinson's Patients: A Functional Gradient Boosting Approach. In *AIME*.

[6] Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. 2011. Datum-wise classification: a sequential approach to sparsity. In *ECML PKDD*. 375–390.

[7] FICO. 2018. Explainable machine learning challenge. (2018).

[8] Tianshi Gao and Daphne Koller. 2011. Active classification based on value of classifier. In *NIPS*.

[9] Rishabh Krishnan Iyer. 2015. *Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications*. Ph.D. Dissertation.

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

[11] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2018. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* (2018).

[12] Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 912–920.

[13] Charles X Ling, Qiang Yang, Jianning Wang, and Shichao Zhang. 2004. Decision trees with minimal costs. In *ICML*.

[14] D. J. Lizotte, O. Madani, and R. Greiner. 2003. Budgeted learning of Naive-Bayes classifiers *(UAI)*. 378–385.

[15] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *ICML*.

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

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

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

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

[20] Feng Nan and Venkatesh Saligrama. 2017. Adaptive classification for prediction under a budget. In *NIPS*.

[21] Feng Nan, Joseph Wang, and Venkatesh Saligrama. 2015. Feature-budgeted random forest. In *ICML*.

[22] Feng Nan, Joseph Wang, and Venkatesh Saligrama. 2016. Pruning random forests for prediction on a budget. In *NIPS*.

[23] Feng Nan, Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. 2014. Fast margin-based cost-sensitive classification. In *ICASSP*.

[24] Sriraam Natarajan, Srijita Das, Nandini Ramanan, Gautam Kunapuli, and Predrag Radivojac. 2018. On Whom Should I Perform this Lab Test Next? An Active Feature Elicitation Approach.. In *IJCAI*.

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

[26] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.

[27] Thomas Rückstieß, Christian Osendorfer, and Patrick van der Smagt. 2011. Sequential feature selection for classification. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 132–141.

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

[29] Victor S Sheng and Charles X Ling. 2006. Feature value acquisition in testing: a sequential batch test algorithm. In *ICML*.

[30] Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *NIPS*.

[31] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. 2015. Efficient learning by directed acyclic graph for resource constrained prediction. In *NIPS*.

[32] Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. 2013. Cost-sensitive tree of classifiers. In *ICML*.

[33] Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. 2012. The greedy miser: learning under test-time budgets. In *ICML*.

[34] Z. Zheng and B. Padmanabhan. 2002. On active learning for data acquisition *(ICDM)*. 562–569.