Cost Aware 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

Motivated by clinical tasks where acquiring certain features such as FMRI or blood tests can be expensive, 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. Our experiments on three real-world medical tasks demonstrate the efficacy and effectiveness of our proposed approach in minimizing costs and maximizing 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. 2020. Cost Aware Feature Elicitation. In *San Diego '20: Knowledge-infused Mining and Learning, 2020, San Diego, CA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In supervised classification setting, every instance has a fixed feature vector and a discriminative function is learnt on such fixed-length feature vector and it's corresponding class variable. However, a lot of practical problems like healthcare, network domains, designing survey questionnaire [19, 20] etc has an associated feature acquisition cost. In such domains, there is a cost budget and getting all the features of an instance can be very costly. As a result, many cost sensitive classifier models [2, 8, 24] have been proposed in literature to incorporate the cost of acquisition into the model objective during training and prediction.

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. Consider a patient visiting a doctor for some potential diagnosis of a disease. For such a patient, information like age, gender, ethnicity and other demographic features are easily available at zero cost. However, various lab tests that the patient needs to undergo incurs cost. So, a training model should be able to identify the most relevant (i.e. those which are most informative, yet least costly) lab tests that are required for each *specific* patient. The intuition of this work 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. Based on this intuition, we find out similar data points in the observed feature space and identify the important feature subsets of these similar instances by employing a greedy information theoretic feature selector objective.

Our **contributions** in this work are as follows: (1) 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. (2) account for acquisition cost in both the feature selector objective and classifier objective to balance the trade-off between acquisition cost and model performance. (3) empirically demonstrate the effectiveness of the proposed approach on three real-world medical data sets.

2 RELATED WORK

The related work 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 in literature. A lot of work has been done in tree based models [5, 16, 17, 26–28] by adding cost term to the tree objective function in either decision trees or ensemble methods like gradient boosted trees. All these methods aim to build an adaptive and complex decision tree boundary by considering trade-off between performance and test-time feature acquisition cost. While we are similar in motivation to these approaches, our methodology is different in the sense that we do not consider tree based models. Instead our approach aims to find local feature subsets using an information theoretic feature selector for different clusters of training instance build in a lower dimensional space.
Adaptive classification and dynamic feature discovery: Our work also draws inspiration from Nan al.'s work [15] where they learn a high performance costly model and approximate the model's performance adaptively by building a low cost model and 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 because we do not maintain a high cost model which is costly to build and 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 [18] where they select varying feature subsets for test instance using neighbourhood information of the training data. While calculating the neighborhood

San Diego '20, 2020, San Diego, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

information from training data is similar to building clusters in our approach, the training neighborhood for our method is on just the observed feature space. Moreover, we incorporate the neighbourhood information in the training algorithm whereas Nan et al.'s work is a prediction time algorithm. Ma et al. [10] also address this problem of dynamic discovery of features based on generative modelling and Bayesian experimental design.

Feature elicitation using Reinforcement learning: There is another line of work along the sequential decision making literature [4, 9, 22] to model the test time elicitation of features by learning the optimal policy of test feature acquisition. Along this direction, our work aligns with the work of Shim et al. [25] where they jointly train a classifier and RL agent together. Their classifier objective function is similar to our method with a cost penalty, however they use a Deep RL agent to figure out the policy. We on the other hand use localised feature selector to find the important feature subsets for the underlying training clusters in the observed feature space.

Active Feature Acquisition: Our problem set-up is also inspired by work along active feature acquisition [13, 14, 19, 23, 29] where certain feature subsets are observed and rest are acquired at a cost. While all the above mentioned work follow this problem set up during training time and typically use active learning to seek informative instances at every iteration, we use this particular setting for test instances. Unlike their work, all the training instances in our work are fully observed and the assumption is that the feature acquisition cost has already being paid during training. Also, we address a supervised classification problem instead of an active learning set up. Our problem set up is similar to Kanani et al. [6] as they also have partial test instances, however their problem is that of instance acquisition where the acquired feature subset is fixed. Our method aims at discovering variable length feature subsets for various underlying clusters.

Our contributions: Although the problem of prediction time feature elicitation has been explored in literature from various directions and with various assumptions, we come up with an intuitive solution to this problem and formulate 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 .

3 COST AWARE FEATURE ELICITATION

3.1 Problem setup

Given: A dataset $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ with each $x_i \in \mathbf{R}^d$ as the feature set. Each feature has an associated cost r_i.

Objective: Learn a discriminative model which is aware of the feature costs and can balance the trade-off between feature acquisition cost and model performance.

We make an additional assumption here that there is a subset of features which have 0 cost. These could be, for example, demographic information (e.g. age, gender, etc) in a medical domain which are easily available/less cumbersome to obtain 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). The goal is to use these observed features to find similar instances in the training set and identify the important feature subsets for each of these clusters based on a feature selector objective function which balances the trade-off between choosing the important features and the cost at which these features are acquired.

3.2 Proposed solution

As a first step, we cluster the training instances based on just the observed zero cost feature set $\mathcal{O}$. The intuition is that instances with similar features will also have similar characteristics in terms of which elicitable features to order. For example, in a medical application, whether to request for a blood test or a ct-scan will depend on factors such as age, gender, ethnicity and whether patients with similar demographic features had requested these tests. Also, since the feature set $\mathcal{O}$, comes at zero cost, we assume that for unseen test instances, this feature set is observed.

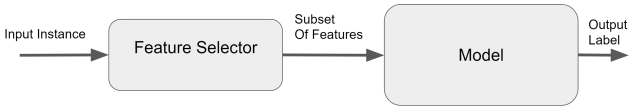

Figure 1: Optimization framework for the proposed problem

We propose a model which consists of a parameterized feature selector module $F(X, \mathbf{E}^{c_i}, \alpha)$ which takes in a set of input instances E^{c_i} belonging to the cluster c_i based on the feature set $\mathcal{O}$ and produces a subset X of most important features for the classification task. The feature selection model is based on an information- theoretic objective function and is augmented with the feature cost to account for the trade off between model performance and acquisition cost at test-time. The output feature subset from the feature selector module are used to update the parameters of the classifier. The optimization framework is shown in Figure 1

Information theoretic Feature selector model: The feature selector module selects the best subset of features for each cluster of training data based on an information theoretic objective score. Since at test time, we do not know the elicitable feature subset $\mathcal{E}$ (since the goal of feature selection is in the first place to find the truly necessary features for learning). Hence we propose to use the closest set of instances in the training data to the current instance. Since we assume that the training data has already been elicited, we have all the features observed in the training data. We compute this distance just based on the observed feature set $\mathcal{O}$. We cluster the training data based on the observed features into m clusters $c_1, c_2, \cdots c_m$. Next, we use the Minimum-Redundancy-Maximum Relevance (MRMR) feature Selection paradigm [1, 21]. 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. The feature selection module is a function of the parameters of the cluster to which a set of instances belong and is defined as:

$$F(X, \mathbf{E}^{c_i}, \alpha_{c_i}) = \underbrace{\alpha^1_{c_i} \sum_{\mathcal{E}_p \in X} I(\mathcal{E}_p; Y)}_{\text{max. relevance}} \\ \underbrace{- \sum_{\mathcal{E}_p \in X} \left(\alpha^2_{c_i} \sum_{\mathcal{E}_j \in X} I(\mathcal{E}_j; \mathcal{E}_p) - \alpha^3_{c_i} \sum_{\mathcal{E}_j \in X} I(\mathcal{E}_p; \mathcal{E}_j | Y) \right)}_{\text{min. redundancy}} \\ \underbrace{- \alpha^4_{c_i} \sum_{\mathcal{E}_p \in X} c(\mathcal{E}_p)}_{\text{cost penalty}} \quad (1)$$

where $I(\mathcal{E}_p; Y)$ is the mutual information between the random variables $\mathcal{E}_p$ (feature) and Y (target). In the above equation, the feature subset X is grown greedily using a greedy optimization strategy maximizing the above objective function. In equation 1, $\mathcal{E}_p$ denotes a single feature from the elicitable set $\mathcal{E}$ that is considered for evaluation based on the subset X grown so far. The first term is the mutual information between each feature 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 and the features already added to the feature subset X. This value needs to be minimized for selecting informative features. The third term is called the conditional redundancy [1] 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. For this work, we do not learn the parameters α_{c_i} for each cluster, instead fix these parameters to 1. We leave the learning of these parameters to future work.

In the problem setup, since the 0 cost feature subset is always present, we always consider the observed feature subset $\mathcal{O}$ 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 features with these observed features. It is to be noted that in cases where the observed features are not discriminative enough of the target, the feature selector module ensures that the elicitable features with **maximum relevance** to the target variable are picked.

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

$$X^i_\alpha = \text{argmax}_{X \subseteq \mathcal{E}} F(X, \mathbf{E}^{c_i}, \alpha) \quad (2)$$

For a given instance (x, y), we denote $L(x, y, X, \theta)$ as the loss function using a subset X 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:

$$\min_\theta \sum_{i=1}^n L(x_i, y_i, X^i_\alpha, \theta) + \lambda_1 c(X^i_\alpha) + \lambda_2 ||\theta||^2 \quad (3)$$

where λ_1 and λ_2 are hyper-parameters. 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 and updates the parameter accordingly. The classifier objective also consists of a cost term denoted by $c(X^i_\alpha)$ 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 ignored because the elicited feature subset adheres to a fixed cost and hence, this term is constant.

3.3 Algorithm

We present the algorithm for **Cost Aware Feature Elicitation** (CAFE) in Algorithm 1. CAFE takes as input set of training examples $\mathbf{E}$, the zero cost feature set $\mathcal{O}$, the elicitable feature subset $\mathcal{E}$, a cost vector $M \in \mathbf{R}^d$ and a budget B. Each element in the training set $\mathbf{E}$ consists of a tuple (x, y) where $x \in \mathbf{R}^d$ is the feature vector and y is the label.

The training instances $\mathbf{E}$ are clustered based on just the observed feature set $\mathcal{O}$ using K-means clustering (`Cluster`). For every cluster c_i, the training instances belonging to the cluster is assigned to the set $\mathbf{E}^{c_i}$ and is passed to the Feature Selector module (lines 6-8). The `FeatureSelector` function takes $\mathbf{E}^{c_i}$, parameter α, the feature subsets $\mathcal{O}$ and $\mathcal{E}$, cost vector M and a predefined budget B as input and returns the most important feature subset $\mathbf{X}^{c_i}_\alpha$ corresponding to a cluster c_i. A greedy optimization technique is used to grow the feature subset X of every cluster based on the feature selector objective function defined in Equation 1. The `FeatureSelector` terminates once the budget B is exhausted or the mutual information score becomes negative. Once all the important feature subsets are obtained 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 for the clusters to which the training instances belong (lines 12-18). All the remaining features are imputed by using either 0 or any other imputation model before training the model. The final training model $\mathtt{G}(\mathbf{E}_{\mathcal{O} \cup X_\alpha}, \alpha, \theta)$ is an unified model used to make predictions for a test-instance consisting of just the observed feature subset $\mathcal{O}$.

4 EMPIRICAL EVALUATION

We did experiments with 3 real world **medical data sets**. The intuition of CAFE makes more sense in medical domains, hence our choice of data sets. However, the idea can be applied to other domains ranging from logistics to resource allocation task. Table 2 jots down the various features of the data sets used in our experiments. Below are the details of the 3 real data sets, we use for our experiments.

1. **Parkinson's disease prediction:** The Parkinson's Progression Marker Initiative (PPMI) [12] is an observational study where the aim is to identify Parkinson's disease progression from various types of features. The PPMI data set consists of various 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 rest all features as elicitable features which must be acquired at a cost.

Algorithm 1 Cost Aware Feature Elicitation

```
function CAFE($\mathbf{E}, \mathcal{O}, \mathcal{E}, M, B$)
    $\mathbf{E} = \mathbf{E}_{\mathcal{O}\cup\mathcal{E}}$    ▷ E consists of 0 cost features $\mathcal{O}$ and costly features $\mathcal{E}$
    $C$ = Cluster($\mathbf{E}_{\mathcal{O}}$)    ▷ Clustering based on the observed features $\mathcal{O}$
    $\mathbf{X} = \{\varnothing\}$    ▷ Stores best feature subsets of each cluster
    for $i = 1$ to $|C|$ do    ▷ Repeat for every cluster
        $\mathbf{E}^{c_i}$ = GetClusterMember($\mathbf{E}, C, i$)
            ▷ get the data points belonging to each cluster $c_i$
        $\mathbf{X}_{\alpha}^{c_i}$ = FeatureSelector($\mathbf{E}^{c_i}, \alpha, \mathcal{O}, \mathcal{E}, M, B$)
            ▷ Parameterized feature selector for each cluster
        $\mathbf{X} = \mathbf{X} \cup \{\mathbf{X}_{\alpha}^{c_i} \cup \mathcal{O}\}$
    end for
    for $i = 1$ to $|C|$ do    ▷ Repeat for every cluster
        $\mathbf{X}_{\alpha}^{c_i}$ = GetFeatureSubset($\mathbf{X}, i$)
            ▷ Get the feature subset for each cluster $c_i$
        for $j = 1$ to $|\mathbf{E}^{c_i}|$ do    ▷ Repeat for every data point in cluster $c_i$
            Optimize $J(x_j, y_j, \mathbf{X}_{\alpha}^{c_i}, \theta, M)$
                ▷ Optimize the objective function in Equation 3
            Update $\theta$    ▷ Update the model parameter $\theta$
        end for
    end for
    return G($\mathbf{E}_{\mathcal{O}\cup\mathbf{X}_{\alpha}}, \alpha, \theta$) ▷ G is the training model built on E
end function
```

2. **Alzheimer's disease prediction:** 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 as observed and zero cost features and the FMRI image data and cognitive score data as unobserved and elicitable features.
3. **Rare disease prediction** This data set is created from survey questionnaires [11] and the task here is to predict whether a person has rare disease or not. The demographic features are observed while other sensitive questions in the survey regarding technology use, health and disease related meta information is considered to be elicitable.

Evaluation Methodology: All the data sets were partitioned into a 80:20 train-test split. Hyper parameters like the number of clusters on the observed features were picked by doing 5 fold cross validation on all the data sets. The optimal number of clusters picked were 6 for ADNI, 9 for Rare disease data set and 7 for the PPMI data set. For the results reported in Table 1, we considered a hard budget on the number of elicitable features and set it to half of the total number of features in the respective data set. We use K-means clustering as the underlying clustering algorithm. For all the reported results, we use an underlying Support Vector Machine [3] classifier with Radial basis kernel function. Since, all the data sets are highly imbalanced, hence we consider metrics like *recall*, *F1*, *AUC-ROC* and *precision* for our reported results. For the Feature selector module, we used the existing implementation of Li et al. [7] and built upon it. We consider two variants of CAFE:(1) **CAFE** in which we replace the missing and unimportant features of every cluster with 0 and then update the classifier parameters (2) **CAFE-I** where we replace the missing and unimportant features by using an imputation model learnt from the already acquired feature values of other data points. A simple imputation model is used where we replace the missing features with *mode* for categorical features and *mean* for numeric features.

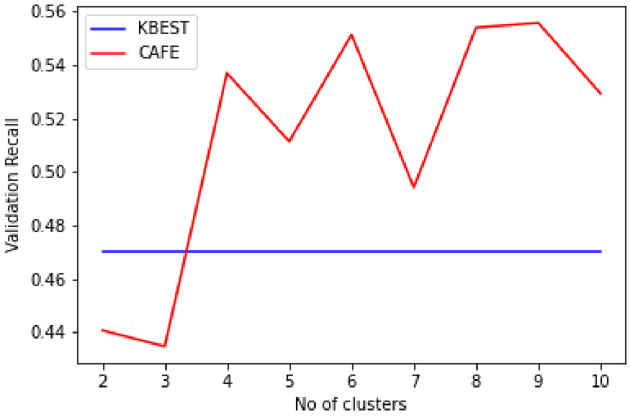

Figure 2: Recall Vs number of clusters for Rare disease for CAFE-I

Baselines: We consider 3 baselines for evaluating CAFE and CAFE-I: (1) using 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 and 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 denoted as KBEST. We keep the value of 'k' to be the same as that used by CAFE. Although some of the existing methods could be potential baselines, none of these methods match the exact setting of our problem, hence we do not compare our method against them.

Results: We aim to answer the following questions:

Q1: How does **CAFE** and **CAFE-I** with hard budget on features compare against the standard baselines?

Q2: How does the cost-sensitive version of CAFE and CAFE-I fare against the cost-sensitive baseline KBEST?

The results reported in Table 1 suggests both CAFE and CAFE-I significantly outperform the other baselines in almost all the metrics for Rare disease and PPMI data set. For ADNI, CAFE and CAFE-I outperform the other baselines in clinically relevant recall metric while KBEST performs the best for the other metrics. The reason for this is that in ADNI, since, the elicitable features are image features and we discretize the image features to calculate the information gain for the feature selector module, the granular level feature information is lost because of this discretization and hence the drop in performance. For the experiments in Table 1, we keep the budget to be approximately half of the total number

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

Data set	Algorithm	Recall	F1	AUC-ROC	AUC-PR
Rare disease	OBS	**0.647**	0.488	0.642	0.347
	RANDOM	0.57 ± 0.064	0.549± 0.059	0.693 ± 0.042	0.421 ± 0.051
	KBEST	0.47	0.457	0.628	0.349
	CAFE	**0.647**	0.628	0.749	0.489
	CAFE-I	**0.647**	**0.647**	**0.759**	**0.512**
PPMI	OBS	0.765	0.685	0.741	0.563
	RANDOM	**0.857 ± 0.023**	0.809 ± 0.015	0.85 ± 0.013	0.712 ± 0.020
	KBEST	0.828	0.807	0.846	0.716
	CAFE	0.846	0.817	0.855	0.726
	CAFE-I	0.855	**0.829**	**0.865**	**0.743**
ADNI	OBS	0.5	0.44	0.553	0.365
	RANDOM	0.711 ± 0,043	0.697 ± 0.082	0.767 ± 0.064	0.592 ± 0.098
	KBEST	0.73	**0.745**	**0.806**	**0.646**
	CAFE	**0.807**	0.711	0.786	0.578
	CAFE-I	0.769	0.701	0.776	0.574

Table 1: Comparison of CAFE against other baseline methods on 3 real data sets

Dataset	# Pos	# Neg	# Observed	# Elicitable
PPMI	554	919	5	31
ADNI	94	287	6	69
Rare Disease	87	232	6	63

Table 2: Data set details of the 3 real data sets used.#Pos is number of positive example, #Neg is number of negative example. # Observed is number of observed features and # Elicitable is the maximum number of features that can be acquired.

of features for all the methods. On an average, CAFE-I performs better than CAFE across all the data sets because of the underlying imputation model which helps in better treatment of the missing values as against replacing all the features by 0. This answers **Q1** affirmatively.

In Figure 3, we compare the cost version of CAFE and CAFE-I against KBEST. Cost version takes into account the cost of individual features and accounts for them as penalty in the feature selector module. Hence, in this version of CAFE, a cost budget is used as opposed to hard budget on the number of elicitable features. We generate the cost vector by sampling each cost component uniformly from (0,1). For PPMI and Rare disease, we can see that cost sensitive CAFE 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 lead to elicitation of just 1 feature, beyond that the information gain was negative, hence the performance of PPMI across various cross budget remains the same. CAFE 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 ADNI data set, CAFE performs better than KBEST only in recall. The reason for this is the same as mentioned above. This helps in answering **Q2** affirmatively.

Lastly, Figure 2 shows the effect of increasing cluster on the validation recall for the Rare disease data set. As can be seen, for smaller number of clusters, the recall is very low and increases to an optimum for 9 clusters. This helps us in understanding the fact that forming clusters based on observed important features helps CAFE in selecting different feature subsets for different clusters, thus helping the learning procedure.

5 CONCLUSION

In this paper, 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 then optimizing the training loss. We show the effectiveness of our approach in real data sets where the problem set up is intuitive. Future work includes learning the parameters of the feature selector module and jointly optimizing the feature selector and model parameters for a more robust framework and adding more constraints to optimization.

ACKNOWLEDGEMENTS

SN & 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.

REFERENCES

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

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

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

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

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

[6] 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).

[7] 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).

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

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

[10] 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*.

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

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

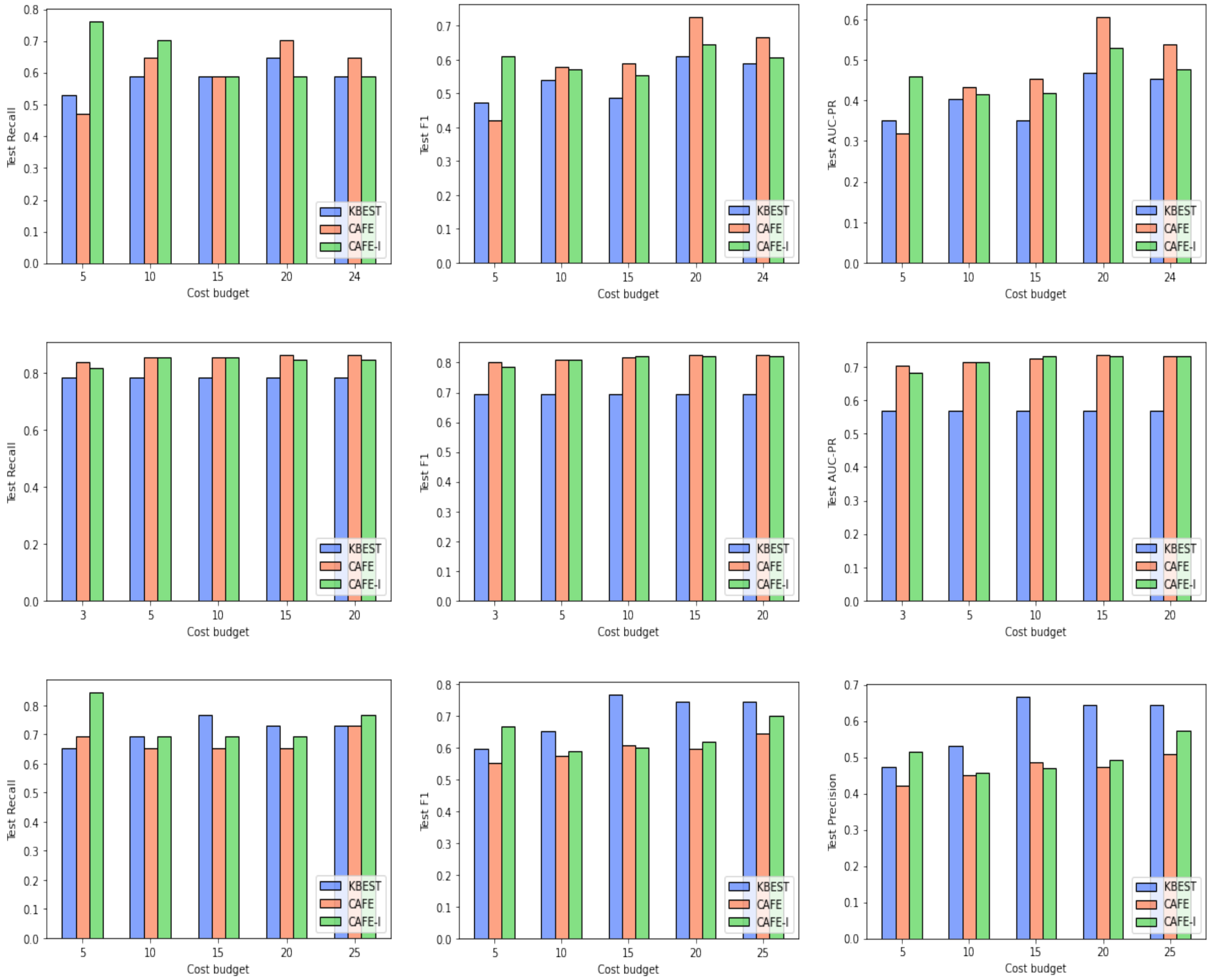

Figure 3: Recall (left), F1 (middle), AUC-PR (right) for (from top to bottom) Rare Disaese, PPMI, and ADNI. The x-axis refers to the cost budget used which leads to the elicitation of different number of features.

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

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

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

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

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

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

[19] 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*.

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

[21] 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.

[22] 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.

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

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

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

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

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

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

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