Nutrition and Health Data for Cost-Sensitive Learning

Mohammad Kachuee, Kimmo Karkkainen, Orpaz Goldstein, Davina Zamanzadeh, and Majid Sarrafzadeh
Department of Computer Science
University of California, Los Angeles (UCLA)

***Abstract*—Traditionally, machine learning algorithms have been focused on modeling dynamics of a certain dataset at hand for which all features are available for free. However, there are many concerns such as monetary data collection costs, patient discomfort in medical procedures, and privacy impacts of data collection that require careful consideration in any health analytics system. An efficient solution would only acquire a subset of features based on the value it provides whilst considering acquisition costs. Moreover, datasets that provide feature costs are very limited, especially in healthcare. In this paper, we provide a health dataset as well as a method for assigning feature costs based on the total level of inconvenience asking for each feature entails. Furthermore, based on the suggested dataset, we provide a comparison of recent and state-of-the-art approaches to cost-sensitive feature acquisition and learning. Specifically, we analyze the performance of major sensitivity-based and reinforcement learning based methods in the literature on three different problems in the health domain, including diabetes, heart disease, and hypertension classification.**

***Index Terms*—Cost-sensitive learning, opportunistic learning, feature acquisition, health data, health informatics**

I. Introduction

Traditional machine learning is focused on modeling dynamics of a dataset consisting of features that are freely available. However, in many real-world problems, especially in the health domain, having access to the value of each feature entails a certain cost which requires careful consideration. This notation of cost is general and may include the actual monetary cost, patient discomfort, privacy impacts, and so forth [1]. Careful consideration of this cost and devising algorithms and methods that consider this notion can be crucially important in health settings as it can reduce the data collection costs and increase the human subject compliance. In a cost-sensitive learning scenario, information is being acquired based on optimizing the balance between the predictive value it provides and the cost entailed by the acquisition.

Recently, cost-sensitive learning has garnered a lot of attention. For instance, sensitivity analysis of predictor models was suggested as a measure of feature importance [2], [3]. Alternatively, reinforcement learning solutions were suggested that formulate the feature acquisition and prediction process as a Markov decision process [4]–[7].

While there has been good progress in the development of these algorithms and methods, there has been little work done on the evaluation and application of them on real-world data in general, and health data in particular. The main reason behind this is that the currently available datasets rarely provide feature costs. Consequently, arbitrary or synthesized cost assignments are frequently being used in the literature which prevents the evaluation these methods in actual use cases such as disease diagnosis.

In this paper, we provide a study of cost-sensitive learning for smart health scenarios. Specifically, we provide a framework for mining datasets from public health records released by CDC. It consists of demographics, examination, questionnaire, and laboratory data for about 100,000 individuals. Furthermore, we propose a methodological way for real cost assignment based on a survey conducted using Amazon Mechanical Turk. Furthermore, we present a comparison of major cost-sensitive learning methods on this dataset and across various problems. The related source code and data for this paper is available online[1].

II. Methodology

A. Data Source

We use the National Health and Nutrition Examination Survey (NHANES) [8] between 1999 to 2016 as our data source. NHANES is an ongoing survey which is designed to assess the well-being of adults and children in the United States. For each year, health data collected from few thousand individuals and it consists of demographics, questionnaire, examination, and laboratory data. Not all data is collected from each individual (e.g., certain blood tests are not used for young children) and there is a slight variation between the information being collected at each year (e.g., the prevalence of disease change over time causing changes on the data collection focus). For more information about NHANES please refer to [8].

B. Data Preparation

We developed a general data processing pipeline which can be used for different tasks and settings. The data preparation starts with loading raw data files associated with each variable in the dataset containing values of that variable for each subject. Please note that we merge columns and rows based on variable and subject identifiers so that, logically, all variables appear as columns and individuals appears as a row. For a certain task, any available variable could be defined as a feature or a target, depending on the task.

The dataset consists of 9385 unique variables of different types including categorical, real-valued, multiple choice, etc.

[1]The related source code and data for this paper is available at https://github.com/mkachuee/OpportunisticData

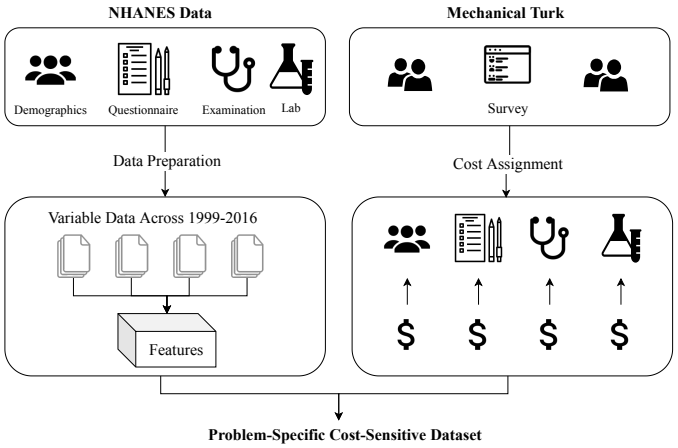


Fig. 1: Visualization of the proposed preprocessing and cost assignment pipeline.

Accordingly, we use different preprocessing functions such as statistical normalization, one-hot encoding, etc. to prepare each variable for further analysis. Also, it should be noted that, for each individual, only a subset of these variables is available and the rest are missing.

To define each certain task (e.g. diabetes classification), a target should be defined as a function of available variables (e.g., blood glucose level categories). On the other hand, two methods are suggested to determine variables to be used as input features and appropriate preprocessing functions : (i) explicitly defining a list of variables and their corresponding preprocessing functions (ii) automatically selecting relevant features by searching over the space of variables and automatically deciding on appropriate preprocessing functions. Regarding the second method, one can limit the number of features being eventually used by setting a threshold on the mutual information between the target and each feature as well as a threshold on the percentage of available (not missing) features for each selected variable.

C. Cost Assignment

In order to assign costs corresponding to each feature, we conducted a survey to collect the level of overall inconvenience that asking for each feature would cause to subjects. Specifically, we collected survey data from 108 individuals using the Amazon Mechanical Turk framework. Since the data is collected from individuals in the United States, we limited our survey to the same population. Table I presents the questionnaire used in this study. Before starting the questionnaire, we asked the turkers to pay attention to the following instructions:

- Please rate each question in terms of the total inconvenience they will cause you (including time burden, financial cost, discomfort, etc.).
- Assume that each item will provide you with useful health information; however, there is no urgency to do any of these.
- The scale is 1 to 10, 10 being the most convenient. Rate each item based on the relative level of inconvenience.
- After completing the sheet, please review and adjust, if necessary.

The median of survey results for each question is used as a level of convenience for each question. In order to convert these values to cost values (i.e., the higher the more expensive), we subtract each convenience value from 11 and consider the resulting value as the cost of acquiring features of that category. Accordingly, the final cost for feature categories corresponding to questions 1 to 4 of Table I is determined to be 2, 4, 5 and 9, respectively. See Fig. 2 for the distribution of answers.

D. Inference

1) Problem Definition: In this paper, we consider the general scenario of supervised classification using a set of features $\boldsymbol{x}_i \in \mathbb{R}^d$ and ground truth target labels $\tilde{y}_i$. However, initially, the values of features are not available and there is a cost for acquiring each feature $(\boldsymbol{c}_j; 1 \leq j \leq d)$. Consequently, for sample i, at each time step t, we only have access to a partial realization of the feature vector denoted by $\boldsymbol{x}_i^t$ consisting of features that are acquired until t. There might be a maximum budget (B) or a user-defined termination condition (e.g., prediction confidence) which limits the features being acquired eventually.

More formally, we define a mask vector $\boldsymbol{k}_i^t \in \{0,1\}^d$ where each element of $\boldsymbol{k}$ indicates if the corresponding feature is available in $\boldsymbol{x}_i^t$. Using this notation, the total feature acquisition cost at each time step can be represented as

$$C_{total,i}^t = (\boldsymbol{k}_i^t - \boldsymbol{k}_i^0)^T \boldsymbol{c} \quad . \tag{1}$$

Furthermore, we define the feature query operator (q) as

$$\boldsymbol{x}_i^{t+1} = q(\boldsymbol{x}_i^t, j), \text{ where } \boldsymbol{k}_{i,j}^{t+1} - \boldsymbol{k}_{i,j}^t = 1 \quad . \tag{2}$$

In this setting, the objective of a cost-sensitive feature acquisition algorithm is to balance the accuracy versus cost trade-off via efficiently acquiring as many features as necessary at test-time.

2) Sensitivity-Based Approach: Sensitivity-based approaches use trained classifier models and select features that have the most influence on the model predictions. This influence can be used as a utility function which measures the importance of having access to each feature value.

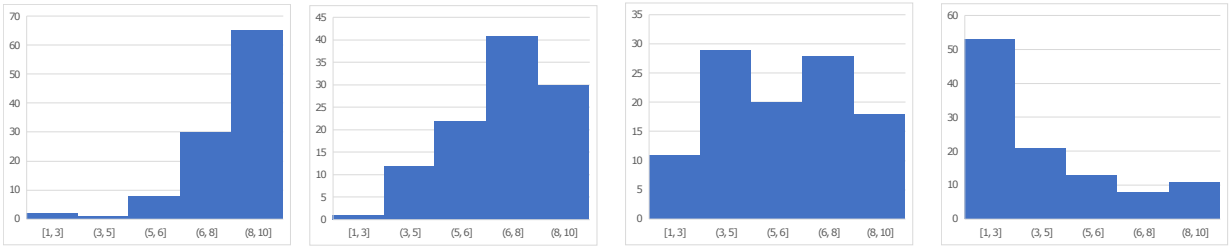
Early *et al.* [2] suggested an exhaustive measurement of expected sensitivity for each feature:

$$\mathbb{E}[U(x^t, j)] = \int p(x_j = r | x^t) U(x^t, x_j = r) dr, \tag{3}$$

where $U(x^t, j)$ is the expected utility of acquiring feature j given the feature vector at t, and with the abuse of notation, $U(x^t, x_j = r)$ is the utility of that feature assuming its value after acquisition would be equal to r. It should be noted that as this method requires modeling joint probability distributions as well as integration over feature values, it is not scalable to datasets consisting of many features.

TABLE I: The questionnaire used in this study.

No.	Question	Answer
1	Convenience of answering general demographics related questions (e.g., age, gender, race, etc.)	1...10
2	Convenience of answering general behavioral/life-style related questions (e.g., smoking habits, sleeping habits, alcohol consumption, drug usage etc.)	1...10
3	Convenience of getting typical examinations such as weight, height, or blood pressure measurement	1...10
4	Convenience of taking a blood or urine test at a lab	1...10



(a) (b) (c) (d)

Fig. 2: Distribution of answers for the level of convenience collecting information about: (a) demographics, (b) behavioral/life-style, (c) medical examinations, and (d) lab tests.

In [3], the authors suggest an approximation of (3) using a binary representation layer in denoising autoencoder architectures:

$$j^t_{sel} = \underset{j\in\{1...d\}|k^t_j=0}{argmax} \frac{\sum_{b=1}^{b=l} |\frac{\partial h(\boldsymbol{x}^t)}{\partial x_{bin_{j,b}}}| \, x'_{bin_{j,b}}}{c^t_j}, \quad (4)$$

where $h(x^t)$ is the classifier outputs and $x_{bin_{j,b}}$ represents binary representation for bth bit of the jth feature. It is worth noting that this approximation leads to much faster computation based on a single forward and backward evaluation of the network.

3) Reinforcement Learning Approach: The cost-sensitive feature acquisition problem can be formulated as a reinforcement learning problem. In this setting, each state would be the features that are acquired at each point. Additionally, each action would be to pay for a certain feature and to acquire its value, transitioning to a new state. Here, the objective would be to learn a policy function that results in an efficient feature acquisition. One possible reward function which is frequently used in the literature [5], [6] is:

$$r(x^t_i, a) = \begin{cases} -\boldsymbol{c}_j & a \text{ is acquiring feature } j \\ 0 & a \text{ is making a correct prediction} \\ -\lambda & a \text{ is making an incorrect prediction} \end{cases}, \quad (5)$$

where λ is a hyperparameter controlling the acquisition cost and prediction accuracy trade-off.

Alternatively, the variations of model certainty weighted by feature costs can be used to define a reward function [7]:

$$r^t_{i,j} = \frac{||Cert(\boldsymbol{x}^t_i) - Cert(q(\boldsymbol{x}^t_i, j))||}{\boldsymbol{c}_j}, \quad (6)$$

where $Cert(x)$ represents the prediction certainty using a feature vector x. This method is shown to offer the state of the art results for cost-sensitive feature acquisition at test-time.

TABLE II: The summary of datasets and experimental settings.

Task	Instances	Features	Classes
Diabetes	92062	45	3
Heart Disease	49509	97	2
Hypertension	22270	31	2

Furthermore, it is highly scalable and applicable to online stream processing.

III. Experiments

In order to evaluate and provide baselines for the proposed dataset, in this section, we define specific cost-sensitive classification tasks and present comparison results for several recent cost-sensitive learning methods. Specifically, we compare: i) a method based on reinforcement learning where a hyper-parameter is balancing the cost vs. accuracy trade-off [5], ii) Opportunistic Learning (OL) [7] a method based on deep Q-learning with variations of model uncertainty as the reward function, iii) a method based on exhaustive measurements of the sensitivity [2], and iv) a method based on approximation of sensitivities using denoising autoencoders (FACT) [3]. Table II presents a summary of dataset tasks we defined including the number of instances, features, and classes for each. Due to space limitation, in the following, we provide a brief explanation of each task. For more details about specific features and setups please refer to the GitHub page.

We defined diabetes as the classification objective to predict the blood glucose categories that fall into the following three categories: normal (blood sugar less than 100), prediabetes (blood sugar between 100-125), and diabetes (blood sugar more than 125). We specifically defined and used 45 relevant features from demographics, questionnaire, examinations, and lab tests. Fig. 3 presents a comparison of results based on this

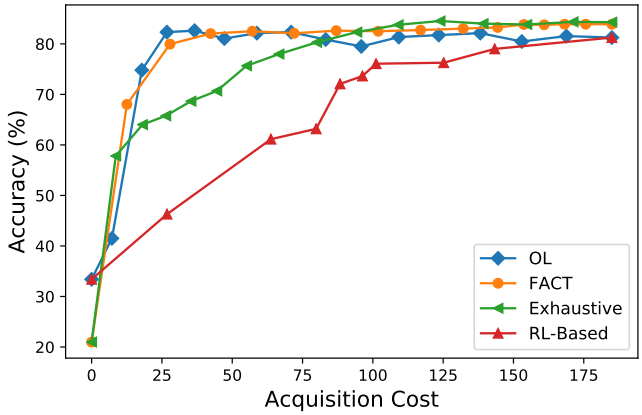


Fig. 3: Accuracy versus cost curve for the diabetes classification task comparing OL [7], FACT [3], Exhaustive [2], and RL-Based [5] methods.

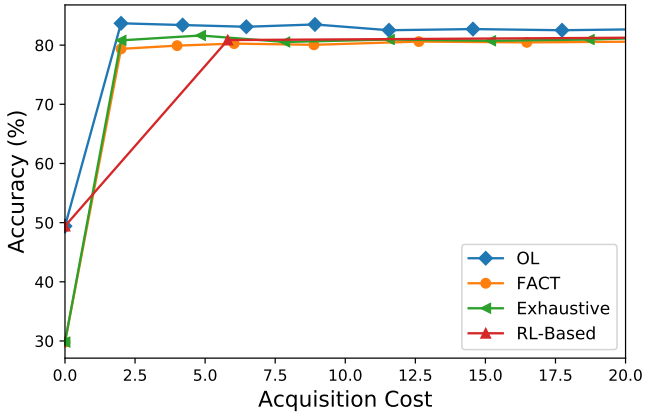


Fig. 5: Accuracy versus cost curve for the hypertension classification task comparing OL [7], FACT [3], Exhaustive [2], and RL-Based [5] methods.

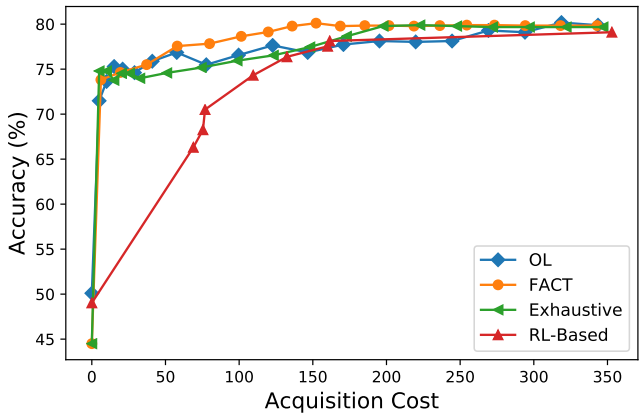


Fig. 4: Accuracy versus cost curve for the heart disease classification task comparing OL [7], FACT [3], Exhaustive [2], and RL-Based [5] methods.

task. As it can be seen from this figure, FACT and OL achieve superior results compared to other work.

As another task, we consider heart disease classification which is defined as binary classification task. Here, positive samples are individuals that reported any heart disease related issue in their history (e.g., heart attack, heart failure, etc.). For this task, we used the automated feature selection method as explained in Section II-B resulting in 97 features. Fig. 4 presents the comparison of results for this task. It can be inferred that this task is relatively simple and most approaches were able to achieve a reasonable performance.

At last, we consider the problem of predicting the existence of hypertension condition in individuals. Specifically, we consider subjects with systolic blood pressure of more than 140 mmHg as positive (hypertensive) class. Fig. 5 shows the performance of different methods on this dataset. It can be inferred from this figure that this task is relatively easy and all methods were able to achieve a reasonable performance. The only exception is the RL-based method which we were not able to find hyper-parameters resulting in cost values in the range of 0 and 5.

IV. Conclusion

In many machine learning applications, especially in healthcare, it is of paramount importance to consider feature acquisition costs. In this paper, we prepared a dataset consisting of nutrition and health data as well as feature acquisition costs for cost-sensitive learning in health domain. The prepared dataset consists of about 10,000 unique variables and can be used in defining various cost-sensitive studies in health. Furthermore, we compared the performance of the state-of-the-art approaches in the literature in three different classification problems including diabetes, heart disease, and hypertension classification.

References

[1] B. Krishnapuram, S. Yu, and R. B. Rao, *Cost-sensitive Machine Learning*. CRC Press, 2011.

[2] K. Early, S. E. Fienberg, and J. Mankoff, "Test time feature ordering with focus: interactive predictions with minimal user burden," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 992–1003.

[3] M. Kachuee, S. Darabi, B. Moatamed, and M. Sarrafzadeh, "Dynamic feature acquisition using denoising autoencoders," *IEEE transactions on neural networks and learning systems*, 2018.

[4] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang, "Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis," in *Advances in Neural Information Processing Systems*, 2018, pp. 7333–7342.

[5] J. Janisch, T. Pevnỳ, and V. Lisỳ, "Classification with costly features using deep reinforcement learning," *arXiv preprint arXiv:1711.07364*, 2017.

[6] H. Shim, S. J. Hwang, and E. Yang, "Why pay more when you can pay less: A joint learning framework for active feature acquisition and classification," *arXiv preprint arXiv:1709.05964*, 2017.

[7] M. Kachuee, O. Goldstein, K. Krkkinen, S. Darabi, and M. Sarrafzadeh, "Opportunistic learning: Budgeted cost-sensitive learning from data streams," in *International Conference on Learning Representations*, 2019.

[8] "National health and nutrition examination survey," 2018. [Online]. Available: https://www.cdc.gov/nchs/nhanes