Efficient Bayes Risk Estimation for Cost-Sensitive Classification

Daniel Andrade
Security Research Laboratories, NEC
s-andrade@cj.jp.nec.com

Yuzuru Okajima
Security Research Laboratories, NEC
y-okajima@bu.jp.nec.com

Abstract

In some real world applications, acquiring covariates for classification can be cost-intensive and should be limited as much as possible. For example, in the medical setting, a doctor cannot just perform all possible types of tests to classify whether the patient has diabetes or not. The decision of classifying or acquiring more covariates before classifying is dependent on the costs of new covariates and the expected optimal cost of misclassification (Bayes risk). However, estimating the latter is a formidable task due to the estimation of a high dimensional probability density and intractable integrals. In this work, we show that for linear classifiers this task can be considerably simplified, leading to a one dimensional integral for which we propose an efficient approximation. Experimental results on three datasets show consistent improvements over previously proposed methods for cost-sensitive classification. We also demonstrate that our proposed Bayes risk estimation procedure can benefit from additional unlabeled data which can be helpful when only small amount of labeled data is available.

1 Introduction

The traditional classification and regression tasks assume that all covariates can be provided at zero cost. In such settings, reducing the number of covariates is motivated by the increase in interpretability and reduction of model complexity (Tibshirani, 1996; Hastie et al., 2015). However, in some applications acquiring covariates incur costs which need to be balanced with the risk of misclassification. For example, in the medical setting, a doctor cannot just perform all possible types of tests to classify whether the patient has diabetes or not. Instead, depending on the observed symptoms of a patient, she might decide to conduct an additional blood examination or, without further tests, she might decide that the patient has no diabetes.

In such applications, the final goal is to minimize the total costs of classification, defined as the sum of the acquired covariates' costs and the expected misclassification costs (Bayes risk). While the objective can be stated formally rather straight forward (see Section 2), the optimization is complicated due to a combinatorially hard selection problem and the evaluation of a multi-dimensional integral for estimating the Bayes risk. While the former problem has some similarity to the problem of variable selection in (generalized) linear regression (see e.g. Tibshirani (1996); O'Hara et al. (2009)), the latter problem is unique to this task. Here in this work, we focus on the latter problem, and show that for linear classifiers the problem can be simplified to a one-dimensional integral for which we propose an efficient estimation. Compared to other methods, our proposed method experimentally leads to a reduction in total classification costs. Furthermore, our proposed method allows to include unlabeled data in a straight forward way, which further reduces costs.

In the next section, we introduce the necessary terminology and clarify the optimal classification strategy, followed by Section 3 where we introduce our proposed method. In Section 4, we compare our proposed method to several previous work on three standard datasets for cost-sensitive classification. In Section 5, we summarize the previous work on cost-sensitive classification. Finally, in Section 6 we summarize our conclusions.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

2 A cost rational selection criteria

Let $L := \{l_1, \ldots, l_c\}$ denote the set of class labels, and c_{y,y^*} the cost of classifying a sample as class y^*, when the true label is y. A decision procedure $\delta^* : \mathbb{R}^p \rightarrow L$ for which

$$\forall \delta : \mathbb{E}_{\mathbf{x},y}[c_{y,\delta(\mathbf{x})}] \geq \mathbb{E}_{\mathbf{x},y}[c_{y,\delta^*(\mathbf{x})}]$$

is called a Bayes procedure. The following procedure δ^* is a Bayes procedure (for a proof see, for example, Theorem 6.7.1 in Anderson (2003)):

$$\begin{aligned} \delta^*(\mathbf{x}) &= \operatorname*{arg\,min}_{y^* \in L} \sum_{y \in L} p(y|\mathbf{x}) \cdot c_{y,y^*} \\ &= \operatorname*{arg\,min}_{y^* \in L} \mathbb{E}_y[c_{y,y^*}] \, . \end{aligned} \tag{1}$$

The expected mis-classification cost of the Bayes procedure, i.e. $\mathbb{E}_{\mathbf{x},y}[c_{y,\delta^*(\mathbf{x})}]$ is called the Bayes risk.

Let us denote by $V := \{1, \ldots, p\}$ the index set of covariates with $V \cap L = \emptyset$. We denote the Bayes procedure for classifying a sample based only on the covariates $S \subseteq V$ by $\delta^*_S : \mathbb{R}^{|S|} \rightarrow L$. That means

$$\delta^*_S(\mathbf{x}_S) = \operatorname*{arg\,min}_{y^* \in L} \sum_{y \in L} p(y|\mathbf{x}_S) \cdot c_{y,y^*} \, . \tag{2}$$

When it is clear from the context, we drop the index on δ^*_S, and just write $\delta^*(\mathbf{x}_S)$ instead of $\delta^*_S(\mathbf{x}_S)$.[1]

2.1 Optimal Procedure

The classical definition of Bayes procedure does not consider the cost of covariate acquisition, and assumes that all covariates are acquired at once. Therefore, let us first formally extend the definition appropriately.

We use the following definition of a decision procedure.

Definition 1. *A function of the form*

$$\pi : \mathbb{R}^p \times 2^V \rightarrow L \cup V \, ,$$

which fulfills, $\forall \mathbf{x} \in \mathbb{R}^p, S \subseteq V$*:*

$$\pi(\mathbf{x}, S) = \pi(\mathbf{x} \odot \mathbf{1}_S, S) \, , \tag{3}$$

$$\pi(\mathbf{x}, S) \in L \cup (V \setminus S) \, , \tag{4}$$

is called a decision procedure.[2]

The condition in Equation (3) means that a decision procedure uses only the covariates that are indexed by S; the condition in Equation (4) means that a decision procedure cannot select a covariate that is already in S. In summary, the decision procedure $\pi(\mathbf{x}, S)$ either classifies the current sample, or selects a new covariate based on the observations $\mathbf{x}_S$. To simplify the notation, we write $\pi(\mathbf{x}_S)$ instead of $\pi(\mathbf{x}, S)$. Furthermore, we denote the cost of acquiring covariate i by c_i.

Given a sample $\mathbf{x}$ with class label y, we denote the loss of a decision procedure π as $l((\mathbf{x}, y), \pi)$. The loss can be computed recursively as follows. Let $l((\mathbf{x}, y), \pi) := l((\mathbf{x}, y), \pi, \emptyset)$, with

$$l((\mathbf{x}, y), \pi, S) = \begin{cases} c_{y,\pi(\mathbf{x}_S)} & \text{if } \pi(\mathbf{x}_S) \in L, \\ c_{\pi(\mathbf{x}_S)} + l((\mathbf{x}, y), \pi, S \cup \{\pi(\mathbf{x}_S)\}) & \text{else.} \end{cases}$$

If not stated otherwise, we assume that all costs are non-negative, i.e. $c_i \geq 0$, and $c_{y,y'} \geq 0$.

Theorem 1. *The decision procedure* π^* *defined by*

$$\pi^*(\mathbf{x}_S) = \operatorname*{arg\,min}_{i \in L \cup (V \setminus S)} \begin{cases} \mathbb{E}_y[c_{y,i}|\mathbf{x}_S] & \textit{if } i \in L \, , \\ c_i + \mathbb{E}_{\mathbf{x}_{V \setminus S}, y} \left[l((\mathbf{x}, y), \pi^*, S \cup \{i\}) | \mathbf{x}_S \right] & \textit{else.} \end{cases}$$

is a Bayes procedure. That means for any other decision procedure π *we have*

$$\mathbb{E}_{\mathbf{x},y}[l((\mathbf{x}, y), \pi^*)] \leq \mathbb{E}_{\mathbf{x},y}[l((\mathbf{x}, y), \pi)] \, .$$

The proof is given in the supplement material. We note that if the covariates are discrete, then we can formulate the problem as a stationary Markov decision process (MDP) where every policy leads to a terminal state (Zubek et al., 2004; Bayer-Zubek, 2004). The Bayes procedure from Theorem 1 is then equivalent to the optimal policy defined by the Bellman updates with the discounting factor set to 1 (Russell and Norvig, 2003).

For continuous covariates, implementing the exact decision procedure π^* is, in general, intractable. The reason is that in order to recursively evaluate the loss, we need to evaluate a sequence of interchanging minimizations and expectations. Therefore, our first relaxation is to pull-out all minimizations which leads to an upper bound:

$$\mathbb{E}_{\mathbf{x},y}[l((\mathbf{x}, y), \pi^*)] \leq \overbrace{\min_{S \subseteq V} \left(\mathbb{E}_{\mathbf{x}_S,y}[c_{y,\delta^*(\mathbf{x}_S)}] + \sum_{i \in S} c_i \right)}^{=:\mathcal{U}}$$

In the following, we denote this upper bound by $\mathcal{U}$. However, evaluating this upper bound is still computationally difficult for two reasons: First, iterating

[1] Remark about our notation: we denote by bold font a column vector, e.g. $\mathbf{x} \in \mathbb{R}^p$, and a column vector indexed by a set $A \subseteq V$ denotes the corresponding sub-vector, e.g. $\mathbf{x}_A \in \mathbb{R}^{|A|}$.

[2] $\odot$ denotes the Hadamard product, and $\mathbf{1}_S \in \mathbb{R}^p$ is the vector that is one in all positions indexed by S, and zero otherwise.

over all subsets $S \subseteq V$ is intractable for even moderate number of covariates. Second, evaluation of the expectation $\mathbb{E}_{\mathbf{x}_S,y}[c_{y,\delta^*(\mathbf{x}_S)}]$ requires to estimate the probability density of $\mathbf{x}_S$ and to solve an analytically intractable integral of dimension $|S|$.

3 Adaptive Cost-sensitive Forward Selection

For evaluating $\mathcal{U}$, instead of considering all subsets $S \subseteq V$, we limit our search to $S \in \mathfrak{S} = \{S_1, S_2, \ldots S_q\}$, which is such that $S_1 \subseteq S_2 \subseteq S_3 \ldots S_q \subseteq V$. Here we assume that $\mathfrak{S}$ is given, and helpful in the sense that it leads to a tight upper bound of $\mathcal{U}$. Later in Section 3.2, we explain a heuristic to find such a set of subsets.

Before we proceed, let us introduce our definition of future costs. Let $A \subseteq V$ and $S \subseteq V \setminus A$, then we define

$$F_{\mathbf{x}_A}(S) := \overbrace{\mathbb{E}_{\mathbf{x}_S,y}\left[c_{y,\delta^*(\mathbf{x}_{A\cup S})}|\mathbf{x}_A\right]}^{\text{(conditional) Bayes risk}} + \overbrace{\sum_{i \in S} c_i}^{\text{covariate costs}} \quad . \quad (5)$$

$F_{\mathbf{x}_A}(S)$ is the expected total additional cost of classification when we have already acquired the covariates A, and are planning to acquire additionally the covariates S before classifying. In particular, the upper bound $\mathcal{U}$ can be expressed as $\min_{S \subseteq V} F_{\mathbf{x}_\emptyset}(S)$.

Our approximation of the Bayes procedure π^* from Theorem 1 is given in Algorithm 1. First, we acquire all covariates indexed by S_1, and then check whether acquiring any additional covariates from $S_2 \backslash S_1, \ldots S_q \backslash S_1$ reduces the total cost of classification in expectation. If that is case, we acquire the covariates in $S_2 \backslash S_1$, and proceed analogously. If the total cost of classification is not expected to decrease with more covariates, we stop and classify based on the covariates acquired so far.

Algorithm 1: Adaptive Cost-sensitive Forward Selection (AdaCOS) for classifying a test sample.

```
Input: $S_1, \ldots, S_q$
$S_0 := \emptyset$
for $i \in \{1, \ldots, q-1\}$ do
    acquire $\mathbf{x}_{S_i \backslash S_{i-1}}$
    if $\forall j \in \{i+1, .., q\} : F_{\mathbf{x}_{S_i}}(S_j \setminus S_i) \geq F_{\mathbf{x}_{S_i}}(\emptyset)$
     then
        output class $\delta^*(\mathbf{x}_{S_i})$
    end
end
```

The algorithm is adaptive in the sense that the expected future costs $F_{\mathbf{x}_A}(S)$ depend on the covariates $\mathbf{x}_A$ observed so far. Therefore, we see that the effectiveness of the Algorithm hinges on the non-trivial task of calculating $F_{\mathbf{x}_A}(S)$.

3.1 Bayes Risk Estimation

The main challenge in evaluating the future costs is to estimate the multi-dimensional integral in $\mathbb{E}_{\mathbf{x}_S,y}\left[c_{y,\delta^*(\mathbf{x}_{A\cup S})}|\mathbf{x}_A\right]$. By assuming that the conditional class probability $p(y|\mathbf{x}_{A\cup S})$ can be modeled by a logistic regression model, we will show that it is possible to reduce the multi-dimensional integral into a one-dimensional with an effective approximation.

Let us denote, by $\boldsymbol{\beta}$ and τ, the regression coefficients and intercept, respectively, of the logistic regression model for $p(y|\mathbf{x}_{A\cup S})$. Furthermore, for simplicity, we assume that there are only two class labels $\{0, 1\}$, and $c_{0,0} = c_{1,1} = 0$.[3]

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_S,y}\left[c_{y,\delta^*(\mathbf{x}_{A\cup S})}|\mathbf{x}_A\right] \\
&= \mathbb{E}_{\mathbf{x}_S}\left[\sum_y c_{y,\delta^*(\mathbf{x}_{A\cup S})} p(y|\mathbf{x}_{A\cup S})|\mathbf{x}_A\right] \\
&= \mathbb{E}_{\mathbf{x}_S}\left[c_{0,\delta^*(\mathbf{x}_{A\cup S})} p(y=0|\mathbf{x}_{A\cup S})|\mathbf{x}_A\right] \\
&\quad + \mathbb{E}_{\mathbf{x}_S}\left[c_{1,\delta^*(\mathbf{x}_{A\cup S})} p(y=1|\mathbf{x}_{A\cup S})|\mathbf{x}_A\right].
\end{aligned}$$

Since

$$\begin{aligned}
\delta^*(\mathbf{x}_{A\cup S}) = \arg\min[&p(y=1|\mathbf{x}_{A\cup S}) \cdot c_{1,0}, \\
&p(y=0|\mathbf{x}_{A\cup S}) \cdot c_{0,1}],
\end{aligned}$$

we have

$$\begin{aligned}
\delta^*(\mathbf{x}_{A\cup S}) = 1 &\Leftrightarrow \frac{p(y=1|\mathbf{x}_{A\cup S}) \cdot c_{1,0}}{p(y=0|\mathbf{x}_{A\cup S}) \cdot c_{0,1}} \geq 1 \\
&\Leftrightarrow \frac{g(\boldsymbol{\beta}^T \mathbf{x}_{A\cup S} + \tau) \cdot c_{1,0}}{(1 - g(\boldsymbol{\beta}^T \mathbf{x}_{A\cup S} + \tau)) \cdot c_{0,1}} \geq 1 \\
&\Leftrightarrow e^{\boldsymbol{\beta}^T \mathbf{x}_{A\cup S} + \tau} \geq \frac{c_{0,1}}{c_{1,0}} \\
&\Leftrightarrow \boldsymbol{\beta}^T \mathbf{x}_{A\cup S} \geq \log(\frac{c_{0,1}}{c_{1,0}}) - \tau \\
&\Leftrightarrow \boldsymbol{\beta}_S^T \mathbf{x}_S \geq \log(\frac{c_{0,1}}{c_{1,0}}) - \tau - \boldsymbol{\beta}_A^T \mathbf{x}_A \\
&\Leftrightarrow z \geq z^*,
\end{aligned}$$

where we defined $z := \boldsymbol{\beta}_S^T \mathbf{x}_S$, and $z^* := \log(\frac{c_{0,1}}{c_{1,0}}) - \tau - \boldsymbol{\beta}_A^T \mathbf{x}_A$. We see that $\delta^*(\mathbf{x}_{A\cup S})$ depends only on z (random variable) and z^* (fixed). In the following, to simplify notation, let us denote by $h(z)$ the conditional distribution $p(z|\mathbf{x}_A)$, and by g the sigmoid function.

[3] Extension for allowing non-zero costs for $c_{y,y}$ is straight-forward, and omitted here.

We thus have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_S}\left[c_{1,\delta^*(\mathbf{x}_{A\cup S})}p(y=1|\mathbf{x}_{A\cup S})|\mathbf{x}_A\right]\\
&= \mathbb{E}_{\mathbf{x}_S}\left[c_{1,\delta^*(\mathbf{x}_{A\cup S})}g(\boldsymbol{\beta}_A^T\mathbf{x}_A+\boldsymbol{\beta}_S^T\mathbf{x}_S+\tau)|\mathbf{x}_A\right]\\
&= \mathbb{E}_z\left[c_{1,\delta^*(z,z^*)}g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)|\mathbf{x}_A\right]\\
&= \int c_{1,\delta^*(z,z^*)}g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\\
&= \int_{-\infty}^{z^*} c_{1,0}g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\\
&\quad + \int_{z^*}^{\infty} c_{1,1}g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\\
&= c_{1,0}\int_{-\infty}^{z^*} g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\\
&\quad + c_{1,1}\int_{z^*}^{\infty} g(z+\boldsymbol{\beta}_S^T\mathbf{x}_A+\tau)h(z)dz\\
&= c_{1,0}\int_{-\infty}^{z^*} g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\,.
\end{aligned}$$

And, analogously, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_S}\left[c_{0,\delta^*(\mathbf{x}_{A\cup S})}p(y=0|\mathbf{x}_{A\cup S})|\mathbf{x}_A\right]\\
&= c_{0,1}\int_{z^*}^{\infty} h(z)dz - c_{0,1}\int_{z^*}^{\infty} g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\,.
\end{aligned}$$

Thus the remaining task is to evaluate the following integral

$$\begin{aligned}
&\int_{a'}^{b'} g(z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau)h(z)dz\\
&= \int_{a'+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau}^{b'+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau} g(u)h(u-\boldsymbol{\beta}_A^T\mathbf{x}_A-\tau)du\,. \qquad (6)
\end{aligned}$$

We assume that $h(z) = p(z|\mathbf{x}_A)$ can be well approximated by a normal distribution with mean μ_z and variance σ^2. We defer the explanation of how to estimate μ_z and σ^2 to Section 3.1.1.

The integral in Equation (6) has no analytic solution. One popular strategy is to approximate the sigmoid function g by the cumulative distribution function of the standard normal distribution Φ, as in Gaussian process classification (Rasmussen and Williams, 2006). However, it turns out that this approximation is not applicable here, since a' or b' is bounded in our case. Instead, we use here the fact that the sigmoid function can be well approximated with only a few number of linear functions. In order to facilitate notation, let us introduce the following constants:

$$\begin{aligned}
a &:= a'+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau\,,\\
b &:= b'+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau\,,\\
\mu &:= \mu_z+\boldsymbol{\beta}_A^T\mathbf{x}_A+\tau\,.
\end{aligned}$$

Then we can write the integral in Equation (6) as

$$\int_a^b g(u)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du\,. \qquad (7)$$

Let us define the following piecewise linear approximation of the sigmoid function:

$$g(u) \approx \sum_{t=1}^{\xi+2}\Big(1_{[b_{t-1},b_t]}(u)\big(m_t u + v_t\big)\Big)\,,$$

where for $1 \leq t \leq \xi+1$, we set $b_t := -10 + \frac{20}{\xi}(t-1)$, and for $1 \leq t \leq \xi$, we set

$$m_{t+1} := \frac{g(b_{t+1})-g(b_t)}{b_{t+1}-b_t}\,, v_{t+1} \quad := g(b_t) - m_{t+1}b_t\,,$$

and

$$\begin{aligned}
&b_0 := -\infty\,,\ m_1 := 0\,,\ v_1 := g(b_1)\,,\\
&b_{\xi+2} := +\infty\,,\ m_{\xi+2} := 0\,,\ v_{\xi+2} := g(b_{\xi+1})\,,
\end{aligned}$$

and ξ is the number of linear approximations, which is, for example, set to 40. A comparison with the approximation $\Phi(\sqrt{\frac{\pi}{8}}u)$ is shown in Figure 1. That means for a relatively few number of linear approximations, we can achieve an approximation that is more accurate than the Φ-approximation. More importantly, as we show below, this allows for a tractable calculation of the integral in Equation (7), which is *not* the case when using the Φ-approximation. Then we have

$$\begin{aligned}
&\int_a^b g(u)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du\\
&= \int_a^b \sum_{t=1}^{\xi+2}\Big(1_{[b_{t-1},b_t]}(u)\big(m_t u + v_t\big)\Big)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du\\
&= \sum_{t=1}^{\xi+2} m_t\int_{\max(a,b_{t-1})}^{\min(b,b_t)} u\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du\\
&\quad + v_t\Phi_{\max(a,b_{t-1})}^{\min(b,b_t)}\,,
\end{aligned}$$

where we denote by $\Phi_l^o := \int_l^o \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du$ which can be well approximated with standard implementations. The remaining integral can also be expressed by Φ using the substitution $u-\mu := r$, we have

$$\begin{aligned}
&\int_l^o u\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(u-\mu)^2}du\\
&= \int_{l-\mu}^{o-\mu} r\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}r^2}dr + \mu\int_{l-\mu}^{o-\mu}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}r^2}dr\\
&= \frac{\sigma}{\sqrt{2\pi}}\big(e^{-\frac{1}{2\sigma^2}(l-\mu)^2} - e^{-\frac{1}{2\sigma^2}(o-\mu)^2}\big) + \mu\Phi_l^o\,.
\end{aligned}$$

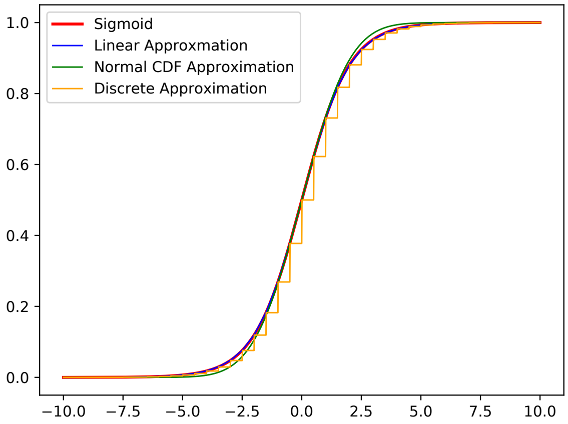

Figure 1: Comparison of Sigmoid function approximations. For the linear function approximation and the discrete bin approximation (Ji and Carin, 2007) we set $\xi = 40$. For the normal CDF approximation we use $\Phi(\sqrt{\frac{\pi}{8}}u)$.

3.1.1 Estimation of μ_z and σ^2

Recall that we assumed that $p(z|\mathbf{x}_A)$ is a normal density with mean μ_z and variance σ^2. In order to estimate μ_z and σ^2, we propose to model z given $\mathbf{x}_A$ as a regression problem with additive noise, where z is the response variable, and $\mathbf{x}_A$ are the explanatory variables. In detail, for learning the regression model from the training data $\{\mathbf{x}^{(k)}\}_{k=1}^{n}$, we prepare a collection of response and explanatory variable pairs of the form $\{(z^{(k)}, \mathbf{x}_A^{(k)})\}_{k=1}^{n}$, where $z^{(k)} = \boldsymbol{\beta}_S^T \mathbf{x}_S^{(k)}$. The important point to note is that for training the regression model, we do not require the class label y. As a consequence, additional to the class-labeled training data, we can exploit unlabeled training data (if available).

For our experiments, we use a standard Bayesian linear regression model with a scaled inverse χ^2 distribution prior on the noise variance (Gelman et al., 2013). However, we note that our choice is not limited to linear regression models, and we could also apply a non-parametric probabilistic regression model like Gaussian process regression (Rasmussen and Williams, 2006).

3.2 Subset Selection

Finally, we describe our heuristic for finding a sequence of subsets $S_1 \subseteq S_2, \ldots S_q \subseteq V$ for which the expected total cost of classification is minimal. We suggest to set $q := p + 1$, and use greedy forward selection as outlined in Algorithm 2.

Algorithm 2: Cost-sensitice forward selection (COS) for finding subsets $S_1 \subseteq S_2, \ldots S_{p+1} \subseteq V$.

```
Input: $\{\mathbf{x}^{(k)}\}_{k=n_l+1}^{n_l+n_u}$, $\{(y^{(k)}, \mathbf{x}^{(k)})\}_{k=1}^{n_l}$
$S_1 := \emptyset$
for $i \in \{1, \ldots, p\}$ do
    $S_{i+1} := \arg\min_{j \in V \setminus S_i} \mathbb{E}_{\mathbf{x}_{S_i}}[F_{\mathbf{x}_{S_i}}(\{j\})]$
end
```

Note that from the definition in Equation 5, we have

$$\mathbb{E}_{\mathbf{x}_S}[F_{\mathbf{x}_S}(\{j\})] = \mathbb{E}_{\mathbf{x}_S}\Big[\mathbb{E}_{x_j,y}[c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}})}|\mathbf{x}_S]\Big] + \sum_{i \in S} c_i \,.$$

In case, where unlabeled data $\{\mathbf{x}^{(k)}\}_{k=n_l+1}^{n_l+n_u}$ is available, we estimate $\mathbb{E}_{\mathbf{x}_S}[F_{\mathbf{x}_S}(\{j\})]$ using

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_S}\Big[\mathbb{E}_{x_j,y}[c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}})}|\mathbf{x}_S]\Big] \\ &= \mathbb{E}_{\mathbf{x}_{S\cup\{j\}}}\Big[\mathbb{E}_y[c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}})}|\mathbf{x}_{S\cup\{j\}}]\\ &\approx \frac{1}{n_u}\sum_{k=n_l+1}^{n_l+n_u}\sum_{y\in L} c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}}^{(k)})} p(y|\mathbf{x}_{S\cup\{j\}}^{(k)})\,, \end{aligned} \tag{8}$$

where the conditional class probability model $p(y|\mathbf{x}_{S\cup\{j\}})$ is learned with the labeled data $\{(y^{(k)}, \mathbf{x}^{(k)})\}_{k=1}^{n_l}$. Otherwise, we use a 10-fold cross-validation estimate:

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_S}\Big[\mathbb{E}_{x_j,y}[c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}})}|\mathbf{x}_S]\Big] \\ &\approx \frac{1}{|\mathscr{A}_f|}\sum_{k\in\mathscr{A}_f}\sum_{y\in L} c_{y,\delta^*(\mathbf{x}_{S\cup\{j\}}^{(k)})} p_f(y|\mathbf{x}_{S\cup\{j\}}^{(k)})\,, \end{aligned}$$

where $\mathscr{A}_f \subseteq \{1, \ldots, n_l\}$ and p_f is trained with $\{1, \ldots, n_l\} \setminus \mathscr{A}_f$. Note that we expect the estimate using the unlabeled data to perform better due to the fact that it uses more data to approximate the joint probability $p(x, y)$.

4 Experiments

We evaluate our method on three datasets that are frequently used for cost-sensitive classification: Pima Diabetes dataset (p = 8, n = 768), MiniBooNE particle identification dataset (p = 50, n = 130065), and the Wisconsin Breast Cancer dataset (p = 10, n = 683), all available at the UCI Machine Learning repository[4]. All datasets are binary classification tasks.

For the Diabetes and Breast Cancer datasets we use 5-fold crossvalidation; for the MiniBooNE dataset, from the whole dataset, we sample 5 random sets which are then split into labeled training data (500), test data

[4] https://archive.ics.uci.edu/ml/index.html

(1000), and unlabeled training data (10000).[5] In order to compare our method with previous work, we use a symmetric misclassification cost $c_{i,j} = c_{j,i}$, which is varied in the range between 100 and 1000. For the MiniBooNE and Breast Cancer dataset we set the covariate acquisition costs to 1, and the costs of correct classification $c_{i,i}$ to 0. For Diabetes, in order to compare our results to the ones reported in (Ji and Carin, 2007; Dulac-Arnold et al., 2012), we use the same covariate costs, and set the costs of correct classification to -50.

As baseline classifier, we use logistic regression with l2-regularization, where the regularization is determined using 10-fold cross-validation on the training data. The "full model" refers to the baseline classifier using all covariates. Due to the limited amount of labeled data, we found that for all datasets this baseline performed better than non-linear classifiers like SVM with RBF-Kernel.

The method COS refers to the method which selects one subset of covariates $S_{i*} \in \mathfrak{S}$ from Algorithm 2, for which the expected total cost of classification is minimal. The set of covariates S_{i*} is then fixed at test time.

The proposed method (AdaCOS) uses the sequence of covariate sets $\mathfrak{S}$, and decides at test time whether to follow the sequence (ask for more covariates) or classify by estimating the total costs as described in Algorithm 1.

We also run two other methods for cost-sensitive classification, namely GreedyMiser (Xu et al., 2012) and its extended version (Nan and Saligrama, 2017) for which the source code is publicly available and executable on a Linux environment.[6] The latter requires the specification of a high accuracy classification model for which we use the full model.

As evaluation measure, we use the average total cost of classification, defined as

$$\text{avg total cost} := \frac{1}{n_t} \sum_{k=1}^{n_t} \left(c_{y_k, y_k^*} + \sum_{i \in S_k} c_i \right),$$

where n_t is the number of test samples; y_k and y_k^* is the k-th true test class and predicted test class, respectively; S_k is the set of covariates that were used by the prediction model for classifying the k-th sample.

The results are shown in Figures 2 and 3. In Table 1 we also compare all methods to the results reported in (Ji and Carin, 2007; Dulac-Arnold et al., 2012). On all three datasets the proposed method AdaCOS shows the smallest total classification cost.

Table 1: Additional comparisons with methods DWSM (Dulac-Arnold et al., 2012) and POMDP (Ji and Carin, 2007). Total costs on Diabetes dataset when the misclassification costs are set to 400 and 800, respectively.

	400	**800**
AdaCOS	71.3 (8.95)	164.78 (14.78)
COS	74.7 (16.35)	170.34 (29.94)
Full Model	98.94 (5.7)	190.09 (10.77)
GreedyMiser	91.36 (14.43)	200.96 (31.93)
AdaptGbrt	90.41 (5.87)	200.92 (17.9)
DWSM	74.0 (-)	181.0 (-)
POMDP	75.0 (-)	180.0 (-)

In Figure 2, right-hand side, we also analyzed the impact of unlabeled data for the MiniBooNE data. The results suggest that a considerable reduction in total classification costs can be achieved by using the estimator in Equation 8.

We also investigated the average costs of covariates, i.e. $\frac{1}{n_t} \sum_{k=1}^{n_t} \sum_{i \in S_k} c_i$ that were acquired by each method. As shown in Figure 4, more (expensive) covariates are selected when the costs of wrong classification get higher. Furthermore, we can conclude that AdaCOS tends to select similar or less (expensive) covariates than other methods while achieving better accuracy.

5 Related Work

Here, we briefly summarize various previous works for cost-sensitive classification.

Markov Decision Process (MDP) Framework The MDP formulation and solution using an action-utility representation (Q-learning) in (Zubek et al., 2004; Bayer-Zubek, 2004) is closest to our approach. Their method also leads to a Bayes procedure, however, they do not provide a formal proof and consider only discrete covariates. The work in (Dulac-Arnold et al., 2011, 2012; Karayev et al., 2013) also uses the MDP framework. However, their proposed method cannot incorporate the uncertainty about the covariate distributions. The work in (Ji and Carin, 2007) tries to model such uncertainties by modeling the cost-sensitive classification problem as a partial observable Markov decision process (POMDP). However, their POMDP formulation can lead to repeatedly selecting

[5] Due to the small size of the Diabetes and Breast Cancer datasets, we did not prepare unlabeled training sets for those.

[6] Available at `http://kilian.cs.cornell.edu/code/code.html` and `https://github.com/fnan/AdaptApprox`.

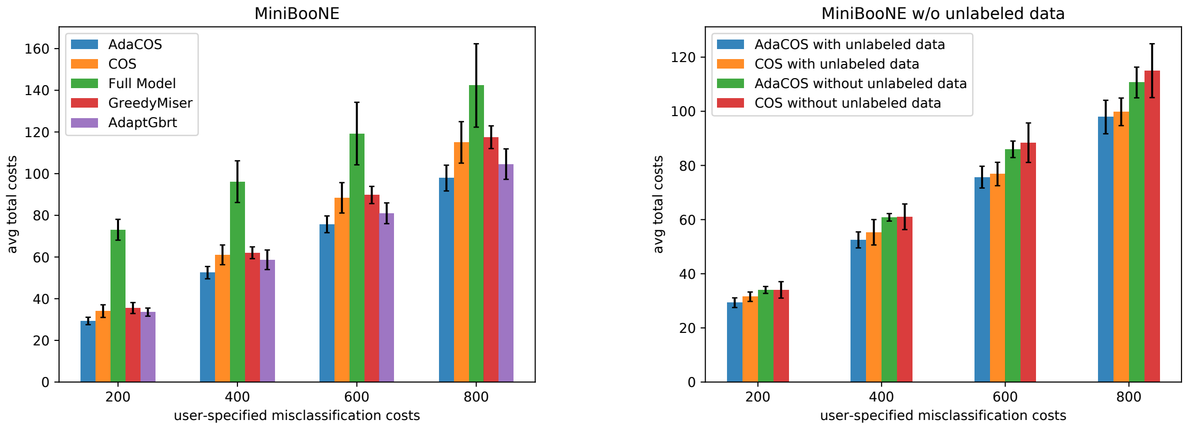

Figure 2: Average total cost of classification on MiniBooNE dataset for user-specified misclassification costs 200, 400, 600, and 800. Comparison between several methods (left), and comparison with and without unlabeled data (right).

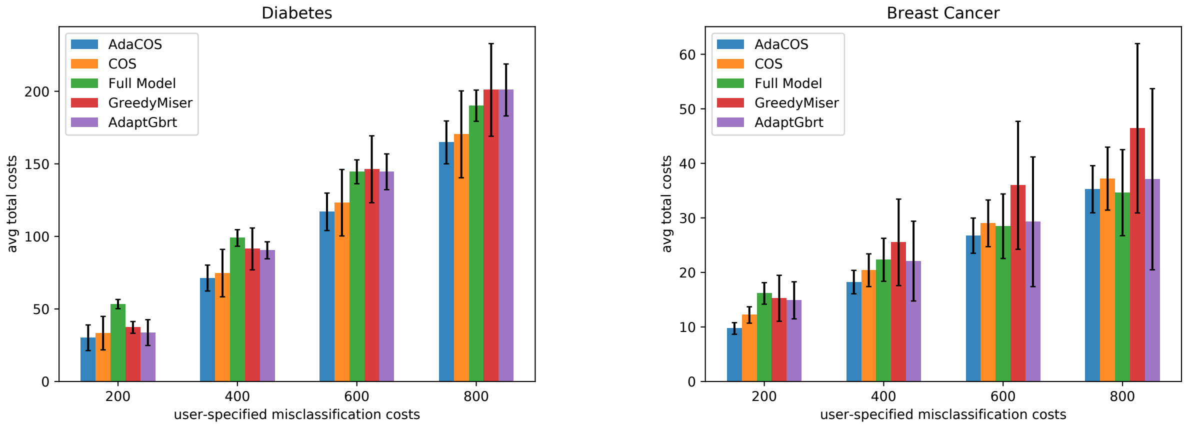

Figure 3: Average total cost of classification on Diabetes (left) and Breast cancer (right) datasets for user-specified misclassification costs 200, 400, 600, and 800.

the same covariates, and as a consequence they need to adapt the stopping criteria.

Reinforcement Learning Approaches Janisch et al. (2017) suggests to use deep reinforcement learning with Q-learning. In contrast to MDP, a discriminative decision maker is learned which does not require an environmental model. Their method performs promising in the domain where huge amounts of labeled training data is available. Alternatively, the work in (Benbouzid et al., 2012) suggests the use of SARSA. The method in (Contardo et al., 2016) also addresses this problem with reinforcement learning.

Discriminative Decision Approach The work in (Wang et al., 2015) proposes an intriguing method for finding a decision procedure that is guaranteed to converge to the Bayes risk given sufficient enough training data. Their idea is to create a Bayes optimal classifier for all possible subsets of covariates, and a directed a-cyclic graph that connects them. They formulate the problem as an empirical risk minimization (ERM) problem, and show that with infinitely many training samples the loss at each node converges to the Bayes risks. However, their method cannot incorporate unlabeled data. The work in (Trapeznikov and Saligrama, 2013; Wang et al., 2014b) uses a similar framework but restricted to a fixed sequential order.

Cost-sensitive Tree Construction The work in (Xu et al., 2012; Nan et al., 2015, 2016; Nan and Saligrama, 2017; Peter et al., 2017) learns a random forest subject to budget constraints on the features. In particular, the methods in (Nan and Saligrama, 2017;

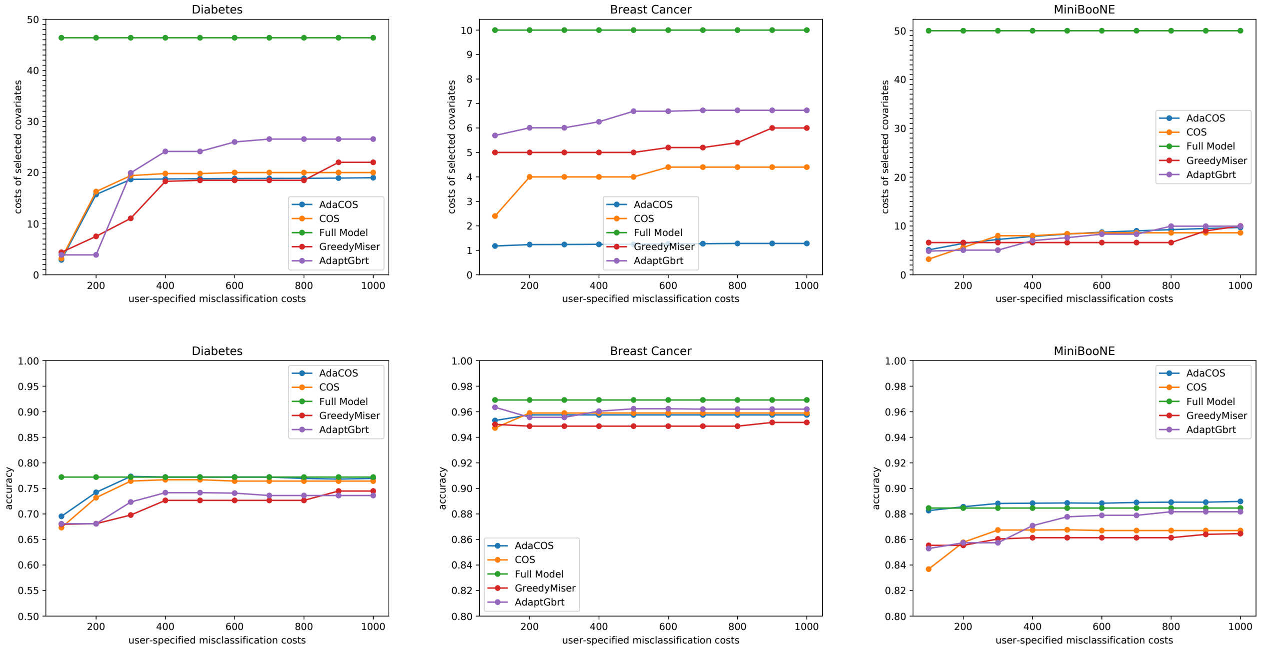

Figure 4: Costs of selected covariates (top row) and accuracy (bottom row) on Diabetes (left) and Breast cancer (right) and MiniBooNE (bottom).

Peter et al., 2017) are considered state of the art for this task. Their usage of gradient boosted decision trees (Friedman, 2001) makes them in particular effective for very large training data. Cost-sensitive decision trees for discrete covariates are also considered in (Sheng and Ling, 2006), and extended to Bayesian Networks in (Bilgic and Getoor, 2007).

Tree of Classifiers The work in (Kusner et al., 2014; Xu et al., 2013) proposes to learn a tree of classifiers that minimizes a convex surrogate loss subject to budget constraints. Wang et al. (2014a) assumes a fixed number of pre-trained classifiers and the goal is to learn a policy that selects one of those classifiers.

Entropy-Based Approaches The work in (Kanani and Melville, 2008; Gao and Koller, 2011; Kapoor and Horvitz, 2009) optimizes a criteria that combines the costs of features with an estimate of the class entropy of the resulting classifier. As such their objective function is different from ours.

Others The work in (Greiner et al., 2002) extends the Probably Approximately Correct (PAC) framework to prove the existence of a cost-sensitive classifier that is with high probability optimal in the sense of providing minimal average total costs. However, they assume a probability distribution over only discrete covariates. The method in (Lakkaraju and Rudin, 2017) is additionally focused on interpretability, and, as a consequence, optimizes an objective function that is different from ours. Imitation learning is also applied to this task by He et al. (2012), but their definition of loss is different from minimizing the total classification costs that we consider here. The work in (Nan et al., 2014) assumes a margin-based classifier and uses a k-nearest neighbor approach to estimate the accuracy of the classifier.

6 Conclusions

In this article, we showed how the optimal covariate acquisition and classification decision could be achieved, in principle (Theorem 1), and proposed an approximation (AdaCOS) which enables an efficient estimation of the Bayes risk. Although our approach is based on strong assumptions (linear sequence of covariates sets, logistic regression model), we confirmed experimentally that our proposed approach outperforms several previous methods. Furthermore, in contrast to previous work, our framework allows to exploit unlabeled data for the Bayes risk estimation. Our experiments confirmed that this can help to further reduce the total cost of classification.

References

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 2003.

Valentina Bayer-Zubek. Learning diagnostic policies from examples by systematic search. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 27–34. AUAI Press, 2004.

Djalel Benbouzid, Röbert Busa-Fekete, and Balázs Kégl. Fast classification using sparse decision dags. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 747–754, 2012.

Mustafa Bilgic and Lise Getoor. Voila: Efficient feature-value acquisition for classification. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1225. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

Gabriella Contardo, Ludovic Denoyer, and Thierry Artières. Sequential cost-sensitive feature acquisition. In *International Symposium on Intelligent Data Analysis*, pages 284–294. Springer, 2016.

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. Datum-wise classification: a sequential approach to sparsity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 375–390. Springer, 2011.

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. Sequential approaches for learning datum-wise sparse representations. *Machine learning*, 89(1-2):87–122, 2012.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Tianshi Gao and Daphne Koller. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems*, pages 1062–1070, 2011.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Russell Greiner, Adam J Grove, and Dan Roth. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 139(2):137–174, 2002.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

He He, Jason Eisner, and Hal Daume. Imitation learning by coaching. In *Advances in Neural Information Processing Systems*, pages 3149–3157, 2012.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. *arXiv preprint arXiv:1711.07364*, 2017.

Shihao Ji and Lawrence Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

Pallika Kanani and Prem Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Advances In Neural Information Processing Systems (NIPS)*, 2008.

Ashish Kapoor and Eric Horvitz. Breaking boundaries: Active information acquisition across learning and diagnosis. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 898–906. Curran Associates Inc., 2009.

Sergey Karayev, Mario J Fritz, and Trevor Darrell. Dynamic feature selection for classification on a budget. In *International Conference on Machine Learning (ICML): Workshop on Prediction with Sequential Models*, 2013.

Matt J Kusner, Wenlin Chen, Quan Zhou, Zhixiang Eddie Xu, Kilian Q Weinberger, and Yixin Chen. Feature-cost sensitive learning with submodular trees of classifiers. In *AAAI*, pages 1939–1945, 2014.

Himabindu Lakkaraju and Cynthia Rudin. Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, pages 166–175, 2017.

Feng Nan and Venkatesh Saligrama. Adaptive classification for prediction under a budget. In *Advances in Neural Information Processing Systems*, pages 4730–4740, 2017.

Feng Nan, Joseph Wang, and Venkatesh Saligrama. Feature-budgeted random forest. In *International Conference on Machine Learning*, pages 1983–1991, 2015.

Feng Nan, Joseph Wang, and Venkatesh Saligrama. Pruning random forests for prediction on a budget. In *Advances in neural information processing systems*, pages 2334–2342, 2016.

Feng Nan, Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. Fast margin-based cost-sensitive classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2952–2956. IEEE, 2014.

Robert B O'Hara, Mikko J Sillanpää, et al. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

Sven Peter, Ferran Diego, Fred A Hamprecht, and Boaz Nadler. Cost efficient gradient boosting. In *Advances in Neural Information Processing Systems*, pages 1550–1560, 2017.

Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. *MIT Press*, 2006.

Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach. 2003.

Victor S Sheng and Charles X Ling. Feature value acquisition in testing: a sequential batch test algorithm. In *Proceedings of the 23rd international conference on Machine learning*, pages 809–816. ACM, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Kirill Trapeznikov and Venkatesh Saligrama. Supervised sequential classification under budget constraints. In *Artificial Intelligence and Statistics*, pages 581–589, 2013.

Joseph Wang, Tolga Bolukbasi, Kirill Trapeznikov, and Venkatesh Saligrama. Model selection by linear programming. In *European Conference on Computer Vision*, pages 647–662. Springer, 2014a.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. An lp for sequential learning under budgets. In *Artificial Intelligence and Statistics*, pages 987–995, 2014b.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. Efficient learning by directed acyclic graph for resource constrained prediction. In *Advances in Neural Information Processing Systems*, pages 2152–2160, 2015.

Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. Cost-sensitive tree of classifiers. In *International Conference on Machine Learning*, pages 133–141, 2013.

Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. The greedy miser: learning under test-time budgets. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1299–1306. Omnipress, 2012.

Valentina Bayer Zubek, Thomas Glen Dietterich, et al. Pruning improves heuristic search for cost-sensitive learning. 2004.