Aalto University

Sundin, Iiris; Schulam, Peter; Siivola, Eero; Vehtari, Aki; Saria, Suchi; Kaski, Samuel

Active Learning for Decision-Making from Imbalanced Observational Data

Published in:
Proceedings of the 36th International Conference on Machine Learning

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

Active Learning for Decision-Making from Imbalanced Observational Data

Iiris Sundin [1] **Peter Schulam** [* 2] **Eero Siivola** [* 1] **Aki Vehtari** [1] **Suchi Saria** [2] **Samuel Kaski** [1]

Abstract

Machine learning can help personalized decision support by learning models to predict individual treatment effects (ITE). This work studies the reliability of prediction-based decision-making in a task of deciding which action a to take for a target unit after observing its covariates $\tilde{x}$ and predicted outcomes $\hat{p}(\tilde{y} \mid \tilde{x}, a)$. An example case is personalized medicine and the decision of which treatment to give to a patient. A common problem when learning these models from observational data is imbalance, that is, difference in treated/control covariate distributions, which is known to increase the upper bound of the expected ITE estimation error. We propose to assess the decision-making reliability by estimating the ITE model's Type S error rate, which is the probability of the model inferring the sign of the treatment effect wrong. Furthermore, we use the estimated reliability as a criterion for active learning, in order to collect new (possibly expensive) observations, instead of making a forced choice based on unreliable predictions. We demonstrate the effectiveness of this decision-making aware active learning in two decision-making tasks: in simulated data with binary outcomes and in a medical dataset with synthetic and continuous treatment outcomes.

1. Introduction

A promising application domain of machine learning is to augment human intelligence in decision-making tasks by providing predictions of outcomes under alternative actions (Schulam & Saria, 2017). To fit a model to this task, we need data recording previous actions a, observed outcomes y, and any features relevant to the context of the

---
[*]Equal contribution [1]Department of Computer Science, Aalto University, Espoo, Finland [2]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. Correspondence to: Iiris Sundin <iiris.sundin@aalto.fi>.

decision, x. Then, the goal is to estimate $p(Y \mid X = x, A = a)$ and, further, individual treatment effect (ITE), $\tau(x) = \mathbb{E}[Y[1] - Y[0] \mid X = x]$, where $Y[a]$ denotes the potential outcome of treatment $A = a$ (Rubin, 1978). ITE provides sufficient information to choose between two actions. The estimation of ITE is susceptible to many error sources (Pearl, 2009; Schulam & Saria, 2017; Mitchell et al., 2018), of which we concentrate on imbalance (Gelman & Hill, 2007).

Imbalance is defined as the difference in covariate distributions in the treated and control groups. Imbalance makes correct model specification essential for avoiding bias in treatment effect estimates (Gelman & Hill, 2007). Mis-specification of the model could be avoided by using non-parametric models, but their variance increases quickly when extrapolating. Recently, imbalance has been shown to increase the upper bound of the model error in estimation of ITE (Shalit et al., 2017). Furthermore, imbalance becomes the more prevalent issue the higher-dimensional the covariate space is (D'Amour et al., 2018).

There are many existing ways to deal with imbalance when learning the average treatment effect (ATE). In causal inference, the most common methods are propensity score matching or weighting (Rosenbaum & Rubin, 1984; Hirano et al., 2003; Lunceford & Davidian, 2004), and modeling the potential outcomes (Imbens & Rubin, 2015; Hernán & Robins, 2018), as well as doubly robust methods which implement both (Bang & Robins, 2005; Funk et al., 2011).

Even though these methods can decrease bias in treatment effect estimates, they will increase variance, and therefore may make the decision-making less reliable. This is especially the case with ITE; For example, in the areas of covariate space where there are more control units than treated, intuition is that the model for the treated outcome either has to generalize from less-representative observations (increasing bias) or extrapolate (increasing variance). Either way, there is higher *uncertainty* about the treated outcome, which makes reliable decision-making difficult. An extreme case of this is illustrated in Fig. 1. A natural question then is, could other data sources be exploited instead of making a forced choice based on insufficient observational data.

This work has three main contributions, which are complementary to each other and can be used independently. First,

we describe how imbalance decreases *decision-making performance* by increasing Type S error rate (Gelman & Tuerlinckx, 2000), which is the probability of the model inferring the sign of the treatment effect wrong. Second, we propose a Bayesian *estimate for the Type S error rate*, which allows quantifying the reliability of a decision-support model. Third, we propose to alleviate the consequences of imbalance by actively collecting more data. To this end, we introduce *decision-making aware active learning criteria* that improve decision-making performance by minimizing the estimated Type S error rate.

Finally, in many cases there are restrictions on what can be measured. For example, in medicine it is in general not ethical to do an experiment on a patient in order to get information to improve the treatment plan of another patient. Therefore, regular active learning would not be possible and, instead, any new information has to be acquired indirectly. For this reason, we introduce the idea of *counterfactual elicitation* which means soliciting indirect observations about counterfactual outcomes, that is, what would have been the outcome had x been treated with a' instead of a. We demonstrate the effectiveness of the proposed active learning criteria applied to counterfactual elicitation.

Main claims of this paper: 1. Reliability of decision-making is hindered by imbalance in data. 2. Bayesian estimate of the probability of error is a good estimate for the decision-making performance. 3. If there is a way to acquire more data (a simulator, new experiment, ask an expert), *decision-making error can be minimized by decision-making aware active learning*.

Technical contributions

1. We state sufficient conditions under which imbalance will cause decision-making error, measured in Type S error rate.

2. We introduce the principle of decision-making aware active learning, and propose a decision-making aware acquisition function. Our formulation allows both continuous and discrete outcomes.

3. We propose two types of queries for counterfactual elicitation. (a) Observation is a scalar-valued point estimate of a counterfactual outcome from a noisy oracle. (b) Observation is a (potentially erroneous) pairwise comparison between factual and counterfactual outcomes for one unit in the training data.

4. We show empirically that the proposed estimate of the Type S error rate has strong correlation with the observed Type S error rate, and that the active learning that aims at minimizing the Type S error rate increases the decision-making performance faster than standard active learning methods.
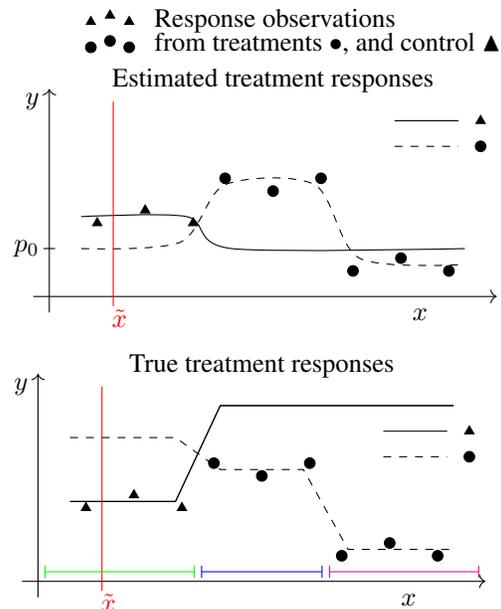

Figure 1. An example decision-making task is to choose a treatment for a specific $\tilde{x}$ (red mark). Upper graph shows the posterior means of the potential outcome models given observations, and the prior mean p_0. The true responses are in the lower graph, which shows that there are three regions with different response types (marked in green, blue and magenta). Lower y is better. The problem is that there are no observations about treatment ● in the green region, which causes it to appear to be the best choice for $\tilde{x}$, although that is incorrect.

2. Related Work

Active learning is commonly used to acquire class labels during model training to improve classification performance, by selecting unlabeled training instances for humans to label. The selection criteria are usually based on uncertainty and correlations of the training instances, see e.g. Fu et al. (2013) for a survey. Some active learning works also consider richer input than just labels, for example about importance of features (Brooks et al., 2015; Ribeiro et al., 2016). Active learning proposed by Javdani et al. (2014) aims at improving automated decision-making by reducing model uncertainty so that the remaining hypotheses are confined to the same decision region. The importance of active learning for decision-making tasks has been noted in other fields, where e.g. Saar-Tsechansky & Provost (2007) developed a heuristic method for deciding which consumers to target in marketing campaigns.

Active learning has also been used to design interventions that improve identifiability of causal networks (Hauser & Bühlmann, 2014). In addition, Bottou et al. (2013) studies carefully how counterfactual inference can be used for active learning. Closest to our work is the work on active

learning with logged data (Yan et al., 2018), which proposes a de-biasing query strategy for a classification task. The difference to our work is that they assume the propensities (probability of revealing the label) to be known.

Using data to estimate the effect of interventions has been extensively studied in the field of causal analysis (see e.g. Pearl (2009); Morgan & Winship (2014); Imbens & Rubin (2015); Hernán & Robins (2018)). In one school of thought, analysts construct a "causal directed acyclic graph" embodying substantive knowledge about the domain, and predict the effect of interventions using operations on the graph (Pearl, 2009; Morgan & Winship, 2014). Another common approach is to model interventions using *potential outcomes*, where separate random variables are constructed to represent the target outcome under each possible action (Neyman, 1923; Rubin, 1978; Neyman, 1990; Imbens & Rubin, 2015; Hernán & Robins, 2018). Our work builds on recent research on using ideas from causal analysis to learn individual treatment effects. The issue of imbalance is discussed by Johansson et al. (2016), and they propose an approach based on empirical risk minimization and domain shift to improve predictions. Xu et al. (2016) estimate the individual treatment effects in a longitudinal setting where individual-specific treatment parameters are refined over time as more observations are collected. Alaa & van der Schaar (2017) use Gaussian processes to model individual-specific outcomes under alternative treatments and prove minimax rates on the risk that the approach achieves. Our work builds on these ideas, and is most closely related to the works by Xu et al. (2016) and Alaa & van der Schaar (2017). Although we also use Gaussian processes to model treatment effects, our work is unique in that it leverages the probabilistic framework to design an active learning algorithm to improve the decision-making process.

3. Problem Formulation

3.1. Setup

Let $p(y[a] \mid \mathbf{x})$ be the distribution, and $\hat{p}(y[a] \mid \mathbf{x})$ our probabilistic model of the the potential outcome $Y[a] \in \mathbb{R}$ of an action $a \in \{0, 1\}$, given covariates $\mathbf{x} \in \mathcal{X}$, e.g. $\mathcal{X} = \mathbb{R}^d$. The decision-maker has a *policy* for choosing which action to take for a unit $\mathbf{x}$, based on its predicted *individual potential outcomes* $\hat{p}(y[a] \mid \mathbf{x})$. The outcome model has been learned in retrospect from observational data, which may be *imbalanced* in the observed actions. The observational data D is a set of n observations $\{y_i, a_i, \mathbf{x}_i\}_{i=1}^n$, where y_i is the observed outcome of action a_i for unit $\mathbf{x}_i$. Imbalance means that the covariate distributions are different in the treated ($a_i = 1$) and control groups ($a_i = 0$).

The *active learning task* is to sequentially improve the decision-making performance under the decision-maker's policy, by making queries about counterfactual outcomes $Y[a] \mid X = \mathbf{x}$. In this work, we use the objective to improve the decisions for a particular target unit $\tilde{\mathbf{x}}$, but the proposed method applies to multiple targets as well.

3.2. Assumptions for Causal Inference

We assume that the unknown policy used to choose actions in the training data only depends on the observed covariates $\mathbf{x} \in \mathcal{X}$. This is equivalent to the no unmeasured confounders assumption (Hernán & Robins, 2018) and implies that all confounders are included in $\mathbf{x}$.

We further assume consistency of potential outcomes, which means that the potential outcomes $p(y[a] \mid \mathbf{x}) = p(y \mid X = \mathbf{x}, A = a)$ can be directly estimated from the observed outcomes in the training data. Regardless of this, imbalance will still cause issues.

4. Methods

Preliminaries. Causal effect of a treatment is the difference between outcomes when a unit $\mathbf{x} \in \mathcal{X}$ is treated and not treated, where $\mathcal{X}$ is the population. Individualized treatment effect is defined as $\tau(\mathbf{x}) = \mathbb{E}[Y[1] - Y[0] \mid X = \mathbf{x}]$, where $a = 1$ means treated and $a = 0$ not treated, i.e. control. Fundamental problem in causal inference is that we cannot observe both potential outcomes for the same unit $\mathbf{x}$. There exists extensive work on how to learn estimates of the individualized treatment effects $\hat{\tau}(\mathbf{x})$ despite this limitation, e.g. Hill (2011); Johansson et al. (2016); Shalit et al. (2017); Alaa & van der Schaar (2017); Wager & Athey (2018).

In this work, the task is to decide which action a to choose for $\tilde{\mathbf{x}}$. Decision-making performance is measured as the probability of correct decision, or equivalently, the proportion of correct decisions in repeated decision-making tasks.

Definition: Type S error rate γ is the probability of the model inferring the sign of the treatment effect wrong; $\gamma \colon \mathcal{M} \times P_{\mathcal{X}, \mathcal{Y}} \to [0, 1]$ where $\mathcal{M}$ is the model space and $P_{\mathcal{X}, \mathcal{Y}}$ is the true treatment effects (distributions over $\mathcal{X} \times \mathcal{Y}$). The expected Type S error rate in $\mathcal{X}$ is $\gamma = \mathbb{E}_{P_{\mathcal{X}, \mathcal{Y}}}[\mathbb{I}(\text{sign}(\hat{\tau}) \neq \text{sign}(\tau))]$, where $\mathbb{I}(A) = 1$ if condition A is true, and 0 otherwise.

The expected proportion of correct decisions in population $\mathcal{X}$ is then $1 - \gamma$, which makes Type S error rate a natural measure of the decision-making performance.

We assume that the decision-maker's policy is to choose the action with the highest expected utility of the outcome. Without loss of generality we will assume that the utility is directly the outcome y (higher better); more sophisticated utilities are discussed in Section 6.

4.1. Effect of Imbalance on Type S Error Rate

In this section we prove that, under certain assumptions, imbalance increases the error rate in decision-making. We start by a sketch of the proof and then continue with details.

Sketch of proof: First, we assume a probabilistic model of potential outcomes, with broad prior distributions. This implies that when the sample size is small, posteriors will be wide. Then, we show that imbalance decreases the expected number of samples locally, therefore increasing the Type S error rate locally. Finally, we provide conditions under which local increase in Type S error rate also increases the expected global Type S error rate.

Assumption 1. (Prior). Assume a broad prior on the expected potential outcomes μ_a: $p(\mu_a) > D > 0 \; \forall \mu_a \in [-K, K]$.

Assumption 2. (Likelihood). Likelihood of observation $p(y_a \mid \mu_a) > C > 0 \; \forall y_a \in [-K, K]$.

Comment. Consequence of Assumptions 1 and 2 is that if sample size is small, the posterior will be wide.

Lemma 1. Given observations on two potential outcomes $\mathcal{D} = \{y_{1,i}\}_{i=1}^{n_1} \cup \{y_{0,j}\}_{j=1}^{n_0}$, probability of Type S error has lower bound $p(\text{“Type S error”}) > 2K^2 D^2 C^{n_1+n_0}$.

Assumption 3. (Covariate distributions). Let $p^a(x) := p(x \mid a)$ be the covariate distribution of group a in covariate space $\mathcal{X}$. Assume $p^a(x)$ are Lipschitz continuous with constant L.

Definition: (Imbalance). Imbalance can be measured using Integral Probability Metric as described by Shalit et al. (2017). Let G be a function family consisting of functions $g : \mathcal{X} \to \mathbb{R}$. For a pair of distributions p, q over $\mathcal{X}$ the Integral Probability Metric is defined as

$$IPM_G(p, q) = \sup_{g \in G} \left| \int_{\mathcal{X}} g(x)(p(x) - q(x))dx \right|. \quad (1)$$

Assumption 4. (Imbalance). Assume there exists non-empty $\Omega = \{x \in \mathcal{X} \mid |p^1(x) - p^0(x)| \geq h\}$ where $h > 0$. (For small enough h this holds if there is any imbalance).

Lemma 2. Let r be the smallest radius $r' > 0$ of a neighborhood $B_{r'}(x_e)$ of $x_e \in \Omega$, such that $|p^1(x) - p^0(x)| = 0$ for some x in the border $\partial B_r(x_e)$. Given assumptions 3 and 4, then $r \geq \frac{h}{2L}$ for all $x_e \in \Omega$.

Corollary of Lemma 2. Mark $P^a_{B_r(x_e)} = \int_{B_r(x_e)} p^a(x)dx$. By Lemma 2 and Assumption 3, $P^a_{B_r(x_e)} > P^{1-a}_{B_r(x_e)}$. This means that the expected number of observations from the group $1 - a$ is lower than from the group a in $B_r(x_e)$.

The following theorem is our main result:

Theorem 1. Let N be the sample size, and a the treatment with the higher number of observations in $B_r(x_e)$, and $x \in \mathbb{R}$. Then the expected probability of Type S error in $B_r(x_e)$ has lower bound

$$p(\text{“Type S error”}) > 2K^2 D^2 C^{N(P^a_{B_r(x_e)} - (1-p(a))\frac{h^2}{2L})}.$$

Comment. Theorem 1 shows that, with fixed r, N and $p(a)$, the larger the local imbalance (h) in $B_r(x_e)$, the higher the Type S error rate in $B_r(x_e)$ is. Higher-dimensional cases are considered in the supplementary.

Now, we have shown that imbalance increases locally the Type S error rate. Then the question remains whether the error rate increases globally as well, or do the local effects cancel out each other. We prove this in one-dimensional case, but we see no reason why the proof would not extend to higher dimensions as well. The following assumption and theorem give conditions under which imbalance increases the global Type S error rate.

Assumption 5. Assume the following balanced and imbalanced settings. In the balanced setting, let $p^1(x) = p^0(x) = p(x)$, and $x \in \mathbb{R}$. Without loss of generality we assume that imbalance arises from a shift in $p^0(x)$, s.t. in the imbalanced setting $p^0(x) = p^1(x) - \eta(x)$, where $\eta(x) \in \mathbb{R}$, and $\int \eta(x)dx = 0$.

Theorem 2. Denote $P_{\eta \geq h} = \int_{\mathcal{X}} \mathbb{I}(\eta(x) \geq h)p_t(x)dx$, where $p_t(x)$ is the covariate distribution in the test set. Given Assumption 5, if $P_{\eta \geq h} > C^{N(1-u)h}$, then imbalance $\eta(x)$ increases the lower bound of the expected global Type S error rate in $\mathcal{X}$.

Proofs in the supplementary.

4.2. Estimated Type S Error Rate

In the Bayesian sense, the model $\hat{p}(y[a] \mid \mathbf{x}, D)$ captures our current understanding of the problem, and therefore the estimated Type S error rate is $\hat{p}(y[1] < y[0] \mid \mathbf{x}, D)$ if the expected effect is positive, that is, if $\mathbb{E}_{\hat{p}(y[1]\mid\mathbf{x},D)}[y[1]] > \mathbb{E}_{\hat{p}(y[0]\mid\mathbf{x},D)}[y[0]]$. Respectively, if the expected effect is negative, then the estimated error rate is $\hat{p}(y[1] > y[0] \mid \mathbf{x}, D)$.

We analyze the properties of the estimated Type S error rate in the model family of linear-parameter regression models, using standard Bayesian linear regression with basis functions. The observation model is $y[a] \mid \mathbf{x} \sim N(w^\top \phi(\mathbf{x}, a), \sigma_0^2)$, where $\phi(\mathbf{x}, a)$ are the basis functions.

The regression weights have Gaussian prior distribution $w \sim N(0, \alpha I)$. Assuming σ_0 and α are known, the posterior predictive distributions of potential outcomes $\hat{p}(\tilde{y}[a] \mid \tilde{\mathbf{x}}, D)$ are Gaussian, with mean and variance

$$\hat{\mu}_a(\tilde{\mathbf{x}}) = \frac{1}{\sigma_0^2}\phi(\tilde{\mathbf{x}}, a)^\top S_a \Phi_a^\top y_a \text{ and}$$

$$\hat{\sigma}_a^2(\tilde{\mathbf{x}}) = \sigma_0^2 + \phi(\tilde{\mathbf{x}}, a)^\top S_a \phi(\tilde{\mathbf{x}}, a), \text{ where}$$

$$S_a^{-1} = \alpha \mathbf{I} + \frac{1}{\sigma_0^2} \Phi_a^\top \Phi_a,$$

where Φ_a is a matrix containing the feature vectors $\phi(\mathbf{x}_i, a)$ of the $\mathbf{x}_i$ that were treated with a, and similarly y_a is a vector of observed outcomes of a.

The treatment effect $\tilde{\tau} \mid \tilde{\mathbf{x}} = \tilde{y}[1] - \tilde{y}[0] \mid \tilde{\mathbf{x}} \sim N(\hat{\mu}_1 - \hat{\mu}_0, \hat{\sigma}_{a=1}^2(\tilde{\mathbf{x}}) + \hat{\sigma}_{a=0}^2(\tilde{\mathbf{x}}))$. For simplicity, assume that the expected treatment effect $\mathbb{E}_{\hat{p}(\tilde{\tau}|\tilde{\mathbf{x}}, D)}[\tilde{\tau}]$ is positive. Then it can easily be shown that the estimated Type S error rate in a test unit $\tilde{\mathbf{x}}$ is

$$\hat{\gamma}(\tilde{\mathbf{x}}) = \text{probit}^{-1}\left(-\frac{|\mathbb{E}_{\hat{p}(\tilde{\tau}|\tilde{\mathbf{x}}, D)}[\tilde{\tau}]|}{\text{Var}(\hat{p}(\tilde{\tau} \mid \tilde{\mathbf{x}}, D))^{\frac{1}{2}}}\right), \quad (2)$$

where probit^{-1} is the cumulative distribution function of normal distribution, and the expectations and variances are over the posterior predictive distribution of $\tilde{\tau} \mid \tilde{\mathbf{x}}$. (The absolute value in (2) makes it to apply also to negative expected treatment effects.) From (2) we see that the estimated Type S error rate will increase if the estimated treatment effect decreases, or if posterior uncertainty (variance) increases. Intuitively this makes sense.

4.3. Decision-Making Aware Active Learning

Our hypothesis is that active learning criteria that reduce the estimated Type S error rate will result in higher decision-making performance. We call active learning criteria that reduce the estimated Type S error *decision-making aware*.

A special example of decision-making aware criteria is *targeted expected information gain* introduced by Sundin et al. (2018). It selects the next query by maximizing the expected information gain of the posterior $\hat{p}(\tilde{y} \mid \tilde{x})$ using KL-divergence. This criterion is related to entropy minimization, because the expected KL-divergence between the current and updated posteriors can be written as $\mathbb{E}[D_{KL}(\hat{p}^*||\hat{p})] = \mathbb{E}[H(\hat{p}^*, \hat{p})] - \mathbb{E}[H(\hat{p}^*)]$, where $\hat{p}^*$ is the updated posterior, and $H(\hat{p}^*)$ its entropy, and $H(\hat{p}^*, \hat{p})$ is the cross entropy between the current and updated posteriors. Reducing the entropy in the posterior therefore reduces variance in eq. (2), which decreases the estimated Type S error rate.

We propose to directly minimize the estimated Type S error rate in eq. (2) with active learning. The criterion is to maximize the estimated reliability of a decision at $\tilde{x}$, that is, $1 - \hat{\gamma}(\tilde{x})$.

Directly minimizing the error can be interpreted as only exploiting what we already know, which lacks in exploration. There is an easy fix, though, which is to add exploration on the error. The exploration-exploitation trade-off is managed by maximizing the expected information gain on the

predictive distribution of Type S error: $\text{Bernoulli}(\hat{\gamma}(\tilde{x}))$ (technically, its relative entropy). The maximization of the information gain is equivalent to minimizing the posterior entropy and consequently the log-loss (Settles, 2012).

4.4. Counterfactual Elicitation

Assuming it is possible to acquire (noisy) observations about counterfactuals in the training data, we can do more, as discussed in the introduction. Denote by $\mathbb{D}$ the set of training examples $\{x_i, a_i, y_i\}_{i=1}^n$ for which the factual outcomes y_i have been observed, and denote by $\mathbb{U}$ the counterfactual examples $\{x_i, 1 - a_i\}_{i=1}^n$, for which the outcomes are unknown. Then we can use active learning to construct a set $\mathbb{L}$ that contains the new observations, and $\mathbb{L}$ will be data which would not normally be available. At each iteration, the algorithm selects $\{x^*, a^*\} \in \mathbb{U}$ to solicit a counterfactual outcome y^*. After this, $\{x^*, a^*, y^*\}$ is added to $\mathbb{L}$, and removed from $\mathbb{U}$. So the optimization problem at each query iteration k becomes

$$x^*, a^* = \underset{\{x,a\} \in \mathbb{U}}{\arg\min} \mathbb{E}_{\hat{p}(y|x,a,\mathbb{D},\mathbb{L})}[\hat{\gamma}_{k+1}(\tilde{x})], \text{ where}$$

$$\hat{\gamma}_{k+1}(\tilde{x}) = \hat{p}(y[a_{\tilde{x}}] < y[1 - a_{\tilde{x}}] \mid \tilde{x}, \mathbb{D}, \mathbb{L}, \{x, a, y\}),$$

and $a_{\tilde{x}}$ is the treatment with the highest expected outcome for $\tilde{x}$: $a_{\tilde{x}} = \arg\max_{a'} \mathbb{E}_{\hat{p}(y[a']|\tilde{x}, \mathbb{D}, \mathbb{L}, \{x,a,y\})}[y[a']]$.

Because the new observations are assumed to come from a different source than the original data, the model has separate noise parameters for the observation models of $\mathbb{D}$ and $\mathbb{L}$.

4.5. Comparative Observations about Counterfactuals

Another possible way in which noisy observations may be available is as comparisons of two counterfactual outcomes. Then, active learning is used to acquire a comparative observation $c \in {0, 1}$, which is a comparison between the expected counterfactual outcomes: $c = 1$ if $\mathbb{E}[Y[1] \mid X = x_i] > \mathbb{E}[Y[0] \mid X = x_i]$, else $c = 0$. At each iteration, the algorithm selects $\{x^*\} \in \mathbb{U}$ to solicit comparative observation c^*. After this, $\{x^*, c^*\}$ is added to $\mathbb{L}$, and removed from $\mathbb{U}$.

5. Experiments

We run three sets of experiments. First, we show that imbalance correlates with both the estimated and observed Type S error rate in simulated data. Second, we evaluate the performance of the proposed decision-making active learning criterion in simulated and semi-synthetic medical data. Last, we compare the performance to other active learning criteria that are applicable to the decision-making task, assuming a fixed query budget.
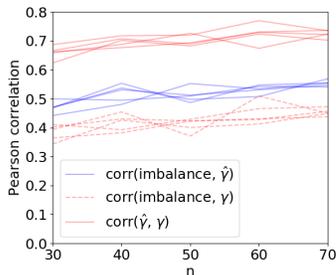
Figure 2. Imbalance correlates with the observed (γ) and estimated ($\hat{\gamma}$) Type S error rates on wide range of sample sizes n. Each line shows correlations between imbalance, estimated Type S error rate and observed Type S error rate in 1200 data sets. Separate lines show variation in 5 repeated experiments.

5.1. The Observed and Estimated Type S Error Rate in Imbalanced Data

In this section, we show empirically how imbalance affects the reliability of decision-making, and see that the estimated Type S error rate correlates with the observed error rate. The data are generated such that the outcome model contains interaction between a and x, and that the treatment effect is either saturating or increasingly increasing.

The outcome model generation is repeated 200 times, and for each outcome model we generate 6 training sets. The training data generation differs in the propensity scores, resulting in different levels of imbalance in the training data sets. Details of the data generation process are in the Supplementary. We measure imbalance using the Maximum Mean Discrepancy (MDD) (Gretton et al., 2012). We model the potential outcomes using two independent Gaussian Processes with squared exponential kernel.

Fig. 2 shows moderate correlation between imbalance and the Type S error rate, and that the correlation is quite constant across a wide range of sample sizes. This is empirical evidence of Theorem 1. Furthermore, the estimated Type S error rate $\hat{\gamma}$ and observed Type S error rate γ have strong correlation, around 0.7. Plotting $\hat{\gamma}$ against γ shows that low estimated error indicates low observed error (figure in the Supplementary). This suggests that the estimated Type S error rate is a good indicator of the prospective decision-making performance.

5.2. Decision-Making Aware Active Learning

We evaluate the performance of decision-making aware active learning (D-M Aware) using counterfactual elicitation in two cases. The first is synthetic data with high local imbalance and difficult outcome function shapes with interaction between a and x, similar to those in Fig. 1. The second is a semi-synthetic medical data set IHDP (Hill, 2011), commonly used in causal inference.

5.2.1. SIMULATED DATA WITH BINARY OUTCOMES

In this experiment, we study the proposed active learning approach in simulated data. Binary outcome y indicates the occurrence of an adverse effect, and the decision-making task is therefore to choose the treatment that results in a lower probability of the adverse effect. The setting is similar to that in the Fig. 1.

Synthetic data. The outcome $y \in \{0, 1\}$ is Bernoulli distributed with parameter $\theta_{x,a}$, given a one-dimensional covariate $x \in \mathbb{R}$ and treatment $a \in \{0, 1\}$. The data are generated from a logistic regression model with interaction between a and 3 radial basis functions (RBF) $\phi(x)$, s.t. $\theta_{x,a} = \text{logit}^{-1}(\mathbf{w}_0^\top \phi(x) + \mathbf{w}_1^\top \phi(x)a)$. Imbalance is induced to the training data by making the better treatment more likely for each x. Training sample size is 30. Details, such as the weights of the radial basis functions, are in the Supplementary.

Model and learning. We model the data with a logistic regression model $p(y \mid x, a) \sim \text{Bernoulli}(\theta_{x,a})$, where $\theta_{x,a}$ has the same form as in the data generation process. The model is fit using a probabilistic programming language Stan (Stan Development Team, 2017; Carpenter et al., 2017). We assume that the RBF centers and length-scale are known, so that only $\mathbf{w}_0$ and $\mathbf{w}_1$ need to be learned.

5.2.2. MEDICAL SEMI-SYNTHETIC DATA WITH CONTINUOUS OUTCOMES

We evaluate decision-making aware active learning approaches in deciding on medical interventions on real medical data with continuous-valued synthetic outcomes.

The IHDP data set. We use the Infant Health and Development Program (IHDP) dataset from Hill (2011), also used e.g. by Shalit et al. (2017) and Alaa & van der Schaar (2017), including synthetic outcomes, containing 747 observations of 25 features. Technical details are that we use the harder of the two cases in the paper, the "non-overlap case," and predict the non-linear outcome involving interactions ("Response Surface C"). These data come from a real randomized experiment, and imbalance has been produced by removing a part of the treated population. We evaluate the performance in leave-one-out cross-validation, but in order to make the problem even more realistically hard, for each of the 747 target units we choose randomly 100 observations as training examples.

Model and learning. We fit separate GPs to the outcomes of each treatment with GPy[1] (version 1.9.2), and use mixed noise likelihood to learn the noise in the obser-

---

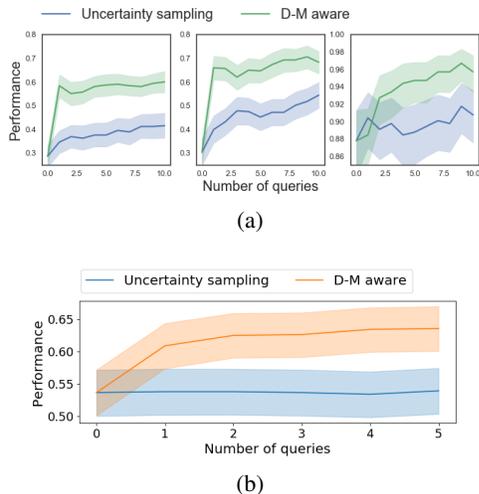[1]Toolbox available at: https://sheffieldml.github.io/GPy

(a)

(b)

Figure 3. Active learning that minimizes the estimated Type S error (D-M aware) increases the decision-making performance, measured in proportion of correct decisions. Shaded areas show the 95% bootstrap confidence interval. (a) Results of the simulated data with binary outcomes, in 100 repetitions averaged over 3 target units in each response type regions; the columns correspond to green (leftmost), blue (middle) and magenta (rightmost) regions in the Fig. 1. (b) Results in IHDP data as a function of the number of counterfactual queries to a simulated, noisy oracle, averaged over 747 decision-making tasks.

vations acquired by active learning. We use an exponentiated quadratic kernel with a separate length-scale parameter for each variable, and optimize the hyperparameters using marginal likelihood. Details of the priors are in the Supplementary. We use Gauss-Hermite quadrature of order 32 to approximate the expectations in D-M aware, Targeted-IG, and EIG.

5.2.3. RESULTS: EVALUATION OF DECISION-MAKING AWARE ACTIVE LEARNING

First, we evaluate the performance of the proposed active learning that minimizes the estimated Type S error rate (D-M aware) on the decision-making performance, as measured by the proportion of correct decisions. Fig. 3 shows that our method improves decision-making performance efficiently, compared to the baseline of uncertainty sampling. The difference to the baseline is statistically significant in Fig. 3(a) two out of three cases in the simulated data, and Fig. 3(b) in the semi-synthetic medical data (IHDP) (based on 95% bootstrap confidence intervals).

5.2.4. RESULTS: COMPARISON OF ACTIVE LEARNING CRITERIA IN DECISION-MAKING TASKS

Next, we compare the performance of D-M aware to two widely-used earlier active learning approaches, uncertainty

sampling and maximum expected information gain (Culotta & McCallum, 2005; Roy & McCallum, 2001) (EIG), and also include results for the previously introduced special case of decision-making aware active learning, targeted-IG (see Section 4.3). In order to make the methods comparable, we use the variant of D-M aware that also explores (see Section 4.3), as do the EIG and targeted-IG. The performance of the non-exploring variant of D-M aware is slightly lower (not shown) in all cases.

Simulated data. Fig. 4(a) shows that the decision-making aware active learning criteria, D-M aware and Targeted-IG, are the fastest to improve decision-making performance, and achieve significant increase in correct decisions after just one query in two cases out of three. After five solicited counterfactuals, Expected Information Gain (EIG) has reached comparable but still lower performance. Uncertainty sampling does not perform well in this example because it concentrates queries to the areas with estimated probability of adverse effect being close to 0.5, instead of close to the target units. Active learning reduces both imbalance and the estimated Type S error rate. Interestingly, D-M aware method achieves good decision-making performance regardless of having only little effect on imbalance, which may be due to local querying that reduces the local imbalance. In contrast, EIG reduces imbalance the most, which is expected as the criterion selects queries that are beneficial to the whole population.

IHDP data. The results in the IHDP data are similar as in the simulated data; The decision-making performance improves fastest with D-M aware and Targeted-IG, compared to EIG and uncertainty sampling (see Fig. 4(b)). D-M aware and Targeted-IG achieve statistically significant improvement in decision-making performance already with one query (based on 95% bootstrap confidence intervals).

5.2.5. COMPARATIVE FEEDBACK

Last, we demonstrate the use of comparative observations for counterfactual elicitation. The setting is the same as in Section 5.2.1, with the difference that the query is about which treatment has lower Bernoulli parameter. We fit the outcomes with GPs using Stan (Stan Development Team, 2017; Carpenter et al., 2017), which allows the model to learn both from direct and comparative observations. The results in Fig. 5 show that comparative observations increase decision-making performance efficiently in the simulated data setting. We note that the results with comparative feedback are better than those in Fig. 4(a), because here the queries give information about $\theta_{x,1} > \theta_{x,0}$, thus providing more information than the direct observations on outcome $y[a] \mid x \sim \text{Bernoulli}(\theta_{x,a})$.
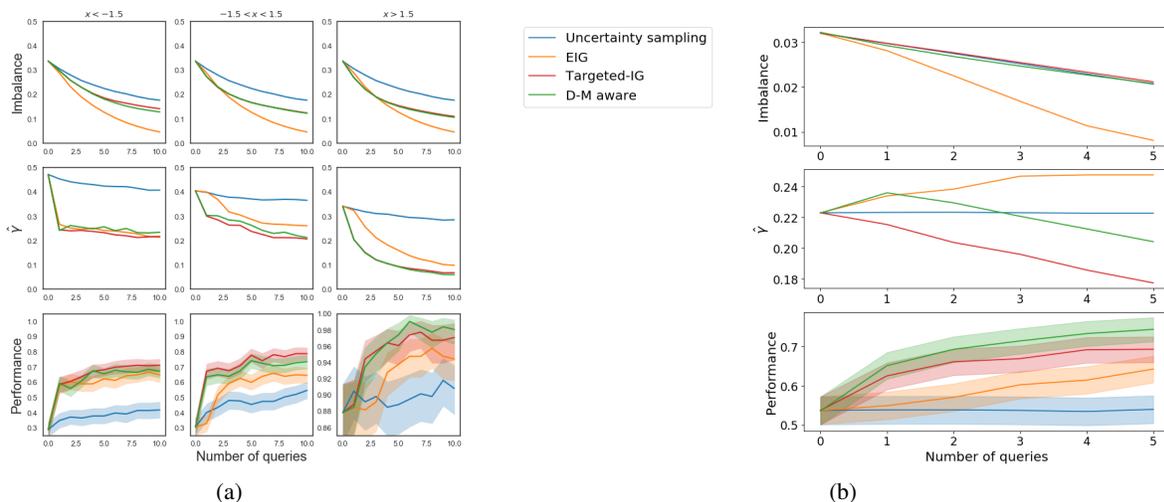
(a)   (b)

Figure 4. Comparison of active learning criteria as a function of number of queries. (a) Simulated data (b) IHDP data. The topmost panel in each shows imbalance, middle panel the estimated Type S error rate, and the lowest panel the proportion of correct decisions. Information-gain-based approaches are more effective than uncertainty sampling, and the decision making-aware criteria D-M aware and Targeted-IG are the best. Shaded areas show the 95% bootstrap confidence interval.
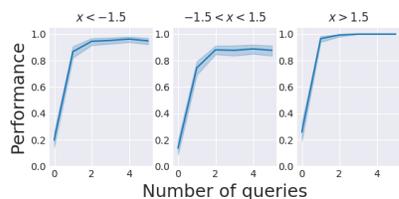

Figure 5. Comparative observations on the Bernoulli parameter in the simulated data is effective in increasing the proportion of correct decisions. The results are averaged over 3 target units in each response type regions and 100 repetitions.

6. Discussion and Conclusions

As machine learning systems are being integrated into human decision-making workflows, it is increasingly important that deployed models are correct and reliable. Predicting likely outcomes of various actions for decision support is especially challenging because the system's success is measured by its ability to correctly forecast the *effect of interventions*. This is more difficult than the classical sense of generalization in machine learning, where the goal is to have low risk under the distribution from which the training data was sampled, meaning that both distributions of x and $a \mid x$ stay the same (Pearl, 2009). In this paper we focused on the effect that imbalance in the training data can have on the reliability of comparisons of $\hat{p}(y[a] \mid \mathbf{x})$. We propose to improve the reliability by active learning that aims at maximizing the estimated reliability. In our experiments, this decision-making aware active learning outperforms standard methods in decreasing the error rate in decision-making.

The most computationally expensive step in the proposed approach is computing the expectation of the information gain, with complexity at worst N times that of re-training the model if no analytic solution is available. For example, a sparse GP model would have $O(N^2 M^2)$ complexity where M is the number of inducing points. The complexity of re-training GP models can be reduced using Kalman-filter based implementation with sequential updates for soliciting new observations. Furthermore, the full algorithm can be approximated by only including k nearest neighbours of the target unit $\tilde{x}$, using the model's intrinsic distance measure, when computing the expected utilities. Our preliminary results indicate, however, that the performance suffers if k is too low (see additional results in the Supplementary).

In this work, we assumed that the utility of an action a is directly the outcome $y[a]$. In case the utility is a function of the outcome, $U(y[a])$, our method applies by defining the Type S error rate as the probability of inferring erroneously which utility is the highest.

Future work includes soliciting observations from multiple sources, developing fast approximations that will allow computing the decision-making based utility efficiently for complex models, and studying human experts as one source of information.

One interesting application of the method is in personalized medicine, where counterfactual elicitation could solicit practitioners' knowledge to the model. For example, a clinician in a hospital has access to medical records of previous patients, and may also have personal experience about some of them. These data are rarely included in training sets of the models, but active learning and counterfactual elicitation could allow leveraging this additional source of information to infer more accurately about the future.

Acknowledgements

References

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task Gaussian Processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3424–3432. Curran Associates, Inc., 2017.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., and Simard, P. Featureinsight: Visual support for error-driven feature ideation in text classification. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pp. 105–112. IEEE, 2015.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. URL https://www.jstatsoft.org/v076/i01.

Culotta, A. and McCallum, A. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pp. 746–751, 2005.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2018.

Fu, Y., Zhu, X., and Li, B. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

Funk, M. J., Westreich, D., Wiesen, C., Strmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.

Gelman, A. and Tuerlinckx, F. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Hauser, A. and Bühlmann, P. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4): 926–939, 2014.

Hernán, M. and Robins, J. *Causal Inference*. Chapman & Hall, 2018.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Imbens, G. and Rubin, D. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Javdani, S., Chen, Y., Karbasi, A., Krause, A., Bagnell, D., and Srinivasa, S. S. Near optimal Bayesian active learning for decision making. In *AISTATS*, volume 2, pp. 17, 2014.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning (ICML)*, 2016.

Lunceford, J. K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

Mitchell, S., Potash, E., and Barocas, S. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Morgan, S. and Winship, C. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

Neyman, J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Neyman, J. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1990.

Pearl, J. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

Ribeiro, M., Singh, S., and Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144. ACM, 2016.

Rosenbaum, P. R. and Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79 (387):516–524, 1984.

Roy, N. and McCallum, A. Toward optimal active learning through Monte Carlo estimation of error reduction. In *International Conference on Machine Learning (ICML)*, pp. 441–448, 2001.

Rubin, D. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp. 34–58, 1978.

Saar-Tsechansky, M. and Provost, F. Decision-centric active learning of binary-outcome models. *Information Systems Research*, 18(1):4–22, 2007.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pp. 1697–1708, 2017.

Settles, B. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 2012.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Stan Development Team. Pystan: the python interface to stan, version 2.16.0.0, 2017. URL http://mc-stan.org.

Sundin, I., Peltola, T., Micallef, L., Afrabandpey, H., Soare, M., Mamun Majumder, M., Daee, P., He, C., Serim, B., Havulinna, A., Heckman, C., Jacucci, G., Marttinen, P., and Kaski, S. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34(13):i395–i403, 2018.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Xu, Y., Xu, Y., and Saria, S. A Bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference (MLHC)*, pp. 282–300, 2016.

Yan, S., Chaudhuri, K., and Javidi, T. Active learning with logged data. In *International Conference on Machine Learning*, pp. 5517–5526, 2018.