Recovering The Propensity Score From Biased Positive Unlabeled Data

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, Elke Rundensteiner
Worcester Polytechnic Institute, Worcester, MA
{wgerych, twhartvigsen, ljbuquicchio, emmanuel, rundenst}@wpi.edu

Abstract

Positive-Unlabeled (*PU*) learning methods train a classifier to distinguish between the positive and negative classes given only positive and unlabeled data. While traditional PU methods require the labeled positive samples to be an *unbiased* sample of the positive distribution, in practice the labeled sample is often a *biased* draw from the true distribution. Prior work shows that if we know the likelihood that each positive instance will be selected for labeling, referred to as the *propensity score*, then the biased sample can be used for PU learning. Unfortunately, no prior work has been proposed an inference strategy for which the propensity score is identifiable. In this work, we propose two sets of assumptions under which the propensity score can be uniquely determined: one in which no assumption is made on the functional form of the propensity score (requiring assumptions on the data distribution), and the second which loosens the data assumptions while assuming a functional form for the propensity score. We then propose inference strategies for each case. Our empirical study shows that our approach significantly outperforms the state-of-the-art propensity estimation methods on a rich variety of benchmark datasets.

Introduction

Typical binary classification assumes access to a dataset of labeled positive and negative examples during training. However, in practice fully-labeled training data are often unavailable. Instead, we may have only a small set of labeled examples any many *unlabeled* examples. As the negative class is often highly diverse, it tends to be prohibitively expensive, if not impossible, to obtain a sufficient labeled sample of the negative class, causing only *some* positive instances to be labeled (Bekker and Davis 2020; Hammoudeh and Lowd 2020). Such *Positive Unlabeled* (PU) data is characteristic of a wide variety of domains, such as healthcare (a lack of diagnosis does not mean someone does not have the disease) and images (most annotators are only asked to label objects that *are* in an image, not the infinite number of objects that are *not*).

PU learning assumes that a typically-unknown and complex labeling mechanism decides *which* positive instances are labeled. This mechanism is usually an imperfect human annotator with inherent and unobserved biases. The vast majority of recent methods for learning from PU data model this labeling mechanism by assuming that all positive instances are equally likely to be labeled (Bekker and Davis 2020). This is overly simplistic and disregards all *biases* in the labeling mechanism, which naturally lead to certain data points being labeled. For instance, this assumption ignores the fact that individuals with health insurance (who are more likely to visit doctors) are more likely to be be diagnosed than individuals without health insurance. The biased labeling mechanism can also be much more socially innocuous: objects in the foreground of an image are more likely to be labeled than objects in the background.

The key idea of our work is to recover the true complex and biased labeling mechanism by identifying the likelihood that a given positive instance is labeled. Recovering this labeling mechanism is essential for biased PU learning, as success would allow us to train a classifier that distinguish between the positive and negative classes given *only* biased positive and unlabeled data (Bekker, Robberechts, and Davis 2019). Additionally, learning the biased labeling mechanism allows us to recover the posterior of the positive class, which we can integrate over to obtain the class prior. This is important, as knowing the class prior allows us to compute performance metrics for standard positive-negative classification even when the test set includes only positive and unlabeled instances (Bekker and Davis 2020). Lastly, knowing the labeling mechanism gives insight into *why* certain instances were labeled while others were not, which can be important for explainability tasks.

While one prior work has likewise attempted to learn this labeling mechanism (Bekker, Robberechts, and Davis 2019), known as the *propensity score*, it does not yield an *identifiable* variable and so the learned likelihoods can be nearly arbitrarily incorrect. Our goal thus is to instead develop an approach that produces an identifiable propensity score.

This is a challenging task. As we do not directly observe the propensity score, it must be inferred indirectly from the data. Further, identifying the true propensity score is impossible without additional assumptions; when left unconstrained, an infinite number of propensity score/posterior class probability pairs can explain the observed PU data, as we show in the Preliminaries section.

It is thus crucial to identify scenarios in which the propen-

sity score *is* identifiable, due to its importance for biased PU learning. To yield identifiability, additional assumptions must be made on either the data distribution (specifically, the likelihoods of the positive and negative class) or on the propensity score itself. We thus propose two different estimation procedures: one that makes stronger assumptions on the positive and negative likelihoods but allows for a flexible propensity score, and another that makes stronger assumptions of the propensity score but allows for weaker likelihood assumptions.

We refer to our first case with rigid likelihoods and a flexible propensity score as the *Local Certainty Scenario*. Specifically, we assume that the relationship between the observed features and the true classes is a deterministic function, implying no class ambiguity when observing an instance which is appropriate for many real cases. For instance, an object either is or is not present in a given image, and this fact will not change if the same image is observed again. This matched a classic data assumption from *unbiased* PU learning (Du Plessis and Sugiyama 2014).

We refer to the flexible class likelihood with rigid propensity score as the *Probabilistic Gap Scenario*. In this case, we assume that the relationship between the observed features and the true class is *probabilistic*. To allow for this, we make the reasonable assumption that positive instances that resemble negative instances are less likely to be labeled. This handles the case where there is class overlap or ambiguity of class assignment for a given instance. For instance, a temperature check and symptom questionnaire can only yield the *probability* that someone has COVID-19, but cannot diagnose it with 100% certainty.

This work makes the following contributions:

- We propose the first methods that recover the *exact* propensity scores for PU data.
- We establish two scenarios under which the propensity score can be uniquely identified.
- For each scenario, we propose inference algorithms to identify the propensity score.
- Through a series of extensive experiments we show that our models outperform the state-of-the-art methods by estimating propensity scores more accurately and subsequently making more accurate classifications.

Related Work

Positive Unlabeled data has been researched for well over a decade (Liu et al. 2003). The overwhelming majority of PU works focus on the case where the positively labeled samples are an *unbiased* sample of the true positive distribution (Kiryo et al. 2017; Jain, White, and Radivojac 2017), and do not address the *biased* PU setting that is the focus of our work. These works range from performing classification (Elkan and Noto 2008; Guo et al. 2020) to recovering the positive class prior (Bekker and Davis 2018; Jain, White, and Radivojac 2016; Zeiberg, Jain, and Radivojac 2020).

Recent Positive Unlabeled work has begun to focus on the biased setting, where the labeled positives are a biased sample of the true positives (Jain et al. 2020; Bekker, Robberechts, and Davis 2019; Hammoudeh and Lowd 2020; Gerych et al. 2020). Many make the assumption that the probability of labeling a positive instances follows the order of the class probabilities (He et al. 2018; Youngs, Shasha, and Bonneau 2015; Kato, Teshima, and Honda 2019), which is similar to but slightly more general than the assumption made by our *Probabilistic Gap Scenario* method. However, unlike our work these methods can not and do not recover the labeling mechanism (propensity score), and generally focus on making accurate *binary* classification decisions. A few other works relax this label ordering assumption (Hammoudeh and Lowd 2020; Na et al. 2020), but likewise do not learn the labeling mechanism and thus do not address the task that is the focus of this work.

The work that is most closely related to ours is the *SAR-EM* method of (Bekker, Robberechts, and Davis 2019). This paper focuses on our goal of recovering the propensity score. *SAR-EM* employs an expectation-maximization algorithm to jointly find the true class posterior and propensity score by maximizing the probability of the observed data. However, as we show in the following section, there are an infinite number of possible propensity scores over a wide range of values that perfectly explain the observed data, but are far from the true propensity score.

The method proposed in (Jain et al. 2020) differs from the other biased PU methods, as it is designed for recovering the class prior from biased PU data rather than performing classification. This approach assumes that there exists clusters in the distribution of the positive instances, such that the labeling probability is constant *per cluster*. Although not discussed in (Jain et al. 2020), as this method finds the prior of each cluster, it can be used to recover this constant propensity score per cluster. This differs from our work by assuming that the propensity score is constant per cluster, and thus the propensity score can only take on a few discrete values over the whole data space (limited by the number of clusters).

Preliminaries

The goal of Positive-Unlabeled (PU) learning is to map features $x \in \mathcal{X}$ into classes $\mathcal{Y} = \{0, 1\}$ given only positive and unlabeled examples. If for a given x the corresponding $y \in \mathcal{Y}$ is 1 then the class is "positive", otherwise the class is "negative". We assume there is a joint distribution $p(x, y, \ell)$, such that $y \in \{0, 1\}$ is the class and $\ell \in \{0, 1\}$ is the *label indicator*. If $\ell = 1$, then the instance is labeled ($y = 1$). If $\ell = 0$, then the instance is unlabeled (and either $y = 1$ or $y = 0$). The class y is unobserved; thus, it is not straightforward to estimate the *class posterior* $p(y = 1|x)$, while the *label posterior* $p(\ell = 1|x)$ can be estimated by training a *non-traditional classifier* (Elkan and Noto 2008) to predict the probability of ℓ (which is observed) given x.

We assume the common *single training set scenario* (also known as the *censoring scenario*), in which a sample of data is collected from the joint distribution $p(x, y)$. When an instance is from the positive class, it is labeled with probability $p(\ell = 1|x, y = 1)$, and is unlabeled ($\ell = 0$) otherwise. The alternative PU assumption is the *case-control scenario*, in which the unlabeled data is drawn from the marginal $p(x)$ and another sample of labeled data is drawn

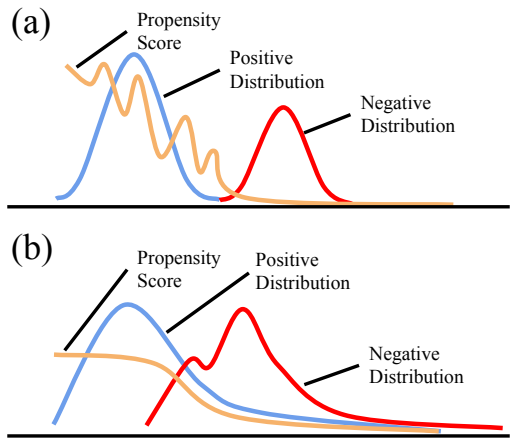

Figure 1: (a) Local Certainty Scenario. Non-overlapping class distributions and arbitrary propensity score. (b) Probabilistic Gap Scenario. Allows for non-overlapping classes, requires propensity score to follow ordering of $p(y = 1|x)$.

from $p(\ell = 1|x, y = 1)$. We describe our method in the single training set scenario as it more commonly used in PU learning (Bekker and Davis 2020). However, like most other PU techniques, it is straightforward to convert between the two.

To-date, the vast majority of the literature has focused on *unbiased* PU learning. This assumes that the labeled samples are an unbiased sample of the positive instances: $p(\ell = 1|x, y = 1) = p(\ell = 1|y = 1)$. However, we instead focus on the more challenging and realistic *biased* PU learning setting, where the likelihood that an instance is labeled depends on the features of that instance. Thus, generally $p(\ell = 1|x, y = 1) \neq p(\ell = 1|y = 1)$.

Recently, a method for directly modeling $p(\ell = 1|x, y = 1)$, known as the *propensity score* and referred to by the symbol e, has been proposed (Bekker, Robberechts, and Davis 2019). Theorem 1 shows the value of modeling the propensity score.

Theorem 1. *Let $\hat{y}$ be the predicted posterior probability of y. Then, $\mathbb{E}[R_{prop}(\hat{Y}|E, L)] = R(Y|X)$, where R is the standard empirical positive-negative risk of the predictions $\hat{y} \in \hat{Y}$ and R_{prop} is the propensity weighted risk defined as*

$$R_{prop}(\hat{Y}|E, L) = \frac{1}{n}\sum_{i=1}^{n} \ell_i \left(\frac{1}{e_i}\delta_1(\hat{y}_1) + (1 - \frac{1}{e_i})\delta_0(\hat{y}_i) \right) + (1 - \ell_i)\delta_0(\hat{y}_i), \tag{1}$$

such that $\delta_1(\hat{y}_i)$ is the cost of predicting $\hat{y}$ assuming that y_i is a true positive, and $\delta_0(\hat{y}_i)$ is the cost assuming y_i is a true negative, L is the set of class labels in the dataset and E is the set of corresponding propensity scores.

This theorem, proven by (Bekker, Robberechts, and Davis 2019), tells us that if the propensity score is known, we can train a classifier to predict the true class y using risk minimization, given only biased positive and unlabeled data. Moreover, this allows us to train a probabilistic classifier to model the class posterior $p(y|x)$, which is useful for uncertainty analysis and for obtaining an estimate for the class prior $p(y)$, which is required to calculate several standard PU classification evaluation metrics (Jain, White, and Radivojac 2017; Bekker and Davis 2020). Note that this is in contrast to most other biased PU methods, which attempt to only make accurate binary predictions for y, rather than modeling the class posterior (Kato, Teshima, and Honda 2019; Youngs, Shasha, and Bonneau 2015). Furthermore, the propensity score is a model of the complex labeling mechanism that decided *which* positive instances were labeled and thus provides information on *why* certain instances were selected to be labeled while others were not.

Existing approaches for calculating the propensity score do so by maximizing the probability of the observed data, treating both the class posterior and propensity score as latent variables. Unfortunately, this does not yield an identifiable propensity score; there are multiple incorrect hypothesis for the propensity score that are as equally likely to be returned by the method as is the true propensity score. In fact, the trivial solution of assuming all positive instances are labeled (treating all unlabeled instances as negative) with a corresponding propensity score of 1 for all positives instances is a valid solution.

The focus of this work is thus to develop approaches that yield an identifiable propensity score. This requires understanding when it is even possible for the propensity score to be identifiable. A natural starting place is to consider the four different standard data assumptions commonly in PU literature (Bekker and Davis 2020): Local Certainty/Separable Classes (Bayes Error of 0 between positive and negative distributions), Positive Subdomain (there is some region A of the feature space determined by partial attribute assignment such that the Bayes error is 0), Positive Function (there is some region A of the feature space determined by an arbitrary function for which the Bayes error is 0), Irreducibility (the negative distribution is not a mixture containing the positive distribution).

Theorem 2. *Let propensity score e be an arbitrary function of x, $e : \mathcal{X} \to (0, 1]$. Let the PU assumption hold (y is unobserved, ℓ and x are observed). Then, e is non-identifiable under the Positive Subdomain, Positive Function, and Irreducibly scenarios.*

Theorem 2, proven in the Appendix, shows that a general propensity score is not identifiable in any of the standard PU data assumptions other than the Probabilistic Gap scenario. Thus, we provide an identifiable propensity score estimation procedure in this setting in the following Local Certainty Propensity Estimation section.

Note that Theorem 2 only holds for a general propensity score. If we make stronger assumptions on the propensity score, we can make less restrictive data assumptions. Thus, we provide an estimation strategy for an identifiable propensity score in the Positive Function assumption in the Probabilistic Gap Propensity Estimation section, in which we assume a linear functional form for the propensity score.

An overview of the difference of assumptions of these two scenarios is shown in Figure 1.

Local Certainty Propensity Estimation

We first describe a method to recover the true propensity score in the Local Certainty scenario. In this setting, we assume the relationship between the observed features and the true class is a deterministic function $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y} = \{0, 1\}$ (where 1 is for the positive class and 0 is for the negative), while allowing the propensity score to be an arbitrary function $e : supp(p(x|y = 1)) \rightarrow \{q \in \mathbb{R} | 0 < q \leq 1\}$; i.e., an arbitrary function from the feature space of positive instances to a probability between 0 and 1. Note that the propensity score is only defined for regions of the feature space where x may be positive, as in PU learning negative instances are never labeled. The deterministic f is a valid assumption when the observed features are sufficient for uniquely determining the class of each instance. For instance, the features observed for an image (i.e. the pixels) are sufficient for determining which objects are in the image.

The first step to identifying the propensity score under local certainty is to express it in terms of the positive distribution and labeled distribution:

$$e = p(\ell = 1|x, y = 1) \tag{2}$$

$$= \frac{p(\ell = 1|x, y = 1)p(x, y = 1)}{p(x, y = 1)} \tag{3}$$

$$= \frac{p(\ell = 1, x, y = 1)}{p(x, y = 1)} \tag{4}$$

$$= \frac{p(\ell = 1, x)}{p(x, y = 1)} \tag{5}$$

$$= \frac{p(\ell = 1)p(x|\ell = 1)}{p(y = 1)p(x|y = 1)} \tag{6}$$

While we can take samples from $p(x|\ell = 1)$, we cannot approximate the above density ratio because we cannot sample from $p(x|y = 1)$ and no existing prior estimation method is applicable for estimating $p(y = 1)$ in our biased PU data[1]. We thus propose to replace the $p(x|y = 1)$ in the denominator of the last line with $p(x)/p(y = 1)$.

At a glance, this seems like an arbitrary and poor approximation of $p(x|y = 1)$. However, note that $p(x) = p(y = 1)p(x|y = 1) + (1 - p(y = 1))p(x|y = 0)$. Further, under the local certainty assumption there is no overlap between the positive and negative assumptions so for a positive instance x, $p(x|y = 0) = 0$ as there is no ambiguity of class belongingness for any observation. Thus, for positive instance x, $p(x) = p(y = 1)p(x|y = 1) + 0$ and so $p(x|y = 1) = p(y = 1)^{-1}p(x)$. This means that making this substitution in the denominator of Equation 6 *will produce the exactly correct propensity score for all positive instances*. Let e^* then be the value when we swap into the denominator of Equation 6:

$$e^* = \frac{p(\ell = 1)p(x|\ell = 1)}{p(x)}. \tag{7}$$

[1]More precisely, there is no previously proposed method for calculating this class prior from biased PU data *under the local certainty assumption*.

Note that the true propensity score, given in Equation 6, is undefined for all negative instances under the Local Certainty Scenario, as the denominator will equal 0 when x is outside of the support of the positive class (which all negative instances will be under the deterministic mapping assumption). However, e^* will be 0 for all negative instances, as $p(x|\ell = 1) = 0$ when x is outside of the support of the positive class by the Positive Unlabeled assumption (i.e., no negative instances are labeled). Additionally, note that e^* will *only* equal 0 for negative instances because e is strictly greater than 0 for all positive instances. Thus,

$$e = \begin{cases} e^* & e^* \neq 0 \\ Undefined & e^* = 0 \end{cases} \tag{8}$$

The candidate propensity score e^* can be estimated from the observed PU data because both $p(\ell = 1)$ and $p(x|\ell = 1)/p(x)$ can be calculated from the observed data. $p(\ell = 1)$ can be easily calculated by simply calculating the ratio of labeled vs unlabeled data. The ratio of $p(x|\ell = 1)$ to $p(x)$ can be estimated using *density ratio estimation* (Rhodes, Xu, and Gutmann 2020; Sugiyama, Suzuki, and Kanamori 2012; Sugiyama 2009) by taking samples from biased positives for $p(x|\ell = 1)$ and samples from the unlabeled instances for $p(x)$. Alternatively, this ratio can be calculated as the posterior probability of the label *indicator*. Thus, the output of a flexible probabilistic classifier trained to distinguish *labeled* vs. unlabeled, rather than positive vs. negative, can be used to obtain an estimate for e^*.

Equation 8 is the theoretically accurate value of the propensity score. However, in practice it is useful to have an estimate of e that is defined for all points when using e to train a classifier to determine the true class y. In this case, we propose directly using e^* as the estimated value of the propensity score e for all points. This substitution introduces no bias when used to train such a classifier:

Theorem 3. *Let $\hat{e}$ be an estimate for propensity score e. Then,*

$$bias(R_{prop}(\hat{Y}|E, L)) = \frac{1}{n}\sum_{i=1}^{n} y_i(1 - \frac{e_i}{\hat{e}_i})(\delta_1(\hat{y}) - \delta_0(\hat{y}))$$

This is shown in (Bekker, Robberechts, and Davis 2019) and the proof is repeated in our appendix for the sake of completeness. Theorem 3 implies that estimated propensity scores of negative instances do not introduce any bias when they are used to estimate $p(y = 1|x)$. Therefore defining the propensity score for negative instances to be 0 (or in fact any value) will not affect the downstream posterior estimates.

Thus, Equation 7 is a good estimate of the propensity score e in the case where the class of each instance is deterministic given the observed features. However, as discussed in our Introduction, for certain tasks the true class can not be determined with 100% certainty. In this scenario, the derivation that leads us to Equation 7 no longer applies, and so e^* may no longer be a good estimate of the propensity score. For this reason, we propose an additional estimation method for e that applies to the probabilistic problem setting, described in the following section.

Probabilistic Gap Propensity Estimation

We now describe how we recover the true propensity score in the probabilistic setting, which we refer to as the *Probabilistic Gap Scenario*. We assume that there is a probabilistic function $f : \mathcal{X} \rightarrow \mathcal{Y}$, such that $f(x) = 1$ with probability $g(x)$, where $g : \mathcal{X} \rightarrow \{q \in \mathbb{R} | \text{ s.t. } 0 < q < 1\}$. Additionally, we assume that The first assumption we now make is that there exists some region of the support of the positive class for which $p(y = 1|x) = 1$; i.e., there is some region without overlap, even if in general there is uncertainty for the class assignment. Informally, one can imagine this being the case for instances very far away from the majority of the negative instances. We note that this corresponds to the common Positive Function data assumption made in prior PU works (Bekker and Davis 2020) and described in our Preliminaries section. Unfortunately, as stated in Theorem 2, the propensity score is unidentifiable in this scenario when there are no additional assumptions made on the propensity score. Thus, we identify assumptions on the functional form of the propensity score in order to yield identifiability.

To this end, we base our propensity score assumption of of the common *invariance of order* assumption made by methods for biased PU learning (He et al. 2018; Youngs, Shasha, and Bonneau 2015; Kato, Teshima, and Honda 2019). This corresponds to $p(\ell = 1|x)$ following the order of $p(y = 1|x)$; i.e., if $p(y = 1|x_1) > p(y = 1|x_2)$ then $p(\ell = 1|x_1) = p(\ell = 1|x_2)$. However, this ordering assumption is slightly too weak to yield identifiability.

Proposition 1. *Let the Positive Function scenario and invariance of order assumption hold. Then, the propensity score is not identifiable*

Proof. We show that e is non-identifiable by constructing multiple valid propensity score and class posterior pairs (both of which are latent variables that determine each other, as $p(y = 1|x) \cdot e = p(\ell = 1|x)$). Let the class posterior hypothesis be $p(y = 1|x) = p(\ell = 1|x)^{1/N}$, and let the corresponding propensity score $p(\ell = 1|y = 1, x) = p(\ell = 1|x)^{(N-1)/N}$. In this case, for any positive integer N greater than 1, $p(\ell = 1|x)$ will follow the order of $p(y = 1|x)$, and the propensity score/class posterior pair are valid. Thus, the propensity score is not identifiable under the invariance of order assumption. □

We thus modify the invariance of order assumption slightly by assuming that the propensity score is a linear function of the class posterior; i.e., $e = k \cdot p(y = 1|x)$. Intuitively, this corresponds to assuming that positive instances that have less class ambiguity, or are more typical of the positive class, have a higher likelihood of being labeled. This is a reasonable assumption for many real-world applications. For instance, consider a human annotator who is tasked with labeling the objects in an image. Objects that are obscured, blurred, in the background, or otherwise difficult to identify will be less likely to be labeled than objects clearly in the foreground of the image.

Let x_i be an instance in the feature space where $p(y_i = 1|x_i) = 1$, as we assume exists according to Positive Function. In this case, the corresponding e_i would be equal to k, as $e_i = k \cdot p(y = 1|x_i) = k \cdot 1$. This implies that $p(\ell_i = 1|x_i)$ would likewise equal k, as $p(\ell_i = 1|x_i) = e_i \cdot p(y_i = 1|x_i) = k \cdot 1$.

Moreover, this implies that $p(\ell_i = 1|x_i) = Sup_{x^* \sim \mathcal{X}}[p(\ell = 1|x^*)]$. Therefore, we can obtain the value of k by modeling the posterior probability of the *label indicator* ℓ, and then finding the value that maximizes this probability. Thus, if $h_\ell(x)$ is a model of $p(\ell = 1|x)$,

$$k = Sup_{x^* \sim \mathcal{X}}[h_\ell(x)]. \tag{9}$$

Next, using Equation 9 we can e from h_ℓ:

$$\begin{aligned} e \cdot p(y = 1|x) &= h_\ell(x) \\ \frac{e^2}{k} &= h_\ell(x) \\ e &= \sqrt{k \cdot h_\ell(x)} \\ e &= \sqrt{Sup_{x^* \sim \mathcal{X}}[h_\ell(x)] \cdot h_\ell(x)} \end{aligned}$$

Therefore, we can obtain an estimate of the propensity score in this setting by first training a model of the posterior of the *label indicator*, such that the propensity score is the root of this posterior probability scaled by a constant that is also obtained from the label posterior.

Experiments

We study the effectiveness of our Local Certainty and Probabilistic Gap methods on several benchmark datasets. We use a Gaussian Process Classifier (Rasmussen 2003) to model the label indicator posterior necessary for each method. As implied, we calculate Equation 7 by using the posterior of the label indicator rather than more-complex density ratio estimates. Additional experiments are available in the supplementary materials.

Experimental Setup

Compared methods. We compare against a method that assumes constant labeling as a baseline. Specifically, we employ the state-of-the-art class prior estimation TiCE (Bekker and Davis 2018) to find an estimate of $p(y = 1)$, and then obtain an estimate of the propensity score by obtaining $p(\ell = 1|y = 1)$ from the estimated $p(y = 1)$. We call this baseline *Constant*, as it assumes that the propensity score is constant for all positive instances. We also compare against the two existing PU methods that estimate the propensity score: SAR EM (Bekker, Robberechts, and Davis 2019) and the method proposed in (Jain et al. 2020) which we refer to as *Cluster*. We use the publicly-available code for TiCE[2] and SAR EM[3] and implement *Cluster* ourselves.

Datasets. We use several standard benchmark datasets from the UCI Machine Learning Repository (Dua and Graff 2017): `Yeast` (Horton and Nakai 1996), `Bank` (Dua and Graff 2017), `Wine` (Aeberhard, Coomans, and De Vel 1994), `HTRU_2` (Lyon et al. 2016), `Occupancy` (Candanedo and Feldheim 2016), and `Adults` (Kohavi 1996).

[2] https://dtaid.cs.kuleuven.be/software/tice

[3] https://dtai.cs.kuleuven.be/drupal/software/sar

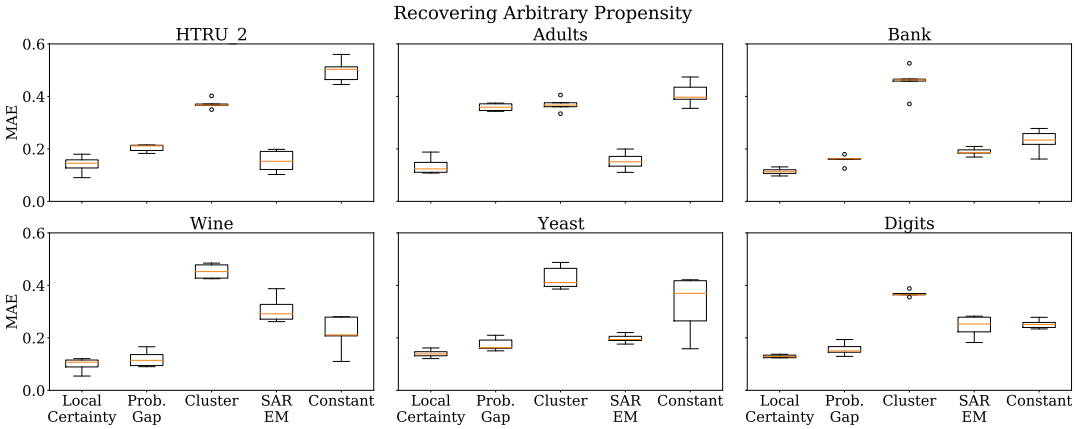

Figure 2: Results for recovering the propensity score for the arbitrary propensity setting. Our Local Certainty method significantly outperforms state-of-the-art on 4/6 datasets, and is never itself outperformed.

We likewise use two real-world datasets: `Yelp Reviews` (Sajnani et al. 2012) and `PASCAL VOC 2007` (Everingham et al. 2007).

These datasets were chosen to have a wide variety in dataset domains, size, and cardinality. Multi-class datasets were converted into binary classification problems when necessary as was done by (Jain et al. 2020). Random 70/30 train/test splits were used for each dataset. Each experiment was repeated 10 times in order to obtain confidence intervals. Additional dataset details are available in the supplemental materials.

Recovering the Propensity Score

We evaluate the ability of each method to recover the propensity scores in two settings: One in which the true propensity score is an arbitrary complex function and there is little to no overlap between classes (the Local Certainty Scenario) and the setting where the Probabilistic Gap assumptions are met (classes overlap and the propensity score is a scaled version of the class posterior). We refer to these as *Arbitrary Propensity* and *Scaled Propensity*, respectively.

Arbitrary Propensity. In this experiment, we decide which positives are labeled according to an arbitrarily complex propensity score. Specifically, we cluster the distances of the positive instances from the mean into 20 bins or clusters using k-means (such that the clustering is on the *distances*, not the positions of the points in the feature space). Each bin is assigned a random propensity score between 0.1 and 0.9. Ten trials are run per dataset and bins are randomly assigned for each. This creates a very complex propensity function to test the ability of the Local Certainty method to recover the propensity score without assuming a specific functional form. Additional details on the experimental setup are available in the supplemental materials. Results of these experiments on each dataset are shown in Figure 2, which reports the MAE between the estimated propensity scores and the true propensity score.

As Figure 2 shows, our Local Certainty method significantly outperforms all other methods for all but one dataset where our method ties SAR EM. As expected, the Probabilistic Gap method does not perform particularly well (usually in third place), as the assumptions of this method are not met in this problem setting. SAR EM is generally in second place, which is again expected as, unlike the other non-Local Certainty methods, it does not require a particular functional form for the propensity score. However, SAR EM is likely to converge to an incorrect estimate of the propensity score even when the classes are separable. Our method corrects for this case and so consistently outperforms SAR EM.

Note that these datasets were not modified to enforce the class separability assumption of the Local Certainty scenario. However, previous work has shown that classifiers have been able to achieve close to 100% accuracy on these datasets, indicating that there is already naturally little or no class overlap.

Scaled Propensity. We next evaluate the performance of each method when the Probabilistic Gap assumptions are met: the classes overlap and the propensity score is a constant multiple of the class posterior (such that an instance more typical of the positive class is more likely to be labeled). To this end, we introduce class overlap to these benchmark datasets. This is achieved by applying Borderline SMOTE (Han, Wang, and Mao 2005) to generate samples along the boundary of the positive and negative class, such that negative samples were generated on the positive side and vice versa. We apply this and ensure a roughly 30% class overlap for each dataset.

The ground truth propensity score in this setting is determined by first training a probabilistic classifier logistic regression model) to find the posterior of the positive class. Then, the propensity score was determined as the posterior model multiplied by a constant k, where k was randomly sampled from 0.3 to 0.8. k was re-sampled for each run, for ten runs per dataset.

Our findings are shown in Figure 3. Our Probabilistic Gap method significantly outperforms all other methods on each dataset, indicating that this approach does indeed more accurately recover the propensity score in this scenario.

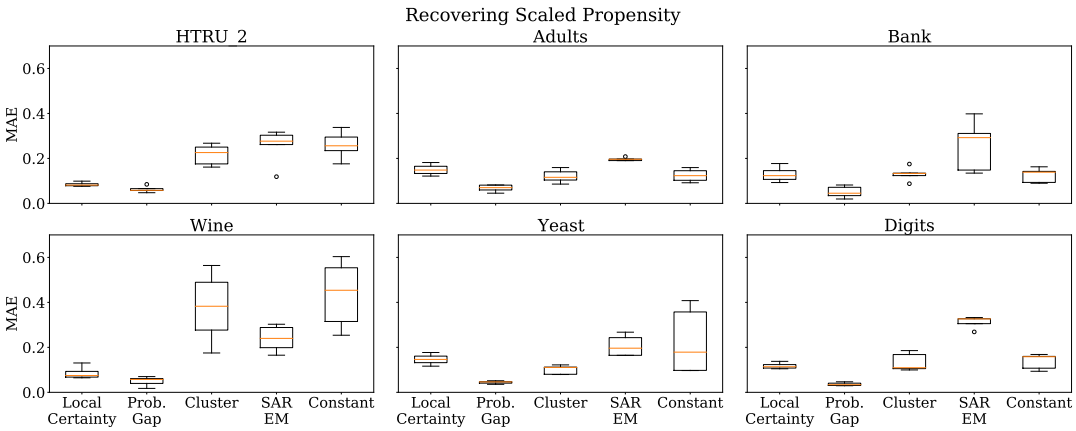

Figure 3: Results for recovering the propensity score for the scaled propensity setting. Our Probabilistic Gap method always significantly outperforms the state-of-the-art.

Dataset	Prop. Func.	LC (Ours)	PG (Ours)	Cluster	SE	Constant
PASCAL VOC	Arbitrary	**0.13**+/-0.03	0.31+/-0.04	0.45+/-0.04	0.20+/-0.04	0.30+/-0.13
PASCAL VOC	Scaled	0.10+/-0.05	**0.09**+/-0.05	0.84+/-0.08	0.12+/-0.08	0.74+/-0.08
Yelp	Arbitrary	**0.15**+/-0.02	0.18+/-0.02	0.29+/-0.04	0.21+/-0.07	0.25+/-0.03
Yelp	Scaled	0.08+/-0.04	**0.04**+/-0.02	0.28+/-0.03	0.09+/-0.03	0.12+/-0.06

Table 1: Mean absolute error between true and recovered propensity score for each method on two real-world datasets.

Interestingly, we observe that our Local Certainty method is 2nd best, meaning that the Local Certainty method is robust to its class separability assumption being broken.

Real-World Datasets. The previous experiments were conducted on benchmark datasets that were modified to meet our assumptions. To show the robustness of our approach and utility for real-world datasets, we perform experiments on two additional real-world datasets: PASCAL-VOC 2007 (featurized using a pre-trained Resnet-18 model (He et al. 2016)) and the Yelp Reviews dataset. The data was not modified to meet our data assumptions (i.e.,we did not ensure separable classes, or add new data using SMOTE). For each dataset, we performed experiments with either the arbitrary propensity score or the linear propensity score, as was done with the other datasets in our paper. Table 1 shows the results of these experiments. Our Local Certainty method still outperforms all other methods on both real-world datasets for the arbitrary propensity score experiment, and performs second only to our other method (Probabilistic Gap) for the linear propensity score (as expected). This shows that our methods still outperform the state-of-the-art even on complex, real-world datasets.

Utility of Recovered Propensity Scores: Using Propensity Sores For Classification

As discussed in the Preliminaries section, the propensity score can be used for classification. We thus illustrate the utility of our estimated propensity scores by comparing the class posteriors obtained from our estimated propensity scores to those obtained form the state-of-the-art propensity estimation methods. We achieve this by training a down-stream classifier for each dataset using the propensity-weighted risk (Equation 1). We utilize the propensity scores obtained in the the "Recovering the Propensity Score" experiments; thus, the dataset preparation and details for those experiments hold true for this experiment as well.

The results, shown in Table 2 and Table 3, demonstrate that our Local Certainty method produces a down-stream classifier with the lowest (best) error for the Arbitrary Propensity setting (Table 2), and our Probabilistic Gap method produces the best classifier in 5/6 of the datasets in the Scaled Propensity setting (Table 3). This shows that our methods nearly always result in the training of a more accurate classifier than the state-of-the-art propensity-recovering PU methods.

Additional Experiments

We include additional experiments in the Appendix. First, we show that the Probabilistic Gap method does perform well as long as the propensity score follows the *order* of the class posterior, even if the true propensity is not exactly a constant multiple of the class posterior (thus, the method still produces good results even when its somewhat strong assumption is broken). Additionally, we show how the performance of our methods is degraded when a poor estimate of the label indicator posterior is used, showing the need to pick a robust posterior model.

Conclusion

This work proposes a significant step forward for PU learning in the biased setting, which is more realistic than much

Dataset	HTRU 2	Adult	Bank	Wine	Yeast	Digits
LC (Ours)	**0.04**+/-0.00	**0.35**+/-0.02	**0.06**+/-0.01	**0.24**+/-0.02	**0.44**+/-0.00	**0.21**+/-0.01
PG (Ours)	0.10+/-0.00	0.40+/-0.00	0.22+/-0.01	0.36+/-0.02	0.47+/-0.00	0.33+/-0.01
Cluster	0.05+/-0.00	0.37+/-0.00	0.09+/-0.00	0.48+/-0.01	0.46+/-0.00	0.23+/-0.00
SE	0.10+/-0.05	0.37+/-0.01	0.09+/-0.01	0.47+/-0.05	0.46+/-0.01	0.33+/-0.03
Constant	0.43+/-0.00	0.70+/-0.01	0.17+/-0.01	0.34+/-0.02	0.46+/-0.00	0.29+/-0.01

Table 2: Classification error for arbitrary propensity score scenario

Dataset	HTRU 2	Adult	Bank	Wine	Yeast	Digits
LC (Ours)	0.14+/-0.01	0.75+/-0.01	0.38+/-0.03	0.26+/-0.01	0.45+/-0.01	0.51+/-0.02
PG (Ours)	0.05+/-0.00	**0.25**+/-0.03	**0.30**+/-0.01	**0.25**+/-0.02	**0.41**+/-0.01	**0.27**+/-0.01
Cluster	**0.04**+/-0.00	0.27+/-0.00	0.33+/-0.00	0.78+/-0.01	0.45+/-0.00	0.51+/-0.01
SE	0.14+/-0.08	0.26+/-0.01	0.35+/-0.01	0.51+/-0.06	0.44+/-0.01	0.49+/-0.03
Constant	0.43+/-0.00	0.46+/-0.03	0.39+/-0.01	0.34+/-0.03	0.46+/-0.00	0.54+/-0.01

Table 3: Classification error for scaled propensity score scenario

of the prior work. Specifically, we establish two cases where the propensity score can be uniquely identified, and propose estimation strategies for recovering this propensity score. This is important, as knowing the propensity score allows us to learn the true class posterior and to calculate important quantities such as the class prior. We show that our approach significantly outperforms existing methods, even in cases where our model's assumptions are not completely met. This work also serves as a jumping off point for future research directions for biased PU learning. In particular, we foresee estimation procedures for a propensity score in the probabilistic setting that do not follow the rigid probabilistic gap assumption.

Acknowledgements. This work was funded by the DARPA WASH program HR001117S0032. Results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI.

References

Aeberhard, S.; Coomans, D.; and De Vel, O. 1994. Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition*, 27(8): 1065–1077.

Bekker, J.; and Davis, J. 2018. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bekker, J.; and Davis, J. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4): 719–760.

Bekker, J.; Robberechts, P.; and Davis, J. 2019. Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*, volume 11907 of *Lecture Notes in Computer Science*, 71–85. Springer.

Candanedo, L. M.; and Feldheim, V. 2016. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. *Energy and Buildings*, 112: 28–39.

Du Plessis, M. C.; and Sugiyama, M. 2014. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5): 1358–1362.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213–220.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results http://www. pascal-network. org/challenges. In *VOC/voc2007/workshop/index. html*.

Gerych, W.; Buquicchio, L.; Chandrasekaran, K.; Alajaji, A.; Mansoor, H.; Murphy, A.; Rundensteiner, E. A.; and Agu, E. O. 2020. BurstPU: Classification of Weakly Labeled Datasets with Sequential Bias. *2020 IEEE International Conference on Big Data (Big Data)*, 147–154.

Guo, T.; Xu, C.; Huang, J.; Wang, Y.; Shi, B.; Xu, C.; and Tao, D. 2020. On positive-unlabeled classification in GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8385–8393.

Hammoudeh, Z.; and Lowd, D. 2020. Learning from Positive and Unlabeled Data with Arbitrary Positive Shift. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13088–13099. Curran Associates, Inc.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887. Springer.

He, F.; Liu, T.; Webb, G. I.; and Tao, D. 2018. Instance-Dependent PU Learning by Bayesian Optimal Relabeling. *CoRR*, abs/1808.02180.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Horton, P.; and Nakai, K. 1996. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, 109–115.

Jain, S.; Delano, J.; Sharma, H.; and Radivojac, P. 2020. Class Prior Estimation with Biased Positives and Unlabeled Examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4255–4263.

Jain, S.; White, M.; and Radivojac, P. 2016. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2685–2693.

Jain, S.; White, M.; and Radivojac, P. 2017. Recovering True Classifier Performance in Positive-Unlabeled Learning. In Singh, S. P.; and Markovitch, S., eds., *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2066–2072. AAAI Press.

Kato, M.; Teshima, T.; and Honda, J. 2019. Learning from Positive and Unlabeled Data with a Selection Bias. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Kiryo, R.; Niu, G.; du Plessis, M. C.; and Sugiyama, M. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances In Neural Information Processing Systems (NeurIPS)*, 1675–1685.

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, 202–207.

Liu, B.; Dai, Y.; Li, X.; Lee, W. S.; and Yu, P. S. 2003. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, 179–186. IEEE.

Lyon, R. J.; Stappers, B.; Cooper, S.; Brooke, J. M.; and Knowles, J. D. 2016. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1): 1104–1123.

Na, B.; Kim, H.; Song, K.; Joo, W.; Kim, Y.; and Moon, I. 2020. Deep Generative Positive-Unlabeled Learning under Selection Bias. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, 1155–1164. ACM.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

Rhodes, B.; Xu, K.; and Gutmann, M. U. 2020. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*.

Sajnani, H.; Saini, V.; Kumar, K.; Gabrielova, E.; Choudary, P.; and Lopes, C. 2012. Classifying yelp reviews into relevant categories.

Sugiyama, M. 2009. Density ratio estimation: A new versatile tool for machine learning. In *Asian Conference on Machine Learning*, 6–9. Springer.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.

Youngs, N.; Shasha, D. E.; and Bonneau, R. 2015. Positive-Unlabeled Learning in the Face of Labeling Bias. In *IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14-17, 2015*, 639–645. IEEE Computer Society.

Zeiberg, D.; Jain, S.; and Radivojac, P. 2020. Fast Nonparametric Estimation of Class Proportions in the Positive-Unlabeled Classification Setting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 6729–6736. AAAI Press.