Class Prior Estimation with Biased Positives and Unlabeled Examples

Shantanu Jain, Justin D. Delano, Himanshu Sharma, Predrag Radivojac
Khoury College of Computer Sciences
Northeastern University, Boston, MA, U.S.A.

Abstract

Positive-unlabeled learning is often studied under the assumption that the labeled positive sample is drawn randomly from the true distribution of positives. In many application domains, however, certain regions in the support of the positive class-conditional distribution are over-represented while others are under-represented in the positive sample. Although this introduces problems in all aspects of positive-unlabeled learning, we begin to address this challenge by focusing on the estimation of class priors, quantities central to the estimation of posterior probabilities and the recovery of true classification performance. We start by making a set of assumptions to model the sampling bias. We then extend the identifiability theory of class priors from the unbiased to the biased setting. Finally, we derive an algorithm for estimating the class priors that relies on clustering to decompose the original problem into subproblems of unbiased positive-unlabeled learning. Our empirical investigation suggests feasibility of the correction strategy and overall good performance.

Introduction

Learning from positive and unlabeled data refers to a binary classification problem in which all labeled examples are positive and unlabeled examples contain a mix of positive and negative examples (Denis 1998; Denis, Gilleron, and Letouzey 2005; du Plessis, Niu, and Sugiyama 2014; Hsieh, Natarajan, and Dhillon 2015; Chang et al. 2016). Over the past two decades, there has been growing interest in this form of semi-supervised learning with a broad range of applications in text mining, biology, and social networks, to name a few (Liu et al. 2003; Ward et al. 2009; Tran 2013). Positive-unlabeled learning has also been studied and well-understood in numerous problems including matrix completion (Hsieh, Natarajan, and Dhillon 2015), hypothesis testing (Geurts 2011), approximation of posterior distributions (Jain et al. 2016) and performance evaluation (Menon et al. 2015; Jain, White, and Radivojac 2017; Ramola, Jain, and Radivojac 2019).

A standard and major assumption in the positive-unlabeled setting is that one is presented with unbiased i.i.d. samples for both positives and unlabeled data. This assumption, however, may not hold in practice and may have major consequences in all aspects of learning and predictor deployment. For example, during the process of scientific discovery, examples selected for labeling may reflect social biases, the priorities of funding agencies, limitations of instrumentation, or individual preferences of researchers. In molecular biology, one can more easily crystallize small proteins than large proteins which biases the Protein Data Bank resource to be more representative of smaller structured molecules, because highly flexible proteins are difficult to crystallize (Dunker et al. 2001). Similarly, one can more easily collect protein-protein interaction (PPI) data for proteins that are expressed in the yeast nucleus via yeast two-hybrid experiments (Fields and Song 1989). This leads to biases in interaction networks with respect to positives that in this case reflect two over-represented groups of protein pairs: (1) those containing small structured proteins cocrystallized with their partners and (2) those containing proteins easily expressed in a yeast's nucleus. As link prediction is a perennial task in PPI networks, estimating the size of the PPI network based on incomplete data as well as estimating missing interactions will likely propagate initial biases unless corrective measures are applied. In fact, the strategies for mitigating the effects of sample selection bias have long been studied in machine learning (Heckman 1979; Zadrozny 2004; Huang et al. 2006; Cortes et al. 2008; Hsieh, Niu, and Sugiyama 2019); however, very few authors have considered a positive-unlabeled setting (Youngs, Shasha, and Bonneau 2015).

The focus of this work is on nonparametric estimation of class priors; i.e., the fractions of positive and negative examples in unlabeled data, given a sample of positives and a sample of unlabeled examples. This problem has been considered in the past decade with well-developed identifiability theory (Blanchard, Lee, and Scott 2010; Jain et al. 2016) and several well-performing estimation algorithms (Elkan and Noto 2008; Sanderson and Scott 2014; Jain et al. 2016; Ramaswamy, Scott, and Tewari 2016; du Plessis, Niu, and Sugiyama 2017; Bekker and Davis 2018). However, both theory and algorithms considered only the case when both positives and unlabeled data are i.i.d. samples from the underlying data distributions. We intend to relax this assumption and investigate the case where only a biased set of positive examples is available, whereas the unlabeled data remains unbiased.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Most semi-supervised learning techniques rely on unbiased unlabeled data for bias correction strategies, typically via sampling or re-weighting, but consider situations where at least one of the samples from class-conditional distributions is unbiased (Zadrozny 2004; Huang et al. 2006; Cortes et al. 2008; Hsieh, Niu, and Sugiyama 2019). In class-prior estimation, several authors have considered the case of biased class proportions in the labeled data (Latinne, Saerens, and Decaestecker 2001; Vucetic and Obradovic 2001) and developed iterative correction methods. The expectation-maximization (EM) approach by Latinne, Saerens, and Decaestecker (2001) can also be reformulated as minimization of Kullback-Leibler divergence between labeled and unlabeled distributions resulting in optimization of a convex objective (du Plessis and Sugiyama 2012). This formulation further allows distribution matching to be generalized to other distance functions (du Plessis and Sugiyama 2012). All of these techniques consider labeled data that contains both positives and negatives.

In this work, we focus on a more restrictive case of positive-unlabeled learning and begin to study class prior estimation in the presence of a biased positive sample. We introduce three important assumptions, called the "mixing bias" assumption, "ϕ-irreducibility" assumption and "disjoint kernel support" assumption, that allow us to model the bias, develop identifiability theory and propose theoretically grounded bias-correction strategies for estimating class priors. We then conduct experiments and provide evidence that our algorithms give considerable improvement in practice over those that assume unbiased positives, even when the assumptions are not fully satisfied.

The remainder of this work is structured as follows. The Background section introduces notation and gives a brief summary of the theory and algorithms for the unbiased case of positives. The Theoretical Framework section gives rigorous identifiability results and the Estimation Algorithm derives an estimation procedure for bias correction. Experiments and Results summarize our empirical investigation, summarizing the datasets, experimental protocols and results. Finally, the discussion and concluding remarks are provided in Conclusions and Future Work.

Background

We consider a binary classification problem of mapping an input space $\mathcal{X}$ to an output space $\mathcal{Y} = \{0, 1\}$ given a set of positive examples and a set of unlabeled examples. Let f, f_1 and f_0 capture the true distributions of the inputs, inputs from the positive class and the inputs from the negative class, respectively. Let α be the proportion of positives in the true distribution of inputs f. Using machine learning terminology, the mixing proportion α is the class prior for the positive class, f_1 and f_0 are the class-conditional distributions for the positive and negative classes, respectively, and f is the distribution from which unlabeled data is sampled. It follows that for any $x \in \mathcal{X}$, f can be expressed as the following two-component mixture

$$f(x) = \alpha f_1(x) + (1 - \alpha) f_0(x). \tag{1}$$

Generally, class prior (α) estimation is an ill-posed problem due to unidentifiability, in the sense that there are multiple values of α that lead to the same f for a given f_1, obtained by altering the value of f_0. Furthermore, the set of all valid α values is an interval of the form $[0, \alpha^*]$, where $\alpha^* = a_f^{f_1}$;

$$a_f^{f_1} = \sup\{a \in [0, 1] : f = a f_1 + (1 - a) h_0, h_0 \in \mathcal{P}_\mathcal{X}\} \tag{2}$$

and $\mathcal{P}_\mathcal{X}$ is the set of all densities (except for f_1) defined on $\mathcal{X}$. The "irreducibility" assumption that f_0 itself cannot be expressed as a nontrivial mixture containing f_1 as one of the component makes the problem identifiable with α taking its largest value (Blanchard, Lee, and Scott 2010; Jain et al. 2016). More formally,

$$a_{f_0}^{f_1} = 0 \Rightarrow \alpha = \alpha^*. \tag{3}$$

Estimation Procedures

Over the past decade, several algorithms have been proposed for class prior estimation in the case of a representative positive sample (Elkan and Noto 2008; Jain et al. 2016; Ramaswamy, Scott, and Tewari 2016; du Plessis, Niu, and Sugiyama 2017; Bekker and Davis 2018). We briefly describe the AlphaMax algorithm (Jain et al. 2016) that we later use as a base algorithm in the experimental section. AlphaMax is a nonparametric class prior estimation algorithm that first maximizes the log-likelihood of the positive unlabeled data at multiple values of $\alpha \in (0, 1)$ and then estimates α^* as an x-coordinate of the elbow of the maximum log-likelihood versus α curve. Though, in principle, AlphaMax approach can incorporate multidimensional data, it can be computationally prohibitive and lead to suboptimal performance when directly run on high-dimensional data. Fortunately, there exist α^*-preserving univariate transforms that can be used to reduce the data to a single dimension while preserving α^* in the transformed space (Jain et al. 2016). Formally, for an α^*-preserving transform, $\tau : \mathcal{X} \to \mathbb{R}$, it holds that

$$a_{f_\tau}^{f_{\tau,1}} = a_f^{f_1}, \tag{4}$$

where f_τ and $f_{\tau,1}$ are density functions on $\mathbb{R}$ that are obtained as counterparts of f and f_1 after transforming the inputs using τ. In practice, AlphaMax is run on the univariate data transformed by the score function of a so-called nontraditional classifier—a classifier trained to discriminate positive examples against the unlabeled examples treated as negatives (Elkan and Noto 2008). This nontraditional classifier serves as an α^*-preserving transform.

Theoretical Framework

All approaches in positive-unlabeled learning assume that an unbiased sample from f_1 is accessible, however, in the setting for this paper, the set of positively labeled examples is biased. We illustrate this situation in Figure 1.

Assumptions for Identifiability

Let the positively labeled examples be an i.i.d. sample from a biased density, f_1'. We formulate our problem under a

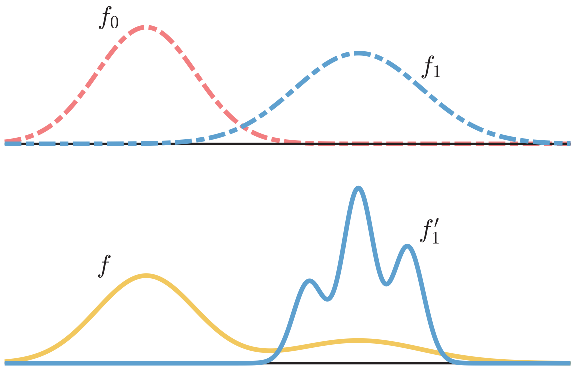

Figure 1: Illustration of the positive-unlabeled problem with bias. The upper panel shows positive (f_1; blue) and negative (f_0; red) class-conditional distributions shown with dashed lines because the samples from these distributions are not available. The lower panel shows the unbiased mixture distribution ($f = \alpha f_1 + (1-\alpha) f_0$, where $\alpha = 0.25$; yellow) and the biased class-conditional distribution of positives (f_1'; blue). The distributions drawn with solid lines illustrate that samples from these distributions are available. The objective of this work is to recover α given a biased positive sample and an unbiased unlabeled sample.

"**mixing bias**" assumption relating f_1' and f_1 via a K-component mixture representations as follows.

$$f_1'(x) = \sum_{i=1}^{K} r_i \phi_i(x), \quad f_1(x) = \sum_{i=1}^{K} \gamma_i \phi_i(x), \tag{5}$$

where $\phi_i(x)$ are density functions, $r_i, \gamma_i \in [0,1]$, $\sum_{i=1}^{K} r_i = 1$ and $\sum_{i=1}^{K} \gamma_i = 1$. The assumption enforces that f_1' and f_1 are mixtures sharing the same components, but having different mixing proportions.

For identifiability, we make a "**ϕ-irreducibility**" assumption that f_0 cannot be expressed as a nontrivial mixture containing ϕ_i; i.e,

$$a_{f_0}^{\phi_i} = 0, \quad \text{for } i = 1, 2, \ldots K. \tag{6}$$

For identifiability, we further make a "**disjoint kernel support**" assumption that ϕ_i's have disjoint supports; i.e.,

$$\text{Supp}(\phi_i) \cap \text{Supp}(\phi_j) = \emptyset, \quad \text{for } i \neq j. \tag{7}$$

Identifiability

In this section we derive that the above assumptions make the mixing proportion estimation a well-posed problem. Our main result in this section is Theorem 1 which shows that if f_0 is irreducible with respect to ϕ_i for $i = 1, 2, \ldots, K$; i.e., $a_{f_0}^{\phi_i} = 0$, then α^* can be uniquely identified.

To derive the results, we define a family of two-component mixtures as follows.

$$\mathcal{F}(\mathcal{P}_0, \mathcal{P}_1) = \{\alpha f_1 + (1-\alpha) f_0\}_{(\alpha, f_1, f_0) \in (0,1) \times \mathcal{P}_1 \times \mathcal{P}_0}, \tag{8}$$

where $\mathcal{P}_1$ and $\mathcal{P}_0$ are families of distributions on $\mathcal{X}$ from where f_1 and f_0 can be picked, respectively.

Let Φ be a family of K unique distributions on $\mathcal{X}$; i.e., $\Phi = \{\phi_i\}_{i=1}^{K}$. Let $\mathcal{M}_\Phi$ be a family of mixtures constructed with components from Φ. In other words, $\mathcal{M}_\Phi$ is the convex hull of Φ; i.e., for $\boldsymbol{a} = [a_i]_{i=1}^{K}$,

$$\mathcal{M}_\Phi = \left\{ \sum_{i=1}^{K} a_i \phi_i \right\}_{\boldsymbol{a} \in \Delta^{K-1}},$$

where Δ^{K-1} is a unit $K-1$ simplex. Thus, $a_i \in [0,1]$ and $\sum_{i=1}^{K} a_i = 1$.

We restate the mixing bias assumption in terms of $\mathcal{M}_\Phi$ as follows

$$f_1' \in \mathcal{M}_\Phi \quad \text{and} \quad \mathcal{P}_1 = \mathcal{M}_\Phi.$$

For the purpose of identifiability theory, we assume that Φ is known. The assumption relies on a presupposition that Φ can be obtained from the decomposition of f_1'. Note that we are not implying that there is a unique decomposition of f_1' giving Φ, but that there is some decomposition which is expressive enough to reconstruct f_1 (see Equation 5).

Next, we consider two choices for $\mathcal{P}_0$

1. $\mathcal{P}_\Phi^{\text{all}} = \mathcal{P}_\mathcal{X} \backslash \mathcal{M}_\Phi$, the set of all possible distributions on $\mathcal{X}$, except those in $\mathcal{M}_\Phi$. Removing $\mathcal{M}_\Phi$ from $\mathcal{P}_\mathcal{X}$, ensures that a mixture in $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$ must have a component outside Φ with nonzero proportion; i.e., $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi) \cap \mathcal{M}_\Phi = \emptyset$.
2. $\mathcal{I}_\Phi = \mathcal{P}_\Phi^{\text{all}} \setminus \mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$. As shown in Lemma 1 (Appendix), $f_0 \in \mathcal{I}_\Phi$ if and only if the ϕ-irreducibility assumption in Equation 6 is satisfied.

$\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$ is the largest conceivable set of nontrivial mixtures under the mixing bias assumption. Since $\mathcal{I}_\Phi \subseteq \mathcal{P}_\Phi^{\text{all}}$, all mixtures in $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ are also present in $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$; i.e., $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi) \subseteq \mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$. However, Theorem 1 shows that the two families contain exactly the same distribution under the disjoint kernel support assumption on Φ. In other words, we get the same set of mixtures even after reducing the parameter space of f_0 from $\mathcal{P}_0$ to $\mathcal{I}_\Phi$. This is useful because, unlike $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$, $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ is identifiable, as shown by the theorem.

Here the identifiability of $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$ and $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ is implicilty defined in terms of parameters (α, f_1, f_0) contained in the parameter space $(0,1) \times \mathcal{M}_\Phi \times \mathcal{P}_\Phi^{\text{all}}$ and $(0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$, respectively. $\mathcal{A}(f, \mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$ is defined as the set of all valid α values that allow f to be expressed as a nontrivial two component mixture with components from $\mathcal{P}_\Phi^{\text{all}}$ and $\mathcal{M}_\Phi$; i.e., $\mathcal{A}(f, \mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi) = \{\alpha \in (0,1) : f = \alpha f_1 + (1-\alpha) f_0, f_1 \in \mathcal{M}_\Phi, f_0 \in \mathcal{P}_\Phi^{\text{all}}\}$

Theorem 1. *Given a Φ that satisfies the disjoint kernel support (Equation 7),*

1. As sets, $\mathcal{F}(\mathcal{P}_\Phi^{all}, \mathcal{M}_\Phi)$ and $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ are equal.[1]

[1] $\mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$ and $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ are not equal as families as they differ in terms of their underlying parameter space.

2. *Every $f \in \mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ has a unique parametrization $(\alpha^*, f_1^*, f_0^*) \in (0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$, i.e., $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ is identifiable.*
3. *For an $f \in \mathcal{F}\big(\mathcal{P}_\Phi^{all}, \mathcal{M}_\Phi\big)$ having parametrization $(\alpha^*, f_1^*, f_0^*) \in (0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$, $\mathcal{A}\big(f, \mathcal{P}_\Phi^{all}, \mathcal{M}_\Phi\big) = (0, \alpha^*]$.*
4. *$\mathcal{F}\big(\mathcal{P}_\Phi^{all}, \mathcal{M}_\Phi\big)$ is unidentifiable.*

Proof.

- **Statement 1** By definition, $f = \alpha f_1 + (1-\alpha) f_0$ for some $\alpha \in (0,1)$, $f_1 \in \mathcal{M}_\Phi$ and $f_0 \in \mathcal{P}_\Phi^{\text{all}}$. Using Lemma 3 (Appendix), f_0 can be expressed as $a h_1 + (1-a) f_0^*$, where $a \in [0,1)$, $h_1 \in \mathcal{M}_\Phi$ and $f_0^* \in \mathcal{I}_\Phi$. Thus, for $\alpha^* = \alpha + (1-\alpha)a$, $f_1^* = \frac{\alpha f_1 + (1-\alpha) a h_1}{\alpha^*}$, $f = \alpha^* f_1^* + (1-\alpha^*) f_0^*$ and consequently $(\alpha^*, f_1^*, f_0^*) \in (0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$ parameterizes f. Thus, $\mathcal{F}\big(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big) \subseteq \mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$, however, $\mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi) \subseteq \mathcal{F}\big(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big)$ is trivially true. Thus, $\mathcal{F}\big(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big) = \mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$
- **Statement 2** Suppose an $f \in \mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ has two distinct parametrizations (α, f_1, f_0), (α^*, f_1^*, f_0^*) in $(0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$. It follows that $\alpha^* f_1^* + (1-\alpha^*) f_0^* = \alpha f_1 + (1-\alpha) f_0$. Since $f_1, f_1^* \in \mathcal{M}_\Phi$, for some $\boldsymbol{\gamma}, \boldsymbol{\gamma^*} \in \Delta^{K-1}$, $f_1 = \sum_{i=1}^K \gamma_i \phi_i$ and $f_1^* = \sum_{i=1}^K \gamma_i^* \phi_i$. The equation can be expressed as

  $$\left(\sum_{i=1}^K (\alpha^* \gamma_i^* - \alpha \gamma_i)_+ \phi_i\right) + (1-\alpha^*) f_0^* \\ = \left(\sum_{i=1}^K (\alpha \gamma_i - \alpha^* \gamma_i^*)_+ \phi_i\right) + (1-\alpha) f_0,$$

  where $a_+ = \max(0, a)$. Denoting $\eta_i = \alpha^* \gamma_i^* - \alpha \gamma_i$, if for some $i \in \{1, \ldots, K\}$, $\eta_i \neq 0$, then by restricting the formula above on Supp(ϕ_i) and dividing by $\phi_i(x)$, we get

  $$|\eta_i| = -\text{sign}(\eta_i)\left((1-\alpha^*)\frac{f_0^*(x)}{\phi_i(x)} - (1-\alpha)\frac{f_0(x)}{\phi_i(x)}\right).$$

  If $\eta_i > 0$, we pick a sequence in Supp(ϕ_i) along which the limit of $\frac{f_0(x)}{\phi_i(x)}$ goes to 0—such a sequence exists due to Lemma 1 and 2 (Appendix). The limit of the RHS along this sequence can not be positive, where as the LHS is a positive number, hence a contradiction. Similarly if $\eta_i < 0$, we pick the sequence such that $\frac{f_0^*(x)}{\phi_i(x)}$ goes to 0 and arrive at the same contradiction. Thus, $\eta_i = 0$, $\forall i \in \{1, \ldots, K\}$ and $(1-\alpha^*) f_0^* = (1-\alpha) f_0$. Taking an integral over $\mathcal{X}$ on both sides implies $\alpha^* = \alpha$ and consequently $f_0^* = f_0$. $\alpha^* = \alpha$ and $f_0^* = f_0$ together imply $f_1^* = f_1$. Thus, the two parameterizations of f are not distinct, hence a contradiction.
- **Statement 3**: For $\alpha \in (0, \alpha^*]$,

  $$f = \alpha^* f_1^*(x) + (1-\alpha^*) f_0^*(x) \\ = \alpha f_1^*(x) + (1-\alpha)\left(\frac{\alpha^* - \alpha}{1-\alpha} f_1^*(x) + \frac{1-\alpha^*}{1-\alpha} f_0^*(x)\right)$$

  Since $\alpha^* < 1$, $\left(\frac{\alpha^*-\alpha}{1-\alpha} f_1^*(x) + \frac{1-\alpha^*}{1-\alpha} f_0^*(x)\right) \notin \mathcal{M}_\Phi$ (from Lemma 4 in the Appendix) and consequently, it is an element of $\mathcal{P}_\Phi^{\text{all}}$. Thus, $(0, \alpha^*] \subseteq \mathcal{A}\big(f, \mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big)$. Suppose for some $\alpha \in (\alpha^*, 1]$, $f_1 \in \mathcal{M}_\Phi$ and $f_0 \in \mathcal{P}_\Phi^{\text{all}}$, $f = \alpha f_1 + (1-\alpha) f_0$. It follows that $\alpha f_1(x) + (1-\alpha) f_0(x) = \alpha^* f_1^*(x) + (1-\alpha^*) f_0^*(x)$. Rearranging terms, adding and subtracting $\alpha^* f_1(x)$, and using the representation of f_1 and f_1^* as elements of $\mathcal{M}_\Phi$,

  $$(\alpha - \alpha^*) f_1(x) + \alpha^* \left(\sum_{i=1}^K (\gamma_i - \gamma_i^*) \phi_i(x)\right) \\ = (1-\alpha^*) f_0^*(x) - (1-\alpha) f_0(x),$$

  where $\boldsymbol{\gamma}, \boldsymbol{\gamma^*} \in \Delta^{K-1}$. Note that there exists $j \in \{1, \ldots, K\}$ such that $\gamma_j \geq \gamma_j^*$ because otherwise $\sum_{i=1}^K \gamma_i < 1$. Restricting the above equation to the support of ϕ_j,

  $$(\text{positive constant}) \times \phi_j(x) = (1-\alpha^*) f_0^*(x) - (1-\alpha) f_0(x).$$

  Dividing both sides by $\phi_j(x)$ and taking limit over a sequence for which $\frac{f_0^*(x)}{\phi_j(x)}$ goes to 0 (such a sequence exists by Lemma 1 and 2 in the Appendix), the RHS is a positive number whereas the LHS is not positive. Hence a contradiction. Thus, an $\alpha > \alpha^*$ cannot be in $\mathcal{A}\big(f, \mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big)$ and $\mathcal{A}\big(f, \mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big) = (0, \alpha^*]$.
- **Statement 4**: The statement follows from Statement 3.

□

Estimation Algorithm

Our approach to estimating the class prior with biased positive-unlabeled data is to decompose the problem into many unbiased positive-unlabeled subproblems by first partitioning the original dataset, then applying AlphaMax to all data subsets in the partition and, lastly, combining the estimates from the subproblem to obtain the final estimates. A pseudo code of the method is given in Algorithm 1.

To motivate the above approach, we derive Theorem 2 which shows that α^* can be expressed in terms of λ_i^*'s, the class priors corresponding to the subproblems. The aim of partitioning the dataset is to obtain a partition of the support of f, $\mathcal{B} = \{B_i\}_{i=1}^K$, that satisfies Supp$(\phi_i) \subseteq B_i$. Note that the disjoint kernel support assumption on Φ is necessary for the existence of such a partition. Under the disjoint kernel support assumption and the mixing bias assumption, the restriction of both f_1 and f_1' on B_i after normalization as a density is ϕ_i; i.e.,

$$\phi_i = \frac{f_1(x)\mathbb{1}_{B_i}(x)}{\int_{B_i} f_1(x)dx} = \frac{f_1'(x)\mathbb{1}_{B_i}(x)}{\int_{B_i} f_1'(x)dx}.$$

Thus, the labeled examples in the i^{th} data subset of the partition is distributed as per ϕ_i. Denoting the restriction of f on the B_i after normalization as ψ_i, the unlabeled data in the i^{th} data subset is distributed as per ψ_i. Now, by representing ψ_i as a two component mixture containing ϕ_i, as one of the components, one can identify $\lambda_i^* = a_{\psi_i}^{\phi_i}$, the maximum value of the proportion ascribed to ϕ_i in all such representations. Alternatively, it is the class prior when the input space is restricted to B_i. However, unlike the unrestricted input space, the distribution of labeled positives restricted to B_i, ϕ_i, is unbiased because the mixing bias assumption only affects the proportion of ϕ_i in f_1'. One can therefore estimate λ_i^* using existing algorithms such as AlphaMax, Elkan-Noto or KM for class prior estimation in the standard positive-unlabeled setting (Jain et al. 2016; Elkan and Noto 2008; Ramaswamy, Scott, and Tewari 2016).

The estimate of λ_i^* can be combined with an estimate of p_i given by the proportion of unlabeled examples lying in the i^{th} subset of the partition, $\frac{\sum_{x \in \mathbb{M}} \mathbb{1}_{B_i}(x)}{|\mathbb{M}|}$, to obtain an estimate of α^* (Equation 9). λ_i^* is also important for inferring the unbiased positive and negative distributions. f_1 can be expressed as a convex combination of ϕ_i's with $\gamma_i = \frac{\lambda_i^* p_i}{\alpha^*}$ as the weights (Equation 11); ϕ_i can be estimated as the density of the positively labeled data in the i^{th} data subset. f_0 can be expressed as a convex combination of $\xi_i = \frac{\psi_i - \lambda_i^* \phi_i}{1 - \lambda_i^*}$'s with $\frac{(1-\lambda_i^*)p_i}{1-\alpha^*}$ as weights (Equation 11), where ψ_i is the restriction of f on B_i after normalization and can be estimated as the density of the unlabeled data in the i^{th} data subset. Though unbiased estimation of f_1^* and f_0^* is possible, we only focus on estimation of α^* in this paper.

Theorem 2. *Let $f \in \mathcal{F}(\mathcal{I}_\Phi, \mathcal{M}_\Phi)$ and $f_1' \in \mathcal{M}_\Phi$. Let (α^*, f_1^*, f_0^*) be the unique parametrization of f in $(0,1) \times \mathcal{M}_\Phi \times \mathcal{I}_\Phi$ —as per Statement 2 in Theorem 1. Consider a size K partition of* Supp(f)*, $\mathcal{B} = \{B_i\}_{i=1}^K$, such that* Supp$(\phi_i) \subseteq B_i$*. Let ψ_i be the density function of the restriction of f on B_i; i.e., $\psi_i(x) = \frac{f(x)\mathbb{1}_{B_i}(x)}{p_i}$, where $\mathbb{1}_B$ is the indicator function of set B and $p_i = \int_{B_i} f(x)dx$. Let $\lambda_i^* = a_{\psi_i}^{\phi_i}$ and $\xi_i = \frac{\psi_i - \lambda_i^* \phi_i}{1-\lambda_i^*}$. It follows that*

$$\alpha^* = \sum_{i=1}^{K} \lambda_i^* p_i \tag{9}$$

$$f_1^* = \frac{1}{\alpha^*} \sum_{i=1}^{K} \lambda_i^* p_i \phi_i. \tag{10}$$

$$f_0^* = \frac{1}{1-\alpha^*} \sum_{i=1}^{K} (1-\lambda_i^*) p_i \xi_i. \tag{11}$$

Proof. $\forall x \in B_i, p_i \psi_i(x) = f(x)$. It follows that

$$p_i(\lambda_i^* \phi_i(x) + (1-\lambda_i^*)\xi_i(x)) = \alpha^* f_1^*(x) + (1-\alpha^*) f_0^*(x)$$

Since $f_1^* \in \mathcal{M}_\Phi$, $f_1^* = \sum_{i=1}^K \gamma_i \phi_i$, for some $\boldsymbol{\gamma} \in \Delta^{K-1}$. Due to the disjoint kernel support assumption on Φ and Supp$(\phi_i) \subseteq B_i$, $f_1^*(x) = \gamma_i \phi_i(x)$ when $x \in B_i$. Thus, after rearranging terms and dividing by $\phi_i(x)$, the equation above can be expressed as

$$p_i \lambda_i^* - \alpha^* \gamma_i = (1-\alpha^*) \frac{f_0^*(x)}{\phi_i(x)} - p_i(1-\lambda_i^*) \frac{\xi_i(x)}{\phi_i(x)}.$$

Taking the limit on both sides along a sequence in Supp(ϕ_i) such that $\frac{f_0^*(x)}{\phi_i(x)}$ goes to 0 (such a sequence exists due to Lemma 1 and 2 in the Appendix), $p_i \lambda_i^* - \alpha^* \gamma_i \leq 0$. Similarly, taking the limit along a sequence in Supp(ϕ_i) such that $\frac{\xi_i(x)}{\phi_i(x)}$ goes to 0 (such a sequence exists due to Lemma 2 in the Appendix), $p_i \lambda_i^* - \alpha^* \gamma_i \geq 0$. Thus

$$p_i \lambda_i^* = \alpha^* \gamma_i \tag{12}$$

It follows that $\sum_{i=1}^K p_i \lambda_i^* = \alpha^* \sum_{i=1}^K \gamma_i$ and consequently, $\alpha^* = \sum_{i=1}^K p_i \lambda_i^*$ because $\boldsymbol{\gamma} \in \Delta^{K-1}$. From Equation 12, $\gamma_i = \frac{p_i \lambda_i^*}{\alpha^*}$. Thus Equation 10 is true. Now, using $f = \sum_{i=1}^K p_i \psi_i$ and the expression for f_1^* in $f_0^* = \frac{f - \alpha^* f_1^*}{1-\alpha^*}$, Equation 11 can be derived. □

Algorithm 1 Algorithm for class prior estimation with biased positives and unlabeled examples.

```
    // maxK specifies the maximum number of clusters.
Require: $\mathbb{M}, \mathbb{C}$, maxK
Ensure: $\alpha^*$
    // Partition the biased positive set by k-means clustering.
    // The number of clusters is picked to be the one giving
    // a clustering with the maximum Silhouette coefficient,
    // up to a maximum of maxK. cPart[i] stores the
    // positives in the $i^{\text{th}}$ cluster.
    cPart ← kMeansSilhouette($\mathbb{C}$, maxK)
    // Obtain cluster centers. centers[i] contains the
    // center of the $i^{\text{th}}$ cluster.
    centers ← mean(cPart)
    // Assign the unlabeled examples to their closest cluster,
    // using distance to the cluster centroid to measure
    // closeness. mPart[i] stores the unlabeled examples
    // in the $i^{\text{th}}$ cluster.
    mPart ← clusterAssignment($\mathbb{M}$, centers)
    for $i = 1, \ldots,$ length(cPart) do
        // Estimate the class prior on the $i^{\text{th}}$ cluster data.
        $\lambda[i]$ ← AlphaMax(mPart[$i$], cPart[$i$])
        // Compute the proportion of unlabeled examples in
        // the $i^{\text{th}}$ cluster.
        p[i] ← $\frac{\text{size(mPart}[i])}{\text{size}(\mathbb{M})}$
    end for
    // Estimate $\alpha^*$ using Equation 9.
    $\alpha^*$ ← dotProduct($\lambda$, p)
```

Partitioning the Data

We start the partitioning by using k-means algorithm to cluster the positively labeled set, $\mathbb{C}$, with the number of clusters chosen to maximize the Silhouette coefficient (de Amorim and Hennig 2015). The unlabeled sample, $\mathbb{M}$, is subsequently partitioned by assigning examples to clusters based on their distance to already computed cluster centroids. In an ideal scenario, the clustering will lead to a partition of the dataset that is consistent with the partition induced by the support of densities in Φ; i.e., for all $i \in \{1, \ldots, K\}$, all points in the dataset that lie in the support of ϕ_i should be in the same data subset of the partition.

In general, however, the partitioning approach is not guaranteed to produce an ideal partition. Fortunately, even if the partition is not ideal, α^* can be estimated reasonably well, as shown in the next section. The efficacy of the estimation depends on the difference between f_1 and its best approximation obtained by taking a convex combination of the densities corresponding to the labeled examples in the data subsets generated from the partitioning.

Experiments and Results

Datasets

Our experiments were carried out on twelve real-life datasets from the UCI Machine Learning Repository (Lichman 2013). When necessary, the original data was converted to binary classification problems on real-valued inputs as follows. Categorical features were encoded by one-hot encoding. Regression problems were converted into classification problems based on the mean of the target variable; i.e., examples with target values higher than the mean were labeled as positive and the remaining examples were labeled as negative. Multi-class classification problems were converted to binary by combining original classes.

To generate biased positive examples and unlabeled data, the positive examples were clustered using k-means, where the number of clusters, K, was determined based on the Silhouette coefficient. For high-dimensional data (dimension greater than 50), we reduced the dimension to 50 by taking Gaussian Random Projections (Dasgupta 2000) before applying k-means clustering. Next, for the i^{th} cluster, a random number ζ_i was picked from Uniform$(0, 1)$. Then, $\lceil \zeta_i \times \text{cluster size} \rceil$ number of positives picked randomly from the cluster were added to the biased positives set, $\mathbb{C}$. Once sampling was completed for all clusters, $\mathbb{C}$ contained the final set of biased positives. The remaining positives were kept in a pool from which the positives for the unlabeled data were drawn randomly. The distribution of the positives in the pool can interpreted as the unbiased distribution of positives. Next, a random number was picked as the class prior, α from Uniform$(0, 1)$. The size of the unlabeled sample was picked as $|\mathbb{M}| = \min(10000, \lfloor {}^{\#\text{pool}}/_{\alpha} \rfloor, \lfloor {}^{\#\text{neg}}/_{(1-\alpha)} \rfloor)$, where #neg and #pool were the number of negative examples in the dataset and the pool size, respectively. This allowed us to pick the largest size of the unlabeled sample while ensuring that enough positives and negatives were available to attain α as the class prior. Then, $\lfloor \alpha |\mathbb{M}| \rfloor$ positives and $\lfloor (1-\alpha) |\mathbb{M}| \rfloor$ negatives were picked randomly from the pool and the negative sample, respectively, and added to the unlabeled set, $\mathbb{M}$. Repeating this process, we generated 50 biased positive-unlabeled datasets per original dataset.

We also generated unbiased positive-unlabeled datasets from each of the twelve real-life datasets to assess the effect of bias correction when, in reality, the positive sample was unbiased. To this end, we first picked a random number, ζ, from Uniform$(0, 1)$ that determined the size of the unbiased positive set. $\lceil \zeta \times \#\text{pos} \rceil$ examples were randomly picked from the set of available positives into the unbiased positive set, where #pos was the total number of positives. The remaining positives were kept in a pool. Once the pool was constructed, the value of α, the size of the unlabeled set, and the examples sampled in the unlabeled set were obtained in the same manner as for the biased case. Fifty unbiased positive-unlabeled datasets were generated from each of the original datasets.

Algorithms

We ran two versions of our algorithm on the biased positive-unlabeled datasets, namely, Corrected and Corrected*, as well as standard class prior estimation for unbiased data that we refer to as Uncorrected. Corrected is an exact implementation of Algorithm 1 with `maxK` intialized to 5. Corrected* implements Algorithm 1 without clustering the data. Instead, the clustering of the positives used to generate the data along with cluster assignments for the unlabeled examples is given to it as input. This allowed us to separately study the effects of suboptimal clustering. Note that for both Corrected and Corrected*, we called AlphaMax on each cluster separately to estimate λ_i^*, which means that a different transform was obtained by training a nontraditional classifier on each cluster. We also ran AlphaMax directly on the entire dataset, as a baseline (Uncorrected), to understand the effect of bias on class prior estimation if no correction was applied. Finally, we also ran Corrected, Corrected* and Uncorrected with the Elkan-Noto algorithm (Elkan and Noto 2008) as another base algorithm, instead of AlphaMax.

Results

The experimental results are summarized in Figure 2 and Figure 3. For most biased datasets, both the Corrected* and Corrected algorithms based on AlphaMax show smaller estimation error compared to the Uncorrected algorithm (Figure 2). However, when the algorithms are based on Elkan-Noto, Corrected* and Corrected do not always lead to better estimates. This can be explained since Elkan-Noto makes the assumption of complete separation between the positive and negative class-conditional distributions, which is rarely the case in real-life datasets.

The algorithms were also ranked for each dataset based on their absolute error. The algorithm with the smallest error was ranked number 1, whereas the one with the largest error was ranked number 3. The average rank of an algorithm was computed for each original dataset by taking the mean of its ranks from all 50 biased datasets generated from the original dataset. The left panel of Figure 3 shows each algorithm's boxplot, constructed with twelve average ranks, one per original dataset, for the AlphaMax base algorithm. Based on the average ranks, the three AlphaMax-based algorithms can be ranked in decreasing order of overall performance as (1) Corrected*, (2) Corrected, and (3) Uncorrected. Their average mean absolute errors (MAEs) across the twelve datasets were 0.0900, 0.1579, and 0.2277, respectively. The superiority of Corrected* over Corrected is expected because it used the same clustering that was used to generate the data. However, Corrected is the more practical algorithm among the two because the clustering that generates the biased data is rarely known.

The mean absolute error (MAE) and ranking were also computed for AlphaMax-based Corrected and Uncorrected procedures on the bias-free datasets. The right panel of Figure 3 shows the boxplots of average

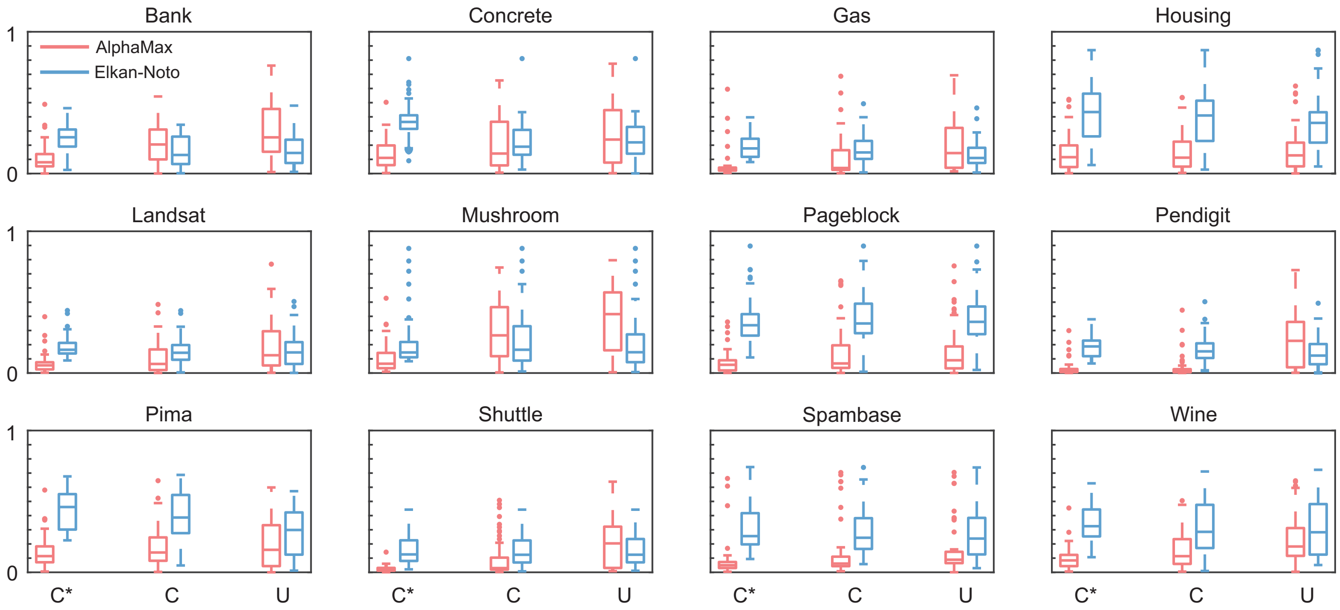

Figure 2: Boxplots of absolute error in class prior estimation for three algorithms (C* = Corrected*, C = Corrected, and U = Uncorrected) on twelve datasets. Each procedure was ran with AlphaMax and Elkan-Noto as the base estimator. Fifty biased positive-unlabeled datasets were generated from each of the twelve original datasets for evaluating the three algorithms.

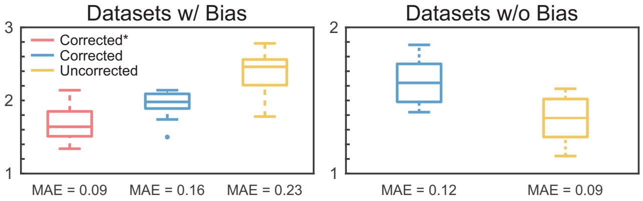

Figure 3: (Left) Boxplots of average rank for AlphaMax-based Corrected*, Corrected, and Uncorrected algorithms on datasets with bias. (Right) Boxplots of average rank for AlphaMax-based Corrected and Uncorrected algorithms on datasets without bias. Both panels also show the average MAE for each algorithm over all datasets.

ranks for Corrected and Uncorrected algorithms on datasets without bias. The algorithm with the smaller error ranked number 1, whereas the one with the larger error ranked number 2. The twelve average ranks per boxplot were computed in the same way as above. Although on the unbiased datasets the average rankings of the Corrected algorithm was lower than the Uncorrected algorithm (Figure 3, right), the comparison based on mean absolute error (MAE) shows that the Corrected algorithm is only slightly inferior. The average MAEs across the twelve datasets for the Corrected and Uncorrected algorithms were 0.1176 and 0.0871, respectively.

Conclusions and Future Work

As demonstrated by our experiments, the existence of bias in the labeled data can adversely affect the estimation of class prior, the central quantity for many tasks in positive-unlabeled learning. The effect of bias on class prior estimation can be corrected under the mixing bias, disjoint kernel support and ϕ-irreducibility assumptions both in theory and in practice as shown by Theorems 1 and 2 and our experimental results. The efficacy of the correction depends on the data satisfying the above assumptions and also on the ability of the algorithm to find a good partitioning. Though the current partitioning strategy always satisfies the disjoint kernel support assumption, it is not explicitly designed to satisfy the mixing bias and ϕ-irreducibility assumptions. This can be an interesting future research direction. Our experimental results also show that in the absence of bias the estimates from the corrected algorithm and the uncorrected algorithm are comparable. Theorem 2 also shows the possibility of inferring the unbiased distribution of positives and negatives. Developing practical algorithms to do so would be critical to other tasks in positive-unlabeled learning such as performance estimation and estimation of posterior probabilities.

Appendix

Lemma 1. $f_0 \in \mathcal{I}_\Phi$ *if and only if* $\forall \phi_i \in \Phi, a_{f_0}^{\phi_i} = 0$.

Proof. (⇒) From Lemma 4 in (Jain et al. 2016), the set of valid mixing proportions in the expression of f_0 as a mixture containing ϕ_i, denoted by $\mathcal{A}\big(f_0, \phi_i, \mathcal{P}^{all}\big)$, is closed from above. Thus, $a_{f_0}^{\phi_i}$ is a valid mixing proportion. If $a_{f_0}^{\phi_i} > 0$, there exists $h_0 \in \mathcal{P}_{\mathcal{X}} \setminus \{\phi_i\}$ such that $f_0 = a_{f_0}^{\phi_i}\phi_i + \left(1 - a_{f_0}^{\phi_i}\right)h_0$. However, this implies that $f_0 \in \mathcal{F}(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi)$, which contradicts $f_0 \in \mathcal{I}_\Phi$. Thus, $a_{f_0}^{\phi_i} = 0$.

(⇐) Suppose the consequent is true. If $f_0 \in \mathcal{F}\big(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\big)$, then for some $a \in (0,1)$, $\boldsymbol{w} \in \Delta^{K-1}$ (the unit $K-1$ simplex), $h_1 = \sum_{i=1}^{K} w_i \phi_i \in \mathcal{M}_\Phi$ and $h_0 \in \mathcal{P}_\Phi^{\text{all}}$, $f_0 = ah_1 + (1-a)h_0$. Let ϕ_i be such that $w_i > 0$. Thus, $f_0 = aw_i\phi_i + (1-aw_i)\frac{\left(a\sum_{j\neq i} w_j\phi_j + (1-a)h_0\right)}{1-aw_i}$; i.e., f_0 can be expressed as a nontrivial two-component mixture containing ϕ_i with nonzero weight. Thus, $a_{f_0}^{\phi_i} > 0$, which contradicts the consequent. Thus, $f_0 \notin \mathcal{F}\left(\mathcal{P}_\Phi^{\text{all}}, \mathcal{M}_\Phi\right)$ and consequently $f_0 \in \mathcal{I}_\Phi$. □

Lemma 2. *For any two density functions f and f_1 on $\mathcal{X}$, $a_f^{f_1} = \alpha$ if and only if there exists a sequence $\{x_j\}_{j=1}^{\infty}$ in* Supp(f_1) *such that* $\lim_{j\to\infty} \frac{f(x_j)}{f_1(x_j)} = \alpha$.

Proof. From Lemma 4 in (Jain et al. 2016), $\alpha = \inf\left\{\frac{f(x)}{f_1(x)} : f_1(x) > 0\right\}$ and $\alpha = a_f^{f_1}$ are equivalent statements. □

Lemma 3. *Any $f_0 \in \mathcal{P}_\Phi^{all}$ can be expressed as $ah_1 + (1-a)f_0^*$, where $a \in [0,1)$, $h_1 \in \mathcal{M}_\Phi$ and $f_0^* \in \mathcal{I}_\Phi$.*

Proof. From Lemma 4 in (Jain et al. 2016), it follows that there exists density ψ_1 such that $a_{\psi_1}^{\phi_1} = 0$ and $f_0 = a_{f_0}^{\phi_1}\phi_1 + \left(1 - a_{f_0}^{\phi_1}\right)\psi_1$. Similarly there exists density ψ_2 such that $a_{\psi_2}^{\phi_2} = 0$ and $\psi_1 = a_{\psi_1}^{\phi_2}\phi_2 + \left(1 - a_{\psi_1}^{\phi_2}\right)\psi_2$. Furthermore, from Lemma 2, since $a_{\psi_1}^{\phi_1} = 0$, there exists a sequence $\{x_j\}_{j=1}^{\infty}$ in Supp(ϕ_1) such that $\lim_{j\to\infty}\frac{\psi_1(x_j)}{\phi_1(x_j)} = 0$. Thus, $\lim_{j\to\infty}\frac{\psi_2(x_j)}{\phi_1(x_j)} = 0$ and consequently, applying Lemma 2, $a_{\psi_2}^{\phi_1} = 0$. Continuing with the same argument on ψ_{k-1} (for $k = 3, 4, \ldots, K$), a two-component mixture decomposition of ψ_{k-1} with ϕ_k as one of the components leads to ψ_k as the second component such that $a_{\psi_k}^{\phi_i} = 0$ for $i = 1, 2, \ldots, k$. In particular, $a_{\psi_K}^{\phi_i} = 0$ for $i = 1, 2, \ldots, K$. Note that from Lemma 1, $\psi_K \in \mathcal{I}_\Phi$. Now recursively substituting mixture representations of ψ_k in $f_0 = a_{f_0}^{\phi_1}\phi_1 + \left(1 - a_{f_0}^{\phi_1}\right)\psi_1$, for $k = 1, 2, \ldots, K$ leads to a mixture representation of f_0 as $\sum_{i=1}^{K} a_i\phi_i + \left(1 - \sum_{i=1}^{K} a_i\right)\psi_K$, where $a_i \in [0,1]$ and $\sum_{i=1}^{K} a_i \in [0,1]$. However, if $\sum_{i=1}^{K} a_i = 1$, then $f_0 \in \mathcal{M}_\Phi$ which leads to a contradiction. Thus, $\sum_{i=1}^{K} a_i < 1$. Thus, we can express f_0 as $ah_1 + (1-a)f_0^*$, where $a = \sum_{i=1}^{K} a_i$, $h_1 = \frac{\sum_{i=1}^{K} a_i\phi_i}{\sum_{i=1}^{K} a_i}$ and $f_0^* = \psi_K$. □

Lemma 4. *For $h_1 \in \mathcal{M}_\Phi$, $f_0 \in \mathcal{I}_\Phi$ and $a \in [0,1)$, $ah_1 + (1-a)f_0^*$ cannot be in $\mathcal{M}_\Phi$.*

Proof. If $ah_1 + (1-a)f_0^* \in \mathcal{M}_\Phi$, then there exists $\sum_{i=1}^{K} \bar{\gamma}_i\phi_i$ with $\bar{\boldsymbol{\gamma}} \in \Delta^{K-1}$ such that $ah_1 + (1-a)f_0^* = \sum_{i=1}^{K} \bar{\gamma}_i\phi_i$. Since h_1 is in $\mathcal{M}_\Phi$, it can be expressed as $\sum_{i=1}^{K} \gamma_i\phi_i$, for some $\boldsymbol{\gamma} \in \Delta^{K-1}$. Thus,

$$a\sum_{i=1}^{K}\gamma_i\phi_i + (1-a)f_0^* = \sum_{i=1}^{K}\bar{\gamma}_i\phi_i$$
$$\Rightarrow \sum_{i=1}^{K}(a\gamma_i - \bar{\gamma}_i)\phi_i = -(1-a)f_0^*$$

If $a\gamma_i = \bar{\gamma}_i$ for all $i = 1, 2, \ldots, K$, then $a = 1$. However, $a < 1$. Thus, there exists an $j \in \{1, 2, \ldots, K\}$ such that $a\gamma_j \neq \bar{\gamma}_j$. Restricting the above equation to the support of ϕ_j leads to

$$(a\gamma_j - \bar{\gamma}_j) = -(1-a)\frac{f_0^*(x)}{\phi_j(x)} \quad \text{for all } x \text{ in Supp}(\phi_j),$$

under the disjoint kernel support assumption. Since $f_0 \in \mathcal{I}_\Phi$, $a_{f_0^*}^{\phi_j} = 0$ follows from Lemma 1. Thus, there exists a sequence $\{x_k\}_{k=1}^{\infty}$ in Supp(ϕ_j) such that limit of the right hand side along the sequence is 0. However, the left hand side is a nonzero constant in the limit. Hence a contradiction. Thus, $ah_1 + (1-a)f_0^* \notin \mathcal{M}_\Phi$ □

References

Bekker, J., and Davis, J. 2018. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI 2018, 2712–2719.

Blanchard, G.; Lee, G.; and Scott, C. 2010. Semi-supervised novelty detection. *J Mach Learn Res* 11:2973–3009.

Chang, S.; Zhang, Y.; Tang, J.; Yin, D.; Chang, Y.; Hasegawa-Johnson, M. A.; and Huang, T. S. 2016. Positive-unlabeled learning in streaming networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2016, 755–764.

Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT 2008, 38–53.

Dasgupta, S. 2000. Experiments with random projection. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI 2000, 143–151.

de Amorim, R. C., and Hennig, C. 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* 324:126–145.

Denis, F.; Gilleron, R.; and Letouzey, F. 2005. Learning from positive and unlabeled examples. *Theor Comput Sci* 348(16):70–83.

Denis, F. 1998. PAC learning from positive statistical queries. In *Proceedings of the 9th International Conference on Algorithmic Learning Theory*, ALT 1998, 112–126.

du Plessis, M. C., and Sugiyama, M. 2012. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of the 29th International Conference on Machine Learning*, ICML 2012, 823–830.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2014. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, NIPS 2014, 703–711.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2017. Class-prior estimation for learning from positive and unlabeled data. *Mach Learn* 106(4):463–492.

Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; and Obradovic, Z. 2001. Intrinsically disordered protein. *J Mol Graph Model* 19(1):26–59.

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*, KDD 2008, 213–220.

Fields, S., and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245–246.

Geurts, P. 2011. Learning from positive and unlabeled examples by enforcing statistical significance. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, AISTATS 2011, 305–314.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–161.

Hsieh, C. J.; Natarajan, N.; and Dhillon, I. S. 2015. PU learning for matrix completion. In *Proceedings of the 32rd International Conference on Machine Learning*, ICML 2015, 2445–2453.

Hsieh, Y. G.; Niu, G.; and Sugiyama, M. 2019. Classification from positive, unlabeled and biased negative data. In *Proceedings of the 36th International Conference on Machine Learning*, ICML 2019, 2820–2829.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Scholkopf, B. 2006. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, NIPS 2006, 600–607.

Jain, S.; White, M.; Trosset, M. W.; and Radivojac, P. 2016. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*.

Jain, S.; White, M.; and Radivojac, P. 2017. Recovering true classifier performance in positive-unlabeled learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI 2017, 2066–2072.

Latinne, P.; Saerens, M.; and Decaestecker, C. 2001. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: evidence from a multi-class problem in remote sensing. In *Proceedings of the 18th International Conference on Machine Learning*, ICML 2001, 298–305.

Lichman, M. 2013. UCI Machine Learning Repository.

Liu, B.; Dai, Y.; Li, X.; Lee, W.; and Yu, P. S. 2003. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, ICDM 2003, 179–186.

Menon, A. K.; van Rooyen, B.; Ong, C. S.; and Williamson, R. C. 2015. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML 2015, 125–134.

Ramaswamy, H. G.; Scott, C.; and Tewari, A. 2016. Mixture proportion estimation via kernel embedding of distributions. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML 2016, 2996–3004.

Ramola, R.; Jain, S.; and Radivojac, P. 2019. Estimating classification accuracy in positive-unlabeled learning: characterization and correction strategies. *Pac Symp Biocomput* 24:124–135.

Sanderson, T., and Scott, C. 2014. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, AISTATS 2014, 850–858.

Tran, P. V. 2013. Supervised link prediction in social networks with positive and unlabeled examples. *Military Operations Research* 18(3):53–62.

Vucetic, S., and Obradovic, Z. 2001. Classification on data with biased class distribution. In *Proceedings of the 12th European Conference on Machine Learning*, ECML 2001, 527–538.

Ward, G.; Hastie, T.; Barry, S.; Elith, J.; and Leathwick, J. 2009. Presence-only data and the EM algorithm. *Biometrics* 65(2):554–563.

Youngs, N.; Shasha, D.; and Bonneau, R. 2015. Positive-unlabeled learning in the face of labeling bias. In *Proceedings of the IEEE International Conference on Data Mining Workshop*, ICDMW 2015, 2375–9259.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, ICML 2004, 114–314.