Learning from Positive and Unlabeled Data with Augmented Classes

Zhongnian Li\(^1\), Liutao Yang\(^2\), Zhongchen Ma\(^3\), Tongfeng Sun\(^1\), Xinzheng Xu\(^1\)\(^*\) and Daoqiang Zhang\(^2\)

\(^1\)Department of Computer Science, China University of Ming and Technogy
\(^2\)Department of Computer Science, Nanjing University of Aeronautics and Astronautics
\(^3\)Department of Computer Science, Jiangsu University
zhongnianli@cumt.edu.cn, xxzheng@cumt.edu.cn

Abstract

Positive Unlabeled (PU) learning aims to learn a binary classifier from only positive and unlabeled data, which is utilized in many real-world scenarios. However, existing PU learning algorithms cannot deal with the real-world challenge in an open and changing scenario, where examples from unobserved augmented classes may emerge in the testing phase. In this paper, we propose an unbiased risk estimator for PU learning with Augmented Classes (PUAC) by utilizing unlabeled data from the augmented classes distribution, which can be easily collected in many real-world scenarios. Besides, we derive the estimation error bound for the proposed estimator, which provides a theoretical guarantee for its convergence to the optimal solution. Experiments on multiple realistic datasets demonstrate the effectiveness of proposed approach.

1 Introduction

Learning from Positive and Unlabeled (PU) data is a type of weakly supervised learning, which has drawn considerable attention in many real-world scenarios. The goal of PU learning is to train a binary classifier by using only positive and unlabeled data without the assistance of negative label, which requires huge costs in some tasks. PU learning has many practical applications, such as, text classification, image annotation, time series categorization, bio-medicine analysis and so on.

Previous researches\(^{[2020b], [2019], [2018], [2019], [2014], [2015], [2014], [2013]}\) focus on handling unlabeled data to solve the PU learning problem. Specifically, a line of effective algorithms aim to extract possible negative examples from unlabeled data\(^{[2003], [2014], [2014]}\), and then train a binary classifier. The performance of this category is heuristic and strongly influenced by selected negative examples. Another category methods train PU learning classifiers by treating unlabeled data as noise negative examples. Thus, the classifier can be trained by utilizing the small weights for negative data\(^{[2014], [2015]}\), which heavily relies on the choice of weight for unlabeled data.

However, it is noteworthy that exist studies on PU learning were in a stable scenario rather than non-stationary environment\(^{[2021], [2015], [2020a], [2020a], [2014]}\), where some examples derived from unobserved classes in training phase might emerge in the testing data\(^{[2020], [2014]}\). In this paper, we focus on PU learning with Augmented Classes (PUAC). Augmented classes classification desire to make reliable prediction, which not only identifies augmented classes but also classifies the data form observed classes accurately in testing phase.

\(^*\)Corresponding author

Preprint. Under review.
Figure 1: One example of Positive and Unlabeled data with Augmented Classes. The augmented class (Tiger) only appear in testing dataset, which is not observed in positive and unlabeled data.

For example, in the task of felidae image annotation utilizing PU learning in the Internet, the user may only label cat images as positive data and the absence label images as unlabeled data as shown in Fig. 1. In the traditional PU learning setting, unlabeled data only consist of cat and non-felidae images. However, since the environment is open and change, there may appear some images from augmented classes, such as tiger images. When an image of tiger comes, exist PU classifiers nearly predict it in the cat or non-felidae class, which will degrade performance of those classifiers. When facing some open and change scenarios in the real world, a mature PU learning algorithm need to work in PUAC setting.

In order to deal with the PUAC problem, we propose an unbiased risk estimator called UPUAC, short for Unbiased risk estimator of Positive and Unlabeled learning with Augmented Classes. UPUAC exploits unlabeled data from the augmented classes distribution, which can be collected easily from many real-world scenarios. More concretely, we rewrite the risk into an equivalent expression by utilizing positive, unlabeled and augmented classes distributions, which will lead to an unbiased risk estimators for PUAC. Besides, we provide a theoretical analysis of estimation error bound which certainly guarantees the estimator converges to the optimal solution. Experimentally, by comparing with existing state-of-the-art PU learning approach, our UPUAC achieves the best classification performance when the augmented classes emerges on multiple realistic datasets. Beyond that, we test proposed UPUAC with some inaccurate class priors, which may be obtain by mixture proportion estimation [du Plessis et al., 2017], Ramaswamy et al. [2016].

2 Preliminaries

In this section, we describe PUAC problem setting and review notations of risk rewrite briefly.

2.1 Problem Setting of PUAC Learning

In traditional PU learning, the learner collects two datasets $D_p = \{x_p^i\}_{i=1}^{n_p}$ and $D_u = \{x_u^i\}_{i=1}^{n_u}$ sampled from positive distribution $P_p(x)$ and unlabeled distribution $P_u(x)$ respectively, where $x \in \mathbb{R}^d$ is a $d$-dimensional feature and $\mathcal{X}$ is the feature space. Let $y_i \in \{p, n\}$ denotes the class label and $\mathcal{Y}' = \{p, n\}$ denotes the label space for traditional PU learning. In the PUAC problem, the learner requires to train a classifier by using the data form the augmented classes distribution, where data from unobserved augmented classes might emerge. Since the number of augmented classes is unknown, the data generated form unobserved augmented classes will be predicted as a single class $a$. In our setup, an unlabeled dataset $D_a = \{x_a^i\}_{i=1}^{n_a}$ sampled from the augmented classes distribution $P_a(x)$, and Let $y_i \in \{p, n, a\}$ denotes the class label, $\mathcal{Y} = \{p, n, a\}$ denotes the label space of PUAC. Let $p_p(x) = p(x|y = p)$, $p_n(x) = p(x|y = n)$, $p_a(x) = p(x|y = a)$ denote the class-conditional densities for positive, unlabeled and augmented classes distributions respectively, $\bar{D} = D_p \cup D_u \cup D_a$.
where denotes the aggregated dataset, \( P(x, y) \) denotes the distribution of \( \overline{D} \), \( \pi_p = p(y = p) \), \( \pi_n = p(y = n) \) and \( \pi_a = p(y = a) \) denote the class prior probabilities for aggregated dataset.

**Data generation process** Let \( \theta_p^p, \theta_n^p, \theta_a^p, \theta_n^a, \theta_a^a \) be class priors for positive, unlabeled and augmented classes distributions. The data collected for PUAC is assumed to be i.i.d sample from the marginal densities as follows:

\[
P_p(x) = \theta_p^p p_p(x) + \theta_n^p p_n(x) + \theta_a^p p_a(x), \\
P_u(x) = \theta_u^p p_p(x) + \theta_u^a p_n(x) + \theta_u^a p_a(x), \\
P_a(x) = \theta_a^p p_p(x) + \theta_a^a p_n(x) + \theta_a^a p_a(x)
\]

(1)

where \( \theta_p^p = 1, \theta_n^p = \theta_a^p = 0, \theta_u^a + \theta_n^a = 1, \theta_a^a = 0 \) and \( \theta_u^a + \theta_n^a + \theta_a^a = 1 \). In the PUAC problem setting, the number of free class priors is three, which could be estimated by various methods that is similar to the traditional PU learning.

Data generation process states that the distribution of PUAC data can be regarded as a mixture of positive, negative and augmented classes with some class priors. It is noteworthy that it does not matter that augmented classes distribution equals to the testing distribution. It means that collecting data from augmented classes distribution is easier than from the testing distribution, which is not a stable distribution during the collecting phase in the non-stationary environments. In the experiments, empirical results confirm the robustness for class distribution shifting in the testing distribution.

### 2.2 Risk Rewrite

In supervised positive, negative and augmented classes classification, let \( f : \mathcal{X} \rightarrow \mathcal{Y} \) denotes the decision function, i.e., \( f \) may be any multi-class classifier. Let \( l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \) be the loss function, and given class priors \( \pi_p = p(y = p) \), \( \pi_n = p(y = n) \) and \( \pi_a = p(y = a) \) the risk of \( f \) is formulated as follows:

\[
R_l(f) = \mathbb{E}_{(x, y) \sim p(x, y)}[l(f(x), y)] = \sum_{i = p, n, a} \pi_i \mathbb{E}_{x \sim p_i}[l(f(x), i)]
\]

(2)

where \( p(x, y) \) denotes the distribution of supervised dataset, \( p_i \) denotes the \( i \)th class-conditional density, i.e., \( p_i(x) = P(x | y = i) \).

In the PUAC problem, since the negative and augmented classes examples are unavailable, we cannot estimate \( \mathbb{E}_{x \sim p_n}[l(f(x), n)] \) and \( \mathbb{E}_{x \sim p_a}[l(f(x), a)] \), which means that Eq[2] is unable to calculate for PUAC directly. Thus, we rewrite the risk by using unlabeled and augmented classes distribution to replace the supervised multi-class risk \( R_l(f) \). The definition of risk rewrite for PUAC is shown as follows.

**Definition 1.** The risk \( R_l(f) \) is rewritable for PUAC setting on the basis of three marginal densities \( P_p, P_u \) and \( P_a \), and if and only if there exist constants \( \alpha_p, \alpha_n, \alpha_a, \beta_p, \beta_n, \beta_a, \gamma_p, \gamma_n \) and \( \gamma_a \), such that for any \( f \) it hold that

\[
R_l(f) = \mathbb{E}_{P_p}[\tilde{l}_p(f(x))] + \mathbb{E}_{P_u}[\tilde{l}_u(f(x))] + \mathbb{E}_{P_a}[\tilde{l}_a(f(x))]
\]

(3)

where \( \tilde{l}_p(f(x)) = \alpha_p l(f(x), p) + \alpha_n l(f(x), n) + \alpha_a l(f(x), a) \), \( \tilde{l}_u(f(x)) = \beta_p l(f(x), p) + \beta_n l(f(x), n) + \beta_a l(f(x), a) \) and \( \tilde{l}_a(f(x)) = \gamma_p l(f(x), p) + \gamma_n l(f(x), n) + \gamma_a l(f(x), a) \) are the corrected loss functions.

If the risk is rewritable for PUAC setting, the PUAC risk can be expressed as expectation over \( p_p(x) \), \( p_n(x) \) and \( p_a(x) \) separately, which builds a bridge to achieve an unbiased risk estimator.

### 3 Learning from PUAC

In this section, we first prove the risk is rewritable for PUAC setting. Then we describe the practical implementation for learning from positive and unlabeled learning with augmented classes. In this paper, all the proofs can be found in Appendix.

#### 3.1 Risk Rewrite for PUAC

Now, we attempt to express the supervised positive, negative and augmented classes risk \( R_l(f) \) on the basis of the three marginal densities given in Section 2.1. According to the Definition 1, if we
find the constants $\alpha_p, \alpha_n, \alpha_a, \beta_p, \beta_n, \beta_a, \gamma_p, \gamma_n$ and $\gamma_a$, we could rewrite the PUAC risk. Fortunately, we show that those constants can be obtained by utilizing class priors. An answer to find constants is given by the following Theorem 2.

**Theorem 2.** Fix class priors $\theta^p, \theta^n, \theta^a, \theta^u$, then the supervised multi-classification $R_\ell(f)$ is rewritable, by letting $\alpha_p = \frac{\theta^p}{\theta^p + \theta^n + \theta^a}$, $\alpha_n = -\frac{\theta^n}{\theta^p + \theta^n + \theta^a}$, $\alpha_a = -\frac{\theta^a}{\theta^p + \theta^n + \theta^a}$, $\beta_p = 0$, $\beta_n = \frac{\theta^n}{\theta^p + \theta^n + \theta^a}$, $\beta_a = -\frac{\theta^a}{\theta^p + \theta^n + \theta^a}$, $\gamma_p = 0$, $\gamma_n = 0$ and $\gamma_a = -\frac{\theta^a}{\theta^p + \theta^n + \theta^a}$.

As the results of above theorem, we can express the PUAC risk using data collected from positive, unlabeled and augmented classes distribution. Let $\tilde{l}_p(f(x)), \tilde{l}_u(f(x))$ and $\tilde{l}_a(f(x))$ be

$$\tilde{l}_p(f(x)) = \frac{\pi_p}{\theta^p} l(f(x), p) - \frac{\theta^n}{\theta^p + \theta^n + \theta^a} l(f(x), n) + \frac{\theta^a}{\theta^p + \theta^n + \theta^a} l(f(x), a)$$

(4)

$$\tilde{l}_u(f(x)) = \frac{\pi_u}{\theta^u} l(f(x), n) - \frac{\theta^n}{\theta^p + \theta^n + \theta^a} l(f(x), n) + \frac{\theta^a}{\theta^p + \theta^n + \theta^a} l(f(x), a)$$

(5)

$$\tilde{l}_a(f(x)) = \frac{\pi_a}{\theta^a} l(f(x), a)$$

(6)

**Proposition 3.** The classification risk can be rewritten equivalently as

$$R_{PUAC}(f) = \mathbb{E}_{x \sim P_x}[\tilde{l}_p(f(x))] + \mathbb{E}_{x \sim P_u}[\tilde{l}_u(f(x))] + \mathbb{E}_{x \sim P_a}[\tilde{l}_a(f(x))]$$

$$= \mathbb{E}_{x \sim P_p}\left[\frac{\pi_p}{\theta_p} l(f(x), p) - \frac{\theta^n}{\theta^p + \theta^n + \theta^a} l(f(x), n) + \frac{\theta^a}{\theta^p + \theta^n + \theta^a} l(f(x), a)\right]$$

(7)

$$+ \mathbb{E}_{x \sim P_u}\left[\frac{\pi_u}{\theta^u} l(f(x), n) - \frac{\theta^n}{\theta^p + \theta^n + \theta^a} l(f(x), n) + \frac{\theta^a}{\theta^p + \theta^n + \theta^a} l(f(x), a)\right] + \mathbb{E}_{x \sim P_a}\left[\frac{\pi_a}{\theta^a} l(f(x), a)\right]$$

The Proposition 3 naturally leads to an unbiased estimator for PUAC. In the empirical minimization framework, Eq(7) is replaced with their empirical as follows:

$$\hat{R}_{PUAC}(f) = \frac{1}{n_p} \sum_{i=1}^{n_p} \tilde{l}_p(f(x)) + \frac{1}{n_u} \sum_{i=1}^{n_u} \tilde{l}_u(f(x)) + \frac{1}{n_a} \sum_{i=1}^{n_a} \tilde{l}_a(f(x))$$

(8)

where the $n_p, n_u$ and $n_a$ denotes the number of examples collected form positive, unlabeled and augmented classes data respectively. In the rest of the paper, the process of obtaining the empirical risk minimizer of Eq(8) i.e., $\hat{f}_{PUAC} = \arg \min_{f \in F} R_{PUAC}(f)$ is named as PUAC learning. In this paper, we propose a ERM-based PUAC learning, and consequently, $\hat{f}_{PUAC}$ can be obtained by optimizing the Eq(8).

**Special case** Consider the Eq(8) by specifying class priors. It is obvious that PUAC problem reduces to Positive and Unlabeled learning, if $\theta^u = 0$ and $\theta^a = 0$ and the augmented classes data is seen as unlabeled examples. Then, we can train the classifier by using the positive and unlabeled learning. Besides, PUAC problem reduces to the standard Positive and Unlabeled learning, if $\theta^a = 0$.

### 3.2 Practical Implementation

In this section, we investigate the practical implementation when the deep model is employed as classifier for PUAC.

**Loss function for PUAC problem** We investigate appropriate choice of the loss function $l$. Typically, the discrete loss is computationally hard, such as, $\mathbb{1}[y \neq p]$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Thus, surrogate losses for multi-class learning with consistency properties are used to replace discrete losses. Some practical examples of common multi-class loss function could be considered which satisfy the consistency proven in [Zhang, 2004].

One group of the loss functions $l$ is the zero-one loss, i.e., $l_{0-1}(f(x), y) = \mathbb{1}[f(x) \neq y]$. Then, Eq(7) is known as the classification error. In the multi-classification, the One-Versus-Rest (OVR) strategy with margin loss $l_{mar}$ has the guarantee that $R_{mar}(f_p, f_n, f_a)$ and $R_{mar}(f_p, f_n, f_a)$ have the same minimizer over all decision functions $\theta^a$. The $l_{mar}$ is formulated as

$$l_{mar}(f_p(x), f_n(x), f_a(x), y) = \phi(f_p(x)) + \sum_{i=p, n, a} \phi(-f_i(x))$$

(9)
where $f_i$ denotes the classifier for $i^{th}$ class, $\phi(z): \mathbb{R} \rightarrow [0, \infty)$ is a binary convex surrogate loss, and $z = yf(x)$ denotes the margin. Many margin loss functions satisfy consistency properties for multi-class problem, such as square loss $\phi(z) = (1 - z)^2$. Then, According to the OVR strategy with margin loss, the loss $\tilde{l}_p(f_p(x), f_n(x), f_o(x))$ is formulated as

$$
\tilde{l}_p(f_p(x), f_n(x), f_o(x)) = \frac{\pi_p}{\theta_p} \phi(f_p(x)) + \phi(-f_n(x)) + \phi(-f_o(x))
$$

Another group of loss for multi-class problem is ordinal regression loss Finocchiaro et al. [2019], Ramaswamy and Agarwal [2016], Pedregosa et al. [2017], e.g. using absolute distance:

$$
l_{ord}(f_o(x), y) = |f_o(x) - y| \quad \forall f_o(x), y \in \{p, n, a\}
$$

where label $p = 1$, $n = 2$ and $a = 3$, $f_o(x)$ denotes the multi-class classifier. Many surrogate losses satisfy consistency properties for ordinal regression loss in multi-class learning, such as absolute surrogate loss. Compared to zero-one surrogate loss with three dimension classifier, the prediction dimension of absolute surrogate is one, which can be employed to develop computationally efficient method Finocchiaro et al. [2019]. Let $\phi_p(f_o(x))$ denotes the absolute surrogate loss, the multi-class loss $l_{abs}$ can be formulated as

$$
l_{abs}(f_o(x), y) = \phi_p(f_o(x)) = |f_o(x) - y| \quad \forall y \in \{p, n, a\}, f_o(x) \in R
$$

We now give the risk minimization of Eq.7 as follows

$$
R_{PUAC,l_{abs}}(f_o) = \mathbb{E}_{x \sim P} \left[ \frac{\pi_p}{\theta_p} |f_o(x) - p| - \frac{\theta_p^p}{\theta_p^p} |f_o(x) - n| + \frac{\theta_n^p}{\theta_p^p} |f_o(x) - a| \right] + \mathbb{E}_{x \sim P} \left[ \frac{\pi_n}{\theta_n} |f_o(x) - n| - \frac{\theta_n^p}{\theta_n^a} |f_o(x) - a| \right] + \mathbb{E}_{x \sim P} \left[ \frac{\pi_a}{\theta_a} |f_o(x) - a| \right]
$$

Given the empirical loss of Eq.13, we can obtain the \(\hat{f}_{PUAC,o}\) by powerful stochastic optimization algorithm for deep models. After obtain the multi-class classifier, we can construct the predictor as $f : X \rightarrow Y$ with $f(x) = \arg \max_{k \in \{p, n, a\} } \hat{f}_{PUAC,o}(x) - k$.

**Class Prior Estimation** It is noteworthy that the optimization of the proposed model requires estimating class priors, when the class priors were assumed to be unknown. In this section, we propose a class prior estimation algorithm from PU and augmented classes datasets.

Since we have samples collected from positive and unlabeled distributions, we can estimate $\theta_p^p$ and $\theta_n^p$ by mixture proportion estimation methods Ramaswamy et al. [2016] easily. For estimating $\theta_n^a$, we employ the kernel embedding proposed by, which use reproducing kernel Hilbert distance to estimate the mixture proportion. If $\theta_n^a$ and $\theta_n^u$ are estimated, class-conditional distribution can be expressed as $p_n(x) = \frac{P_n(x) - \theta_n^u P_u(x)}{\theta_n^u}$. Then, the kernel mapping distribution can be obtained from positive and unlabeled datasets. Since we have some examples collected from augmented classes distribution, the kernel mean distance can be computed between $p_n(x)$ and augmented classes distribution $P_a$ from those datasets. Thus, class prior $\theta_n^a$ can be estimated for PUAC problem.

## 4 Theoretical Analysis

In this section, we first study consistency property of proposed PUAC risk $R_{PUAC,l_{mar}}$. Then, we provide the theoretical analysis of error bound.

Now, we show the Bayes classifiers for OVR strategy with zero-one loss $l_{0-1}$, and show that the PUAC risk $R_{PUAC,l_{mar}}$ is consistent with the supervised multi-class risk $R_{l_{0-1}}$. 

5
Definition 4. (Bayes classifiers for zero-one loss). Let \( f^* : X \rightarrow Y \) be a classifier, \( P(y = c|x) \) denotes the conditional density. Then, the optimal classifier is obtained by minimizing the classification error, which often was referred to the Bayes classifier for zero-one loss given by

\[
f^*(x) = \arg \max_{c \in \{p,n,a\}} P(y = c|x)
\]

The following theorem states that by minimizing the risk of \( R_{PUAC,lm} \), we can obtain the Bayes classifier for zero-one loss.

Theorem 5. Let \( f_p, f_n \) and \( f_a \) denotes binary classifiers for positive, negative and augmented classes, and \( f(x) = \arg \max_{k \in \{p,n,a\}} f_k(x) \), the surrogate loss \( \phi(z) \) is convex, bounded below, differentiable, and \( \phi(z) < \phi(-z) \) when \( z > 0 \), then for any \( \epsilon > 0 \), there exists \( \epsilon > 0 \) such that

\[
R_{PUAC,lm}(f_p, f_n, f_a) \leq R_{PUAC,lm}^* + \epsilon_2 \Rightarrow R_{0-1}(f) \leq R_{0-1}^* + \epsilon_1
\]

where \( R_{PUAC,lm}^* = \min_{f_p, f_n, f_a} R_{PUAC,lm}(f_p, f_n, f_a) \) and \( R_{0-1}^* = \min_f R_{0-1}(f) = R_{0-1}(f^*) \) denotes the Bayes error for multi-class distribution.

Theorem 5 analysis the consistency property for proposed method, which means that we can obtain classifier achieving Bayes rule. According to Definition 4, it is obvious that the Bayes classifiers will achieve optimal classification error under class probabilities shift as usual in the testing distribution. By minimizing the PUAC risk \( R_{PUAC,lm} \), we can get well-behaved classifiers the same as learning with supervised multi-class data.

Now, we analyze the generalization error bounds for the proposed approach implemented by deep neural networks using OVR strategy. Let \( f = (f_p, f_n, f_a) \) denotes classification vector function in the deep network hypothesis set \( \mathcal{F} \). Assume there is \( C_\phi > 0 \), such that \( \sup_z \phi(z) \leq C_\phi \),

\[
K_a = \max(\frac{\pi_n}{\phi_0}, \frac{\phi_0}{\phi_a}, \frac{a \phi_0}{\phi_a}, \frac{\pi_n}{\phi_a}), \quad K_u = \max(\frac{\pi_n}{\phi_0}, \frac{\phi_0}{\phi_a}), \quad K_a = \frac{\pi_n}{\phi_a}.
\]

Let \( L_\phi \) be the Lipschitz constant of \( \phi \), we can establish the following lemma.

Lemma 6. For any \( \delta > 0 \), with the probability at least \( 1 - \delta/2 \),

\[
\sup_{f \in \mathcal{F}} \left| R_p(f) - \hat{R}_p(f) \right| \leq 6K_pL_\phi R_{n_p}(\mathcal{F}) + 3K_pC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_p}}
\]

\[
\sup_{f \in \mathcal{F}} \left| R_u(f) - \hat{R}_u(f) \right| \leq 4K_uL_\phi R_{n_u}(\mathcal{F}) + 2K_uC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_u}}
\]

\[
\sup_{f \in \mathcal{F}} \left| R_a(f) - \hat{R}_a(f) \right| \leq 2K_aL_\phi R_{n_a}(\mathcal{F}) + K_aC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_a}}
\]

where \( R_p(f) = \mathbb{E}_{x \sim P_p}[\hat{f}_p(f(x))] \), \( R_u(f) = \mathbb{E}_{x \sim P_u}[\hat{f}_u(f(x))] \) and \( R_a(f) = \mathbb{E}_{x \sim P_a}[\hat{f}_a(f(x))] \), \( \hat{R}_p(f) \), \( \hat{R}_u(f) \) and \( \hat{R}_a(f) \) denote the empirical risk estimator to \( R_p(f) \), \( R_u(f) \) and \( R_a(f) \) respectively, \( R_{n_p}(\mathcal{F}) \), \( R_{n_u}(\mathcal{F}) \) and \( R_{n_a}(\mathcal{F}) \) are the Rademacher complexities of \( \mathcal{F} \) for the sampling of size \( n_p \), from \( P_p(x) \), the sampling of size \( n_u \) from \( P_u(x) \) and the sampling of size \( n_a \) from \( P_a(x) \).

Based on the Lemma 6, we can obtain the estimation error bound as follows.

Theorem 7. For any \( \delta > 0 \), with the probability at least \( 1 - \delta/2 \),

\[
R_{lm}(\hat{f}_{puac}) - \min_{f \in \mathcal{F}} R_{lm}(f) \leq 12K_pL_\phi R_{n_p}(\mathcal{F}) + 8K_uL_\phi R_{n_u}(\mathcal{F}) + 4K_aL_\phi R_{n_a}(\mathcal{F})
\]

\[
+ 3K_pC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_p}} + 2K_uC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_u}} + K_aC_\phi \sqrt{\frac{\ln(4/\delta)}{2n_a}}
\]

where \( \hat{f}_{puac} \) is trained by minimizing the PUAC risk \( R_{PUAC,lm} \).

Lemma 6 and Theorem 7 show that, with a growing number of positive, unlabeled and augmented classes data, the estimation error of the trained classifiers decreases, which means that the proposed
Table 1: Classification accuracy of each algorithm on benchmark datasets, with varying classes and the number of examples. P = 1, N=3, A=5 means that classes 1, 3, 5 are taken as positive, negative, and augmented classes respectively. # PU and # AC denote the number of training examples in PU and augmented classes datasets. We report the mean and standard deviation of results over 5 trials. The best method is shown in bold (under 5% t-test).

| Dataset       | P, N, A | #PU | #AC | UPU | NPU | MPU | AREA | UPUAC |
|---------------|---------|-----|-----|-----|-----|-----|------|-------|
| MNIST         | 1,3,5   | 17791 | 10703 | 37.77 ± 7.1 | 40.38 ± 1.2 | 73.65 ± 0.58 | 57.27 ± 4.22 | 98.90 ± 0.24 |
| Fashion       | 1,3,5   | 11320 | 6680  | 46.41 ± 6.67 | 41.64 ± 5.15 | 90.39 ± 1.57 | 86.39 ± 0.33 | 98.70 ± 0.11 |
| Kuzushiji     | 5.3,1   | 10625 | 7669  | 42.90 ± 1.84 | 42.33 ± 2.98 | 90.14 ± 1.12 | 83.16 ± 6.09 | 99.47 ± 0.08 |
|               | 3.5,1,2 | 10102 | 14150 | 25.24 ± 2.06 | 27.20 ± 2.30 | 89.66 ± 0.33 | 77.86 ± 1.68 | 99.42 ± 0.13 |
|               | 9,8,7,6 | 10313 | 13670 | 39.13 ± 3.78 | 42.04 ± 1.96 | 91.10 ± 0.41 | 72.68 ± 5.37 | 98.39 ± 0.21 |
|               | 1,3,5   | 10335 | 7665  | 39.98 ± 3.71 | 34.14 ± 11.76 | 94.51 ± 0.10 | 87.70 ± 0.81 | 98.60 ± 0.06 |
| Fashion       | 2,5,8   | 8650  | 9350  | 33.46 ± 1.30 | 37.00 ± 6.90 | 89.43 ± 0.47 | 65.01 ± 3.00 | 98.80 ± 0.08 |
| Kuzushiji     | 9,8,7,6 | 8658  | 9342  | 29.22 ± 1.05 | 27.03 ± 3.05 | 86.45 ± 1.37 | 65.22 ± 1.52 | 97.30 ± 0.11 |
|               | 1,3,5   | 8674  | 9326  | 33.74 ± 1.05 | 31.52 ± 0.51 | 74.72 ± 0.55 | 49.49 ± 1.27 | 94.31 ± 0.52 |
|               | 2,4,6   | 10320 | 7680  | 31.44 ± 3.33 | 31.05 ± 2.56 | 77.36 ± 0.32 | 53.97 ± 0.19 | 93.87 ± 0.50 |
|               | 9,8,6   | 9987  | 8013  | 36.12 ± 6.92 | 35.18 ± 3.66 | 81.34 ± 0.71 | 59.70 ± 1.47 | 96.15 ± 0.33 |

Table 2: Identification accuracy of augmented classes on benchmark datasets, with varying classes and the number of examples. We report the mean and standard deviation of results over 5 trials. P = 1, N=3, A=5 means that classes 1, 3, 5 are taken as positive, negative, and augmented classes respectively. # train AC denotes the number of training examples in augmented classes datasets. The best method is shown in bold (under 5% t-test).

| Dataset       | P, N, A | #train AC | Fashion | Kuzushiji |
|---------------|---------|-----------|---------|-----------|
|               | 5.3,1   | 1.3,5     | 9.8,7   | 9.8,7     |
|               | 9955    | 10703     | 11173   | 7665      |
| MPU           | 84.68±1.33 | 93.55±1.52 | 75.71±6.93 | 91.90±1.34 |
| AREA          | 59.05±1.79 | 72.77±4.50 | 54.01±3.83 | 88.22±2.27 |
| UPUAC         | 98.22±0.48 | 98.39±0.15 | 97.28±0.55 | 96.98±0.31 |

method is consistent. When deep network hypothesis set \( \mathcal{F} \) is fixed and \( \mathcal{R}_n(\mathcal{F}) \leq C/\sqrt{n} \), we have \( \mathcal{R}_{n_p}(\mathcal{F}) = O(1/\sqrt{mp}) \), \( \mathcal{R}_{n_u}(\mathcal{F}) = O(1/\sqrt{n_u}) \) and \( \mathcal{R}_{n_a}(\mathcal{F}) = O(1/\sqrt{n_a}) \), then

\[
n_p, n_u, n_a \rightarrow \infty \implies R_{1mar}(\hat{f}_{puac}) - \min_{f \in \mathcal{F}} R_{1mar}(f) \rightarrow 0
\]

Lemma 6 and Theorem 7 theoretically justify the effective of proposed method in exploiting augmented classes data.

## 5 Experiments

In this section, we experimentally analyze the proposed approach from three aspects: 1) Classification accuracy for each class and identification accuracy for augmented classes comparing with state-of-the-art methods; 2) Robustness for inaccurate training class priors; 3) Accuracy for class distribution shifting in the testing distribution.

### 5.1 Performance Comparison

**Datasets:** In this section, we conduct experiments on five datasets, i.e., MNIST, Fashion-MNIST, Kuzushiji-MNIST, SVNH and CIFAR-10. The MNIST, Fashion-MNIST and Kuzushiji-MNIST datasets consists of 70000 examples which originally have 10 classes. The SVHN dataset consists of 73257 examples with 10 classes. The CIFAR-10 dataset consists of 60000 examples associated with a label form 10 classes. We constructed the positive, unlabeled and augmented classes datasets as follows: we first select three classes form original datasets as positive, negative and augmented classes datasets respectively. Then, we randomly select examples from positive dataset as positive class dataset, examples from positive and negative datasets as unlabeled dataset, and examples from positive, negative and augmented classes as augmented dataset. In the constructed procedure, each example is selected into only one dataset. For performance comparison, we use the original testing datasets as the testing datasets.

**Common Setup:** We conduct experiments using OVR strategy implemented by margin square loss \( \phi(z) = (1 - z)^2 \). As a classifier, we also used neural network with 4 convolutional layers and 2 fully-connected layers for CIFAR-10, neural network with 3 convolutional layers and 2 fully-connected
Table 3: Classification accuracy for inaccurate training class priors on MNIST dataset. $P = 1, N=3, A=5$ means that classes 1, 3, 5 are taken as positive, negative, and augmented classes respectively. $\#PU$ and $\#AC$ denote the number of training examples in PU and augmented classes datasets. We report mean and standard deviation over 3 trials for varying degrees of inaccuracies.

| Dataset | P, N, A | $\#PU$ | $\#AC$ | $\eta_p^u = 0.8$ | $\eta_p^a = 0.9$ | $\eta_n^u = 1$ | $\eta_n^a = 1.1$ | $\eta_a^u = 1.2$ |
|---------|---------|--------|--------|-------------------|------------------|------------------|------------------|------------------|
| MNIST   | 1,3,5   | 10752  | 7542   | 99.44±0.07        | 99.23±0.17       | 99.20±0.13       | 99.30±0.14       | 99.35±0.08       |
|         | 5,3,1   | 9659   | 8635   | 99.16±0.04        | 99.27±0.11       | 99.29±0.08       | 99.22±0.10       | 99.25±0.04       |
|         | 2,4,6   | 9771   | 7947   | 99.09±0.05        | 99.06±0.16       | 99.25±0.07       | 99.00±0.30       | 99.17±0.14       |

| Dataset | P, N, A | $\#PU$ | $\#AC$ | $\eta_p^u = 0.8$ | $\eta_p^a = 0.9$ | $\eta_n^u = 1$ | $\eta_n^a = 1.1$ | $\eta_a^u = 1.2$ |
|---------|---------|--------|--------|-------------------|------------------|------------------|------------------|------------------|
| MNIST   | 1,3,5   | 10752  | 7542   | 98.94±0.38        | 99.13±0.14       | 99.08±0.19       | 99.13±0.20       | 99.10±0.14       |
|         | 5,3,1   | 9659   | 8635   | 98.94±0.38        | 99.13±0.14       | 99.08±0.19       | 99.13±0.20       | 99.10±0.14       |
|         | 2,4,6   | 9771   | 7947   | 98.94±0.38        | 99.13±0.14       | 99.08±0.19       | 99.13±0.20       | 99.10±0.14       |

layers for SVHN and 2 convolutional layers and 2 fully-connected layers for all the MNIST datasets. We used Adadelt Zeiler [2012] for optimization and squared loss for experiments.

There are four contenders, consisting of binary PU learning approaches and multi-class PU learning approaches. We also report the classification accuracy on each class and identification accuracy for augmented classes. The details about the compared methods are described below.

**Binary PU Learning:** UPU [du Plessis et al. 2015] and NNPU [Kiryo et al. 2017] are state-of-the-art binary PU learning approaches. We use the logistic loss for UPU and the sigmoid loss for NNPU. To compare with two binary PU learning approaches fairly, we merge the unlabeled and augmented classes datasets as new unlabeled dataset for binary classifier. Then for the testing phase, the predicted class of classifier is given as $\text{argmax}\{f_p, f_n\}$ and the accuracy of classifier is calculated the same as multi-classes classification.

**Multi-class PU Learning:** MPU [Xu et al. 2017] and AREA [Shu et al. 2020] are state-of-the-art multi-class PU learning approaches. For comparing with two approaches fairly, the positive and unlabeled datasets were treated as two positive classes datasets and augmented classes dataset was treated as unlabeled dataset. In the testing phase, the predicted class is given as $\text{argmax}\{f_p, f_n, f_a\}$.

Table 1 reports the classification accuracy of each algorithm on three datasets. It is obvious that the proposed method achieves the best result of all the binary PU learning and multi-class PU learning.
we conduct the experiments on varying degrees of class distribution shift in the testing phase. Using
we study the robustness for inaccurate training class priors, which can be estimated with mixture
well as the standard deviation of each algorithm. Fig.2 reports the experimental results on SVNH
The similar trends are shown on identification accuracy. Table.2 reports the identification accuracy as
Table 4: Classification accuracy of proposed method for class distribution shifting on MNIST dataset.
\begin{table}[h]
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
\textbf{Dataset} & \textbf{P, N, A} & \textbf{#PU} & \textbf{#AC} & $\eta_p = 0.8$ & $\eta_p = 0.9$ & $\eta_p = 1$ & $\eta_p = 1.2$ & $\eta_p = 1.1$ \\
\hline
MNIST & 7,8,9 & 10109 & 7956 & 98.16±0.33 & 98.70±0.12 & 98.10±0.20 & 98.02±0.25 & 98.10±0.48 \\
& & 9876 & 8239 & 98.95±0.11 & 98.86±0.16 & 98.95±0.15 & 98.63±0.05 & 98.67±0.25 \\
& 3,4,5 & 9955 & 7439 & 99.56±0.12 & 99.70±0.09 & 99.54±0.25 & 99.56±0.20 & 99.61±0.08 \\
\hline
\end{tabular}
\end{table}

methods. Note that the performance of multi-class PU learning approaches MPU and AREA is better
than binary PU learning approaches, since the positive and augmented classes can be learned for
multi-class classifiers. The noise only exists in negative examples for learning multi-class classifier.
The similar trends are shown on identification accuracy. Table.2 reports the identification accuracy as
Moreover, UPUAC is significantly better than multi-class positive and unlabeled learning methods.

5.2 Robustness for Inaccurate Training Class Priors

In above section, we have assumed that the class priors are accessible at the time of training. Here,
we study the robustness for inaccurate training class priors, which can be estimated with mixture
proportion estimation. Without loss of generality, we conduct the experiments on varying degrees of
inaccuracies for class priors in the training phase. Let $\eta_n, \eta_p$ and $\eta_a$ be real number around 1,
$\theta_n^u = \eta_n^u \theta_n^p$, $\theta_p^u = \eta_p^u \theta_p^p$ and $\theta_a^u = \eta_a^u \theta_a^p$ be perturbed class priors. In this section, we draw data by
using $\theta_n^u$, $\theta_p^u$ and $\theta_a^u$ but train models by using $\theta_n^p$, $\theta_p^p$ and $\theta_a^p$ instead. Table 3 shows the classification
accuracy, where training class priors varies from 0.8 to 1.2 under ground-truth priors. We observer that
the proposed model is robust to inaccurate $\theta_n^p$, $\theta_p^p$ and $\theta_a^p$ in mild environment, which prevents
the performance degeneration of proposed approach from misspecified mixture proportions.

5.3 Handling Class Probabilities Shift

In this section, we investigate class distribution shift in the testing data. Without loss of generality,
we conduct the experiments on varying degrees of class distribution shift in the testing phase. Using
the similar setting in section 5.2, let $\eta_p$, $\eta_n$ and $\eta_a$ be real number around 1, $\pi_p^u = \eta_p \pi_p$, $\pi_n^u = \eta_n \pi_n$, and $\pi_a^u = \eta_a \pi_a$ be the testing class probabilities, and we report experimental results on MNIST by
training model using $\pi_p$, $\pi_n$ and $\pi_a$ but testing model using $\pi_p^u$, $\pi_n^u$ and $\pi_a^u$ instead. Table 4 reports
the performance for handling class distribution shifting in terms of classification accuracy. The results prove that our approach can also overcome the class distribution shift problem in the open
environment.

6 Conclusion

In this paper, we investigate the problem of learning from positive and unlabeled data with unobserved
augmented classes by exploiting augmented classes data. We propose an unbiased risk estimator for positive and unlabeled learning with augmented classes. Besides, we provide a theoretical analysis of estimation error bound, which certainly guarantees the estimator converges to the optimal solution. Experiments demonstrated the effectiveness of proposed methods. In the future, we will study multi-
positive and unlabeled data with unobserved augmented classes, which is common in real-world applications. Besides, an interesting future issue is to investigate the advanced method for PUAC without augmented classes data.
References

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In International Conference on Machine Learning, ICML, Stockholmsmässan, Stockholm, Sweden, volume 80, pages 461–470, 2018.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. Mach. Learn., 109(4):719–760, 2020.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. 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, Würzburg, Germany, volume 11907, pages 71–85, 2019.

Qing Da, Yang Yu, and Zhi-Hua Zhou. Learning with augmented class by exploiting unlabeled data. In Proceedings of Conference on Artificial Intelligence, AAAI Québec City, Québec, Canada, pages 1760–1766, 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In Annual Conference on Neural Information Processing Systems, NeurIPS, Montreal, Quebec, Canada, pages 703–711, 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In International Conference on Machine Learning, ICML, Lille, France, volume 37, pages 1386–1394, 2015.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. Mach. Learn., 106(4):463–492, 2017.

Jessica Finocchiaro, Rafael M. Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In Annual Conference on Neural Information Processing Systems, NeurIPS, Vancouver, BC, Canada, pages 10780–10790, 2019.

Lan-Zhe Guo, Zhenyu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In International Conference on Machine Learning, ICML, Virtual Event, volume 119, pages 3897–3906, 2020a.

Tianyu Guo, Chang Xu, Jiajun Huang, Yunhe Wang, Boxin Shi, Chao Xu, and Dacheng Tao. On positive-unlabeled classification in GAN. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA, pages 8382–8390, 2020b.

Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In Annual Conference on Neural Information Processing Systems, NeurIPS, Montréal, Canada, pages 5921–5932, 2018.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In Annual Conference on Neural Information Processing Systems, NeurIPS, Long Beach, CA, USA, pages 1675–1685, 2017.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In Georg Gottlob and Toby Walsh, editors, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, IJCAI, Acapulco, Mexico, pages 587–594. Morgan Kaufmann, 2003.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. 2018.
Fantine Mordelet and Jean-Philippe Vert. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit. Lett.*, 37:201–209, 2014.

Harikrishna Narasimhan, Harish G. Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning, ICML, Lille, France*, volume 37, pages 2398–2407, 2015.

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Annual Conference on Neural Information Processing Systems, NeurIPS, Barcelona, Spain*, pages 1199–1207, 2016.

Fabian Pedregosa, Francis R. Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *J. Mach. Learn. Res.*, 18:55:1–55:35, 2017.

Anh T. Pham, Raviv Raich, Xiaoli Z. Fern, and Jesús Pérez Arriaga. Multi-instance multi-label learning in the presence of novel class instances. In *International Conference on Machine Learning, ICML, Lille, France*, volume 37, pages 2427–2435, 2015.

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *J. Mach. Learn. Res.*, 17:14:1–14:45, 2016.

Harish G. Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning, ICML, New York City, NY, USA*, volume 48, pages 2052–2060, 2016.

Tomoya Sakai and Nobuyuki Shimizu. Covariate shift adaptation on learning from positive and unlabeled data. In *The Thirty-Third Conference on Artificial Intelligence, AAAI, Honolulu, Hawaii, USA*, pages 4838–4845, 2019.

Emanuele Sansone, Francesco G. B. De Natale, and Zhi-Hua Zhou. Efficient training for positive unlabeled learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2584–2598, 2019.

Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Comput.*, 33(5):1234–1268, 2021.

Senlin Shu, Zhuoyi Lin, Yan Yan, and Li Li. Learning from multi-class positive and unlabeled data. In *International Conference on Data Mining, ICDM, Sorrento, Italy*, pages 1256–1261. IEEE, 2020.

Xiu-Shen Wei, Han-Jia Ye, Xin Mu, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Multi-instance learning with emerging novel class. *IEEE Trans. Knowl. Data Eng.*, 33(5):2109–2120, 2021.

Yixing Xu, Chang Xu, Chao Xu, and Dacheng Tao. Multi-positive and unlabeled learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI, Melbourne, Australia*, pages 3182–3188. ijcai.org, 2017.

Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. abs/1212.5701, 2012.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5:1225–1251, 2004.

Yu-Jie Zhang, Peng Zhao, Lanjihong Ma, and Zhi-Hua Zhou. An unbiased risk estimator for learning with augmented classes. In *Annual Conference on Neural Information Processing Systems, NeurIPS, virtual*, 2020.