WHO IS YOUR RIGHT MIXUP PARTNER IN POSITIVE AND UNLABELED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Positive and Unlabeled (PU) learning targets inducing a binary classifier from weak training datasets of positive and unlabeled instances, which arise in many real-world applications. In this paper, we propose a novel PU learning method, namely Positive and unlabeled learning with Partially Positive Mixup (P^{3}Mix), which simultaneously benefits from instance augmentation and supervision correction with a heuristic mixup technique. To be specific, we take inspiration from the directional boundary deviation phenomenon observed in our preliminary experiments, where the learned PU boundary tends to deviate from the fully supervised boundary towards the positive side. For the unlabeled instances with ambiguous predictive results, we select their mixup partners from the positive instances around the learned PU boundary, so as to transform them into augmented instances near to the boundary yet with more precise supervision. Accordingly, those augmented instances may push the learned PU boundary towards the fully supervised boundary, thereby improving the classification performance. Comprehensive experimental results demonstrate the effectiveness of the heuristic mixup technique in PU learning and show that P^{3}Mix can consistently outperform the state-of-the-art PU learning methods.

1 INTRODUCTION

Positive and **U**nlabeled (**PU**) learning refers to a specific binary classification problem, where only a small number of positive training instances are manually annotated but all other instances are unlabeled (Liu et al., 2002). Such kind of datasets naturally arise in many significant real-world scenarios such as product recommendation (Hsieh et al., 2015), deceptive reviews detection (Ren et al., 2014), and medical diagnosis (Yang et al., 2012). For specific example, many diseases, *e.g.,* Alzheimer's disease, Amyotrophic Lateral Sclerosis, and Parkinson's disease, are very infrequent and with long latency, hence only few diagnosed patients are known but a much larger population of undiagnosed individuals may be either diseased or healthy. Treating the diagnosed ones as positive instances and the undiagnosed ones as unlabeled instances results in such PU datasets of medical diagnosis. To meet those practical demands, PU learning has drawn increasing interest from the machine learning community (Bekker & Davis, 2020).

Formally, let $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{0, 1\}$ be the feature representation and category label, respectively, where the positive instance is indicated by $y = 1$ and the negative one by $y = 0$. In the context of PU learning, the training dataset is composed of the sets of positive instances $\mathcal{P} = \{(\mathbf{x}_i, y_i = 1)\}_{i=1}^{n_p}$ and unlabeled instances $\mathcal{U} = \{\mathbf{x}_i\}_{i=n_p+1}^{n_p+n_u}$, where $\mathcal{U}$ contains both positive and negative instances. The target is to learn a binary classifier based on such weak training dataset $\mathcal{P} \cup \mathcal{U}$.

During the past decades, many PU learning methods have been proposed, where, naturally, the essential idea is to estimate the negative instances from the set of unlabeled instances $\mathcal{U}$. Generally, most of existing PU learning methods can be divided into two categories, termed as sample-selection methods and cost-sensitive methods. As the name suggests, the sample-selection methods mainly select reliable negative instances from $\mathcal{U}$ using various heuristic strategies, *e.g.,* Naïve Bayes (Liu et al., 2002), kNN (Zhang & Zuo, 2009), k-means (Chaudhari & Shevade, 2012), and reinforcement learning (Luo et al., 2021); and then apply supervised methods over positive and those reliable negative instances. In contrast, the cost-sensitive methods treat all unlabeled instances as corrupted negative ones, and correct the estimation bias of the objective by employing well-designed mislassi-

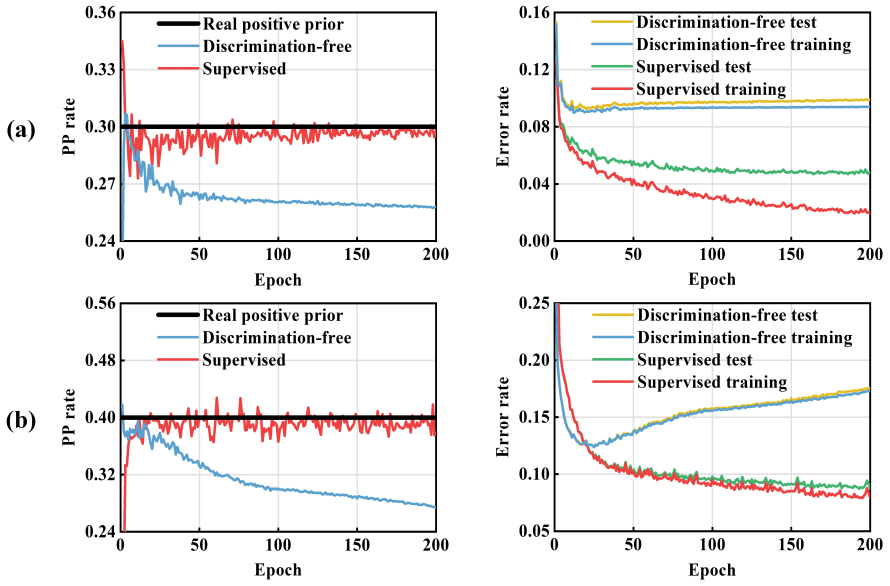

Figure 1: Rates of training instances Predicted as Positive (PP rate) and error rates of disambiguation-free[1] and fully supervised objectives on (a) F-MNIST and (b) CIFAR-10. The disambiguation-free objective suffers from much lower PP rates than the real positive prior as well as the fully supervised objective. This implies the directional boundary deviation phenomenon, which results in higher error rates.

fication risks such as unbiased risk estimator (du Plessis et al., 2014; 2015b; Kiryo et al., 2017) and maximum margin loss (Shi et al., 2018; Gong et al., 2019b; Zhang et al., 2019; Gong et al., 2019a).

Orthogonal to the aforementioned techniques, we note that some PU learning methods such as (Chen et al., 2020a; Wei et al., 2020) have made preliminary attempts to integrate with the art of mixup, *i.e.,* an economic-yet-effective data augmentation method (Zhang et al., 2018). Formally, mixup generates an augmented instance $(\widehat{\mathbf{x}}, \widehat{y})$ with the convex combination of any pair of instances $\{(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)\}$ drawn from the training dataset:

$$\widehat{\mathbf{x}} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \quad \widehat{y} = \lambda y_i + (1-\lambda) y_j, \qquad \lambda \sim \mathrm{Beta}(\alpha, \alpha),\ \alpha \in (0, \infty).$$

Previous studies have indicated that mixup is approximately equivalent to applying adversarial training (Zhang et al., 2021), enabling to improve robustness with even scarce and noisy supervision (Thulasidasan et al., 2019; Carratino et al., 2020; Zhang et al., 2021). Accordingly, it has been successfully used to solve for various learning problems with weak supervision, *e.g.,* semi-supervised learning (Berthelot et al., 2019), noisy label learning (Li et al., 2020), and partial label learning (Yan & Guo, 2020).

Our story and contribution. Inspired by the recent success of mixup in learning with weak supervision, our original goal is to thoroughly investigate the impact of mixup in PU learning. To this end, we begin with a naive *disambiguation-free* objective of PU learning, where all unlabeled instances are treated as *pseudo-negative instances*, denoted by $\widetilde{\mathcal{U}} = \{(\mathbf{x}_i, y_i = 0)\}_{i=n_p+1}^{n_p+n_u}$, and the binary classifier is trained based on $\mathcal{P} \cup \widetilde{\mathcal{U}}$. In preliminary experiments, we found an interesting phenomenon, where the number of training instances predicted as positive by the disambiguation-free classifier tends to be smaller than usual as illustrated in Fig.1. This phenomenon implies that the disambiguation-free boundary tends to deviate from the fully supervised boundary towards the positive side, expressed by a toy example shown in Fig.2(a). We consider that the directional boundary deviation is mainly caused by the *marginal pseudo-negative instances*, which lie between the two boundaries. Such instances are more likely to be positive but actually annotated by negative.

[1]Specially, we apply the early learning regularization (Liu et al., 2020) into the objective to keep it stable.

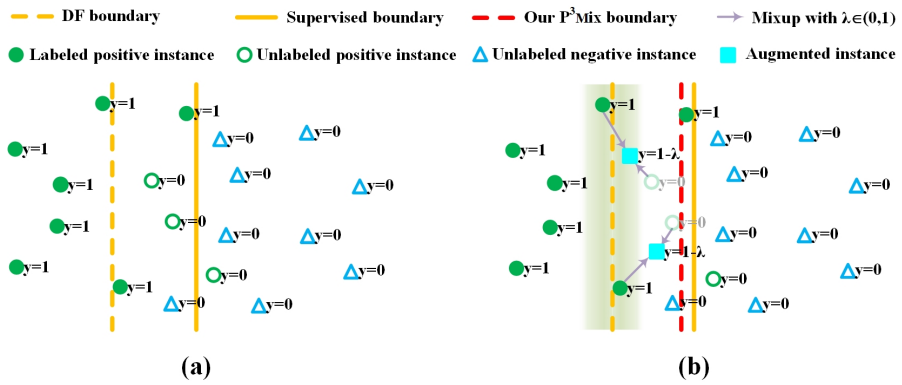

Figure 2: Toy examples of (a) the directional boundary deviation phenomenon and (b) the proposed heuristic mixup for marginal pseudo-negative instances. Best viewed in color.

Motivated by this observation, we extend mixup to a specific heuristic version for PU learning, enabling to achieve instance augmentation and supervision correction simultaneously. Its basic idea is to transform the marginal pseudo-negative instances into augmented instances which are partially positive and yet also lie between the two boundaries, so as to push the learned boundary towards the fully supervised one. This can be achieved by selecting the mixup partners for marginal pseudo-negative instances from the positive instances that are around the learned boundary, as expressed in Fig.2(b). With this insight, we propose a novel PU method, namely **P**ositive and unlabeled learning with **P**artially **P**ositive **Mix**up (**P^{3}Mix**). Generally, P^{3}Mix is easy-to-implement, where, specifically, we can define the marginal pseudo-negative instances using the predictive results and the positive instances around the boundary using the entropy values of predictive results. To evaluate the effectiveness of P^{3}Mix, we conduct a number of experiments on benchmark datasets. Experimental results demonstrate that P^{3}Mix can consistently outperform the state-of-the-art PU methods.

2 THE PROPOSED P^{3}MIX METHOD

In this section, we introduce the proposed **P^{3}Mix** method for PU learning. We first revisit and clarify some important notations: the set of positive instances $\mathcal{P} = \{(\mathbf{x}_i, y_i = 1)\}_{i=1}^{n_p}$ and the set of unlabeled instances $\mathcal{U} = \{\mathbf{x}_i\}_{i=n_p+1}^{n_p+n_u}$. By treating all unlabeled instances as negative, we translate $\mathcal{U}$ into the set of pseudo-negative instances $\widetilde{\mathcal{U}} = \{(\mathbf{x}_i, y_i = 0)\}_{i=n_p+1}^{n_p+n_u}$. Given batches $\mathcal{X}_p \subset \mathcal{P}$ and $\mathcal{X}_u \subset \widetilde{\mathcal{U}}$, the disambiguation-free objective of PU learning can be formulated as follows:

$$\mathcal{L}(\mathcal{X}_p, \mathcal{X}_u; \mathbf{\Theta}) = \frac{1}{|\mathcal{X}_p|} \sum_{(\mathbf{x},y)\in\mathcal{X}_p} \ell\big(f(\mathbf{x};\mathbf{\Theta}), y\big) + \frac{\beta}{|\mathcal{X}_u|} \sum_{(\mathbf{x},y)\in\mathcal{X}_u} \ell\big(f(\mathbf{x};\mathbf{\Theta}), y\big), \tag{1}$$

where $f(\cdot;\mathbf{\Theta})$ is a trainable neural network, *i.e.,* the binary classifier, parameterized by $\mathbf{\Theta}$; $\ell(\cdot,\cdot)$ is the loss function; and β is the coefficient parameter.

To achieve instance augmentation and supervision correction simultaneously, P^{3}Mix transforms $\mathcal{X}_p$ and $\mathcal{X}_u$ into the batches of augmented instances $\widehat{\mathcal{X}}_p$ and $\widehat{\mathcal{X}}_u$ using the proposed heuristic mixup technique. Accordingly, the objective of P^{3}Mix is then expressed as follows:

$$\mathcal{L}(\widehat{\mathcal{X}}_p, \widehat{\mathcal{X}}_u; \mathbf{\Theta}) = \frac{1}{|\widehat{\mathcal{X}}_p|} \sum_{(\widehat{\mathbf{x}},\widehat{y})\in\widehat{\mathcal{X}}_p} \ell\big(f(\widehat{\mathbf{x}};\mathbf{\Theta}), \widehat{y}\big) + \frac{\beta}{|\widehat{\mathcal{X}}_u|} \sum_{(\widehat{\mathbf{x}},\widehat{y})\in\widehat{\mathcal{X}}_u} \ell\big(f(\widehat{\mathbf{x}};\mathbf{\Theta}), \widehat{y}\big), \tag{2}$$

$$\widehat{\mathcal{X}}_p, \widehat{\mathcal{X}}_u = \text{HeuristicMixup}(\mathcal{X}_p, \mathcal{X}_u, \alpha), \tag{3}$$

where $\alpha \in (0, \infty)$ is a hyperparameter of mixup. Next, we describe the details of heuristic mixup.

2.1 TRAINING WITH HEURISTIC MIXUP

Basically, for each instance $(\mathbf{x}_i, y_i) \in \mathcal{X}_p \cup \mathcal{X}_u$ we select a mixup partner $(\mathbf{x}_j, y_j)$ to generate an augmented instance $(\widehat{\mathbf{x}}_i, \widehat{y}_i)$ using the modified mixup operator[2] (Berthelot et al., 2019):

$$\widehat{\mathbf{x}}_i = \lambda' \mathbf{x}_i + (1-\lambda')\mathbf{x}_j, \quad \widehat{y}_i = \lambda' y_i + (1-\lambda') y_j, \quad \lambda' = \max(\lambda, 1-\lambda), \\ \lambda \sim \mathrm{Beta}(\alpha, \alpha), \quad \alpha \in (0, \infty), \tag{4}$$

accordingly forming the augmented instance sets $\widehat{\mathcal{X}}_p$ and $\widehat{\mathcal{X}}_u$.

Our heuristic mixup refers to a guidance of mixup partners selection to refine the imprecise supervision within $\mathcal{X}_u$. We take inspiration from the phenomenon, where the boundary learned by Eq.(1) tends to deviate from the fully supervised boundary towards the positive side as illustrated in Fig.2(a). The marginal pseudo-negative instances $\mathcal{X}_{mpn} \subset \mathcal{X}_u$ lie between the two boundaries, and they are more likely to be positive but actually annotated by negative. To resolve this problem, for each of them we uniformly select a mixup partner from the candidate mixup pool $\mathcal{X}_{cnd} \subset \mathcal{P}$ of positive instances that are around the current learned boundary, so as to generate an augmented instance which is partially positive and yet also lies between the two boundaries as expressed in Fig.2(b). Besides, for positive instances $\mathcal{X}_p$ and other pseudo-negative instances $\mathcal{X}_u \setminus \mathcal{X}_{mpn}$, we uniformly choose their mixup partners from $\mathcal{X}_p \cup \mathcal{X}_u$ The overall mixup partners selection is formulated as follows:

$$(\mathbf{x}_j, y_j) \sim \begin{cases} \mathrm{Uniform}(\mathcal{X}_{cnd}) & \text{if } (\mathbf{x}_i, y_i) \in \mathcal{X}_{mpn}, \\ \mathrm{Uniform}(\mathcal{X}_p \cup \mathcal{X}_u) & \text{if } (\mathbf{x}_i, y_i) \in \mathcal{X}_p \cup \mathcal{X}_u \setminus \mathcal{X}_{mpn}. \end{cases} \tag{5}$$

In what follows, we introduce how to estimate the marginal pseudo-negative instances $\mathcal{X}_{mpn}$ and construct the candidate mixup pool $\mathcal{X}_{cnd}$.

Marginal pseudo-negative instance estimation. Because the fully supervised boundary is exactly unknown, we have to estimate the set of marginal pseudo-negative instances $\mathcal{X}_{mpn}$ from $\mathcal{X}_u$. In this work, we define them as the "unreliable" pseudo-negative instances measured by the predictive scores with thresholding parameter $\gamma \in [0.5, 1]$:

$$\mathcal{X}_{mpn} = \big\{(\mathbf{x}, y=0) | (\mathbf{x}, y=0) \in \mathcal{X}_u, 1-\gamma < f(\mathbf{x}; \mathbf{\Theta}) < \gamma\big\}, \tag{6}$$

where $\gamma = 0.5$ implies $\mathcal{X}_{mpn} = \emptyset$, and $\gamma = 1$ means $\mathcal{X}_{mpn} = \mathcal{X}_u$.

Candidate mixup pool. We maintain a candidate mixup pool $\mathcal{X}_{cnd}$ containing the positive instances around the current learned boundary from $\mathcal{P}$. To be specific, for each positive instance we compute its entropy value of the predictive score, and update the candidate mixup pool with the top-k positive instances as follows:

$$\mathcal{X}_{cnd} = \big\{(\mathbf{x}, y=1) | (\mathbf{x}, y=1) \in \mathcal{P}, \mathcal{H}(f(\mathbf{x}; \mathbf{\Theta})) \in \mathrm{Rank}(\{\mathcal{H}(f(\mathbf{x}_i; \mathbf{\Theta}))\}_{i=1}^{n_p})\big\}, \tag{7}$$

where $\mathcal{H}(\cdot)$ is the entropy, and $\mathrm{Rank}(\cdot)$ outputs a set of positive instances with the top-k maximum entropy values. For efficiency, we update $\mathcal{X}_{cnd}$ in every epoch.

2.2 ROBUSTNESS

The augmented instances within $\widehat{\mathcal{X}}_u$ also suffer from imprecise supervision even using the heuristic mixup. To make P^{3}Mix more robust, we employ two tricks, *i.e.,* early-learning regularization (Liu et al., 2020) and pseudo-negative instance correction.

[2]Because we compute individual loss terms for positive instances and pseudo-negative ones in Eq.(2) appropriately, we define $\lambda' = \max(\lambda, 1-\lambda)$ to guarantee that the feature of each augmented instance $\widehat{\mathbf{x}}_i$ is closer to $\mathbf{x}_i$ than the mixup partner $\mathbf{x}_j$.

Algorithm 1 Training procedure of P^{3}Mix

Input:
$\mathcal{P} \cup \mathcal{U}$: training instances; β: coefficient parameter; γ: thresholding parameter; k: size of the candidate mixup pool; α: hyperparameter of mixup
Output:
$\mathbf{\Theta}$: binary classifier parameters

Initialize $\mathbf{\Theta}$ and the candidate mixup pool $\mathcal{X}_{cnd}$ randomly;
for $t = 1, 2, \cdots, T$ **do**
  Shuffle $\mathcal{P} \cup \mathcal{U}$ into I mini-batches and denote the i-th mini-batch by $(\mathcal{X}_p^i, \mathcal{X}_u^i)$;
  for $i = 1, 2, \cdots, I$ **do**
    Estimate marginal pseudo-negative instances $\mathcal{X}_{mpn}$ using Eq.(6);
    Select the mixup partners for each instance within $\mathcal{X}_p^i \cup \mathcal{X}_u^i$ using Eq.(5);
    Construct $\widehat{\mathcal{X}}_p^i \cup \widehat{\mathcal{X}}_u^i$ by applying Eq.(4) to $\mathcal{X}_p^i \cup \mathcal{X}_u^i$ and their mixup partners;
    Update $\mathbf{\Theta}$ by its gradient $\nabla_{\mathbf{\Theta}} \mathcal{L}(\widehat{\mathcal{X}}_p^i, \widehat{\mathcal{X}}_u^i; \mathbf{\Theta})$ with the Adam optimizer;
  end for
  Update $\mathcal{X}_{cnd}$ using Eq.(7);
end for

Early-learning regularization. We employ the early learning regularization to prevent the memorization of imprecise supervision (Liu et al., 2020). For each mixup instance within $\widehat{\mathcal{X}}_p \cup \widehat{\mathcal{X}}_u$, we estimate an auxiliary target vector $\widetilde{\mathbf{y}}$, and formulate the early-learning regularization below:

$$\mathcal{R}_{elr}(\{(\widehat{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i)\}_{i=1}^{|\widehat{\mathcal{X}}_p|+|\widehat{\mathcal{X}}_u|}; \mathbf{\Theta}) = \frac{1}{|\widehat{\mathcal{X}}_p|+|\widehat{\mathcal{X}}_u|} \sum\nolimits_{i=1}^{|\widehat{\mathcal{X}}_p|+|\widehat{\mathcal{X}}_u|} \log\big(1 - \langle f(\widehat{\mathbf{x}}_i; \mathbf{\Theta}), \widetilde{\mathbf{y}}_i \rangle\big) \tag{8}$$

Here, we estimate the target vector $\widetilde{\mathbf{y}}$ of each instance by using the mean teacher technique (Tarvainen & Valpola, 2017), and incorporate Eq.(8) to Eq.(2).

Pseudo-negative instance correction. We concentrate on the pseudo-negative instances with high confidence to be positive $\big\{(\mathbf{x}, y = 0) | (\mathbf{x}, y = 0) \in \mathcal{X}_u, f(\mathbf{x}; \mathbf{\Theta}) \geq \gamma\big\}$. We directly revise their labels to positive before their corresponding mixup operators.

3 Experiment

3.1 Experimental Settings

Datasets. In the experiments, we employ three prevalent benchmark datasets, including Fashion-MNIST (F-MNIST),[3] CIFAR-10,[4] and STL-10.[5] The dataset statistics are described in Table 2. Note that all benchmark datasets have 10 category labels, and we denote them with integers ranging from 0 to 9 following the default settings in torchvision 0.10.0. For each dataset, we group those category labels into two disjoint sets as positive or negative, and generate two synthetic PU datasets by reversing the definitions of positive and negative labels. Following the protocol of (Chen et al., 2020a), the specific definitions of labels ("positive" vs "negative") are as follows: F-MNIST-1: "1,4,7" vs "0,2,3,4,5,6,8,9", F-MNIST-2: "0,2,3,5,6,8,9" vs "1,4,7"; CIFAR-10-1: "0,1,8,9" vs "2,3,4,5,6,7", CIFAR-10-2: "2,3,4,5,6,7" vs "0,1,8,9"; STL-10-1: "0,2,3,8,9" vs "1,4,5,6,7", STL-10-2: "1,4,5,6,7" vs "0,2,3,8,9". For each dataset, we randomly select 1,000 positive instances from the training set, and treat the remaining instances as unlabeled ones.

Baseline methods. To verify the effectiveness of P^{3}Mix, we utilize 6 PU baseline methods and also the supervised method for comparison. The details of baseline methods are presented below.

- **uPU** (du Plessis et al., 2014)

[3] https://github.com/zalandoresearch/fashion-mnist
[4] http://www.cs.toronto.edu/~kriz/cifar.html
[5] https://cs.stanford.edu/~acoates/stl10

Table 1: Specification of datasets and corresponding backbones.

Dataset	#Train	#Test	Input size	Backbone
F-MNIST	60,000	10,000	28×28	LeNet-5
CIFAR-10	50,000	10,000	3×32×32	7-layer CNN
STL-10	105,000	8,000	3×96×96	7-layer CNN

Table 2: Results of classification accuracy (mean±std). The highest scores among PU methods are indicated in **bold**.

Dataset	F-MNIST-1	F-MNIST-2	CIFAR-10-1	CIFAR-10-2	STL-10-1	STL-10-2
uPU	71.3±1.4	84.0±4.0	76.5±2.5	71.6±1.4	76.7±3.8	78.2±4.1
nnPU	89.7±0.8	88.8±0.9	84.7±2.4	83.7±0.6	77.1±4.5	80.4±2.7
Self-PU	90.8±0.4	89.1±0.7	85.1±0.8	83.9±2.6	78.5±1.1	80.8±2.1
PAN	88.7±1.2	83.6±2.5	87.0±0.3	82.8±1.0	77.7±2.5	79.8±1.4
VPU	90.6±1.2	86.8±0.8	86.8±1.2	82.5±1.1	78.4±1.1	82.9±0.7
PULNS	±	±	±	±	±	±
P^{3}Mix-E	**91.9±0.3**	**89.5±0.5**	88.2±0.4	84.7±0.5	80.2±0.9	**83.7±0.7**
P^{3}Mix-C	**92.0±0.4**	**89.4±0.3**	**88.7±0.4**	**85.9±0.8**	**80.7±0.7**	**83.6±0.6**
Supervised	95.2±0.2	95.2±0.2	91.3±0.3	91.3±0.3	85.6±0.6	85.6±0.6

- **nnPU**[6] (Kiryo et al., 2017)
- **Self-PU**[7] (Chen et al., 2020b)
- **PAN**[8] (Hu et al., 2021)
- **VPU**[9] (Chen et al., 2020a)
- **PULNS** (Luo et al., 2021)
- **Supervised**: The classifiers trained on the fully supervised datasets.

3.2 Classification performance

3.3 Ablation Study

3.4 Sensitivity Analysis

4 Related Work

In this section, we review the representative studies on PU learning, especially the ones most related to P^{3}Mix. Besides, we briefly introduce mixup for data augmentation.

4.1 PU Learning

Survey (Bekker & Davis, 2020)

Discriminative or generative methods

discriminative: sample selection approach, importance reweighting approach (Lee & Liu, 2003; Liu et al., 2003)

Identify reliable negative data (Liu et al., 2002; Yu et al., 2002; Li & Liu, 2003; Peng et al., 2008; Chaudhari & Shevade, 2012; Luo et al., 2021) MarginPU (Gong et al., 2018)

[6] https://github.com/kiryor/nnPUlearning
[7] https://github.com/TAMU-VITA/Self-PU
[8] https://github.com/morning-dews/PAN
[9] https://github.com/HC-Feynman/vpu

Figure 3: Test Error Rate, Precision and Recall on F-MNIST, CIFAR-10 and STL-10.

Figure 4: Number of Instances Predicted as Positive (abbreviated as NIPP) on F-MNIST, CIFAR-10 and STL-10.

cost-sensitive: RP (Northcutt et al., 2017)

unbiased risk estimator: uPU (du Plessis et al., 2014; 2015b), nnPU (Kiryo et al., 2017), Self-PU (Chen et al., 2020b), aaPU (Xu et al., 2019)

maximum margin classification problem, PULD (Zhang et al., 2019), LLSVM (Gong et al., 2019a), LDCE (Shi et al., 2018; Gong et al., 2019b)

VPU (Chen et al., 2020a)

Generative GAN for PU: GenPU (Hou et al., 2018), GanPU (Chiaroni et al., 2018) PAN (Hu et al., 2021), PUGAN (Guo et al., 2020)

Estimating prior (Jain et al., 2016; du Plessis et al., 2015a; 2017; Bekker & Davis, 2018; Yao et al., 2020)

application (Hsieh et al., 2015; Akujuobi et al., 2020)

Mixup: VPU and MixPUL (Chen et al., 2020a; Wei et al., 2020)

4.2 Mixup Augmentation

(Zhang et al., 2018; Verma et al., 2019; Kim et al., 2020; Yun et al., 2019; Hendrycks et al., 2020; Guo et al., 2019) (Thulasidasan et al., 2019; Zhang et al., 2021) (Pang et al., 2020; Chuang & Mroueh, 2021)

5 Conclusion

References

Uchenna Akujuobi, Jun Chen, Mohamed Elhoseiny, Michael Spranger, and Xiangliang Zhang. Temporal positive-unlabeled learning for biomedical hypothesis generation via risk estimation. In *Neural Information Processing Systems*, pp. 4597–4609. 2020.

Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *AAAI Conference on Artificial Intelligence*, pp. 2712–2719. 2018.

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

David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems*, pp. 5050–5060. 2019.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Sneha Chaudhari and Shirish K. Shevade. Learning from positive and unlabelled examples using maximum margin clustering. In *International Conference Neural Information Processing*, pp. 465–473. 2012.

Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *Neural Information Processing Systems*. 2020a.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-pu: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning*, pp. 1510–1519. 2020b.

Table 3: Results of ablative study (mean±std). The highest scores among PU methods are indicated in **bold**.

Dataset	F-MNIST-1	F-MNIST-2	CIFAR-10-1	CIFAR-10-2	STL-10-1	STL-10-2
DF	±	±	±	±	±	±
DF+mixup	±	±	±	±	±	±
P^{3}Mix	±	±	±	±	±	±
DF-E	90.1±0.7	74.2±5.5	82.4±1.6	69.4±3.0	67.3±2.0	75.0±3.7
DF-E+mixup	90.6±0.7	86.1±2.5	85.7±0.7	76.4±0.9	78.3±1.1	79.3±2.3
P^{3}Mix-E	**91.9±0.3**	**89.5±0.5**	**88.2±0.4**	**84.7±0.5**	**80.2±0.9**	**83.7±0.7**
DF-C	89.6±1.8	87.4±2.4	87.2±0.8	69.5±4.1	80.2±3.0	82.7±2.6
DF-C+mixup	91.6±0.3	88.3±1.2	87.7±1.1	69.1±2.4	79.9±3.1	81.6±2.8
P^{3}Mix-C	**92.0±0.4**	**89.4±0.3**	**88.7±0.4**	**85.9±0.8**	**80.7±0.7**	**83.6±0.6**

Florent Chiaroni, Mohamed-Cherif Rahal, Nicolas Hueber, and Frédéric Dufaux. Learning with A generative adversarial network from a positive unlabeled dataset for image classification. In *IEEE International Conference on Image Processing*, pp. 1368–1372. 2018.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*. 2021.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Neural Information Processing Systems*, pp. 703–711. 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pp. 221–236. 2015a.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pp. 1386–1394. 2015b.

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

Chen Gong, Tongliang Liu, Jian Yang, and Dacheng Tao. Large-margin label-calibrated support vector machines for positive and unlabeled learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3471–3483, 2019a.

Chen Gong, Hong Shi, Tongliang Liu, Chuang Zhang, Jian Yang, and Dacheng Tao. Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):918–932, 2019b.

Tieliang Gong, Guangtao Wang, Jieping Ye, Zongben Xu, and Ming Lin. Margin based PU learning. In *AAAI Conference on Artificial Intelligence*, pp. 3037–3044. 2018.

Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *AAAI Conference on Artificial Intelligence*, pp. 3714–3722. 2019.

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*, pp. 8382–8390. 2020.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*. 2020.

Ming Hou, Brahim Chaib-Draa, Chao Li, and Qibin Zhao. Generative adversarial positive-unlabeled learning. In *International Joint Conference on Artificial Intelligence*, pp. 2255–2261. 2018.

Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit S. Dhillon. PU learning for matrix completion. In *International Conference on Machine Learning*, pp. 2445–2453. 2015.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. Predictive adversarial learning from positive and unlabeled data. In *AAAI Conference on Artificial Intelligence*, pp. 7806–7814. 2021.

Shantanu Jain, Martha White, Michael W. Trosset, and Predrag Radivojac. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pp. 5275–5285. 2020.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Neural Information Processing Systems*, pp. 1675–1685. 2017.

Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *International Conference Machine Learning*, pp. 448–455. 2003.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*. 2020.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *International Joint Conference on Artificial Intelligence*, pp. 587–594. 2003.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *International Conference Machine Learning*, pp. 387–394. 2002.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *IEEE International Conference on Data Mining*, pp. 179–188, 2003.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Neural Information Processing Systems*. 2020.

Chuan Luo, Pu Zhao, Chen Chen, Bo Qiao, Chao Du, Hongyu Zhang, Wei Wu, Shaowei Cai, Bing He, Saravanakumar Rajmohan, and Qingwei Lin. PULNS: positive-unlabeled learning with effective negative sample selector. In *AAAI Conference on Artificial Intelligence*, pp. 8784–8792. 2021.

Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Conference on Uncertainty in Artificial Intelligence*. 2017.

Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. In *International Conference on Learning Representations*. 2020.

Tao Peng, Wanli Zuo, and Fengling He. SVM based adaptive learning method for text classification from positive and unlabeled documents. *Knowledge and Information Systems*, 16(3):281–301, 2008.

Yafeng Ren, Donghong Ji, and Hongbin Zhang. Positive unlabeled learning for deceptive reviews detection. In *Conference on Empirical Methods in Natural Language Processing*, pp. 488–498. 2014.

Hong Shi, Shaojun Pan, Jian Yang, and Chen Gong. Positive and unlabeled learning via loss decomposition and centroid estimation. In *International Joint Conference on Artificial Intelligence*, pp. 2689–2695. 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Neural Information Processing Systems*, pp. 1195–1204. 2017.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Neural Information Processing Systems*, pp. 13888–13899. 2019.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. 2019.

Tong Wei, Feng Shi, Hai Wang, Wei-Wei Tu, and Yu-Feng Li. Mixpul: Consistency-based augmentation for positive and unlabeled learning. *arXiv preprint arXiv:2004.09388*, 2020.

Miao Xu, Bingcong Li, Gang Niu, Bo Han, and Masashi Sugiyama. Revisiting sample selection approach to positive-unlabeled learning: Turning unlabeled data into positive rather than negative. *arXiv preprint arXiv:1901.10155*, 2019.

Yan Yan and Yuhong Guo. Partial label learning with batch label correction. In *AAAI Conference on Artificial Intelligence*, pp. 6575–6582. 2020.

Peng Yang, Xiaoli Li, Jian-Ping Mei, Chee Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Gang Niu, Masashi Sugiyama, and Dacheng Tao. Towards mixture proportion estimation without irreducibility. *arXiv preprint arXiv:2002.03673*, 2020.

Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. PEBL: positive example based learning for web page classification using SVM. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239–248. 2002.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision*, pp. 6022–6031. 2019.

Bangzuo Zhang and Wanli Zuo. Reliable negative extracting based on knn for learning from positive and unlabeled examples. *Journal of Computers*, 4(1):94–101, 2009.

Chuang Zhang, Dexin Ren, Tongliang Liu, Jian Yang, and Chen Gong. Positive and unlabeled learning with label disambiguation. In *International Joint Conference on Artificial Intelligence*, pp. 4250–4256. 2019.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*. 2018.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*. 2021.