Who Is Your Right Mixup Partner in Positive and Unlabeled Learning?

Changchun Li[1,2], Ximing Li[1,2], Lei Feng[3,4], Jihong Ouyang[1,2]

[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[3]College of Computer Science, Chongqing University, China
[4]Imperfect Information Learning Team, RIKEN Center for Advanced Intelligence Project, Japan

Background

- Positive and Unlabeled (PU) Learning

PU learning aims to induce a binary classifier from weak training datasets of positive and unlabeled instances.

- Mixup Technique

The mixup is approximately equivalent to applying adversarial training, enabling to improve robustness with even scarce and noisy supervision.

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

Motivation

- Disambiguation-free objective of PU learning

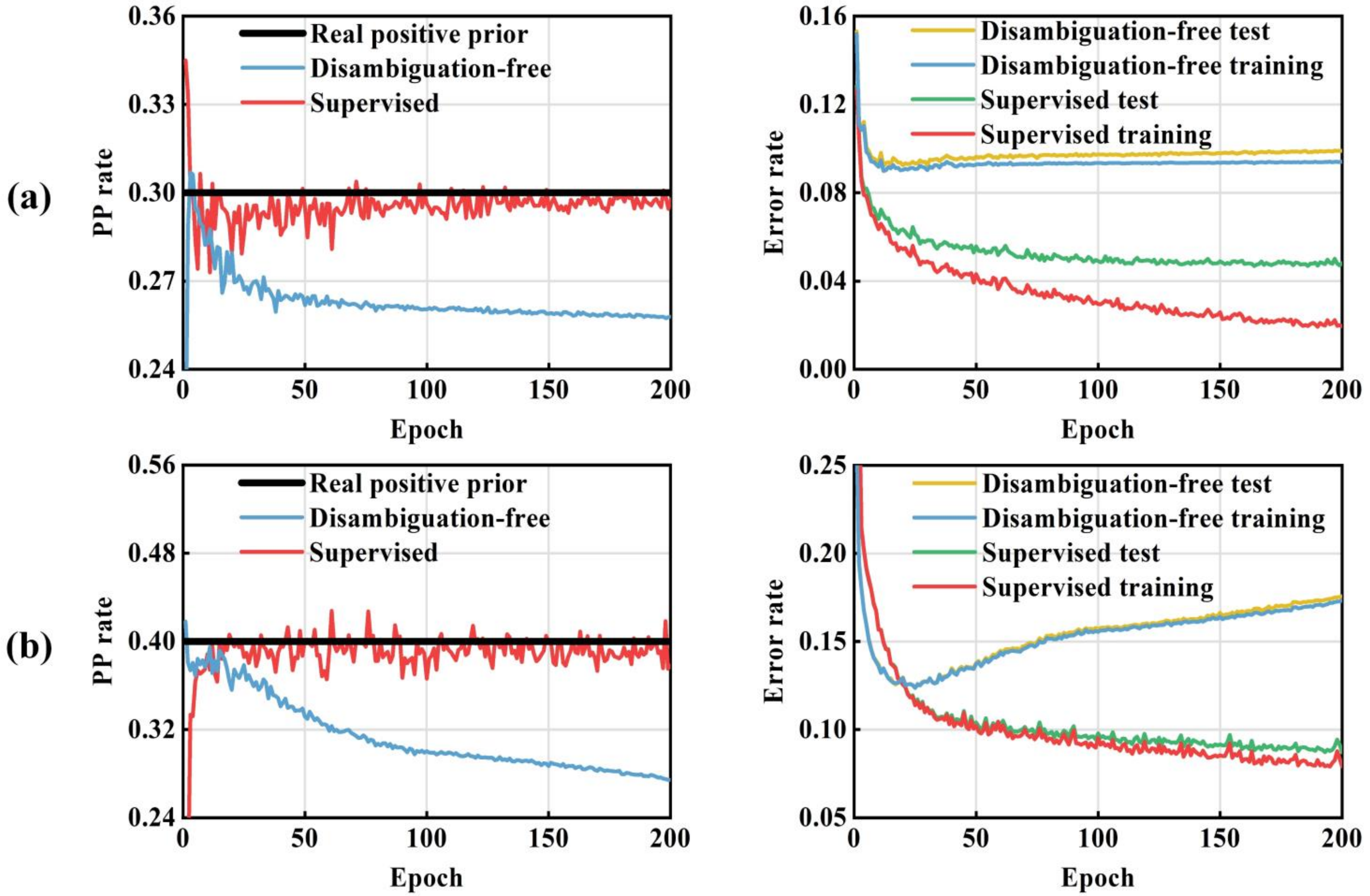
All unlabeled instances are treated as pseudo-negative instances, and the binary classifier is trained based on positive and pseudo-negative instances.

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

Motivation

- Decision boundary deviation phenomenon in PU learning

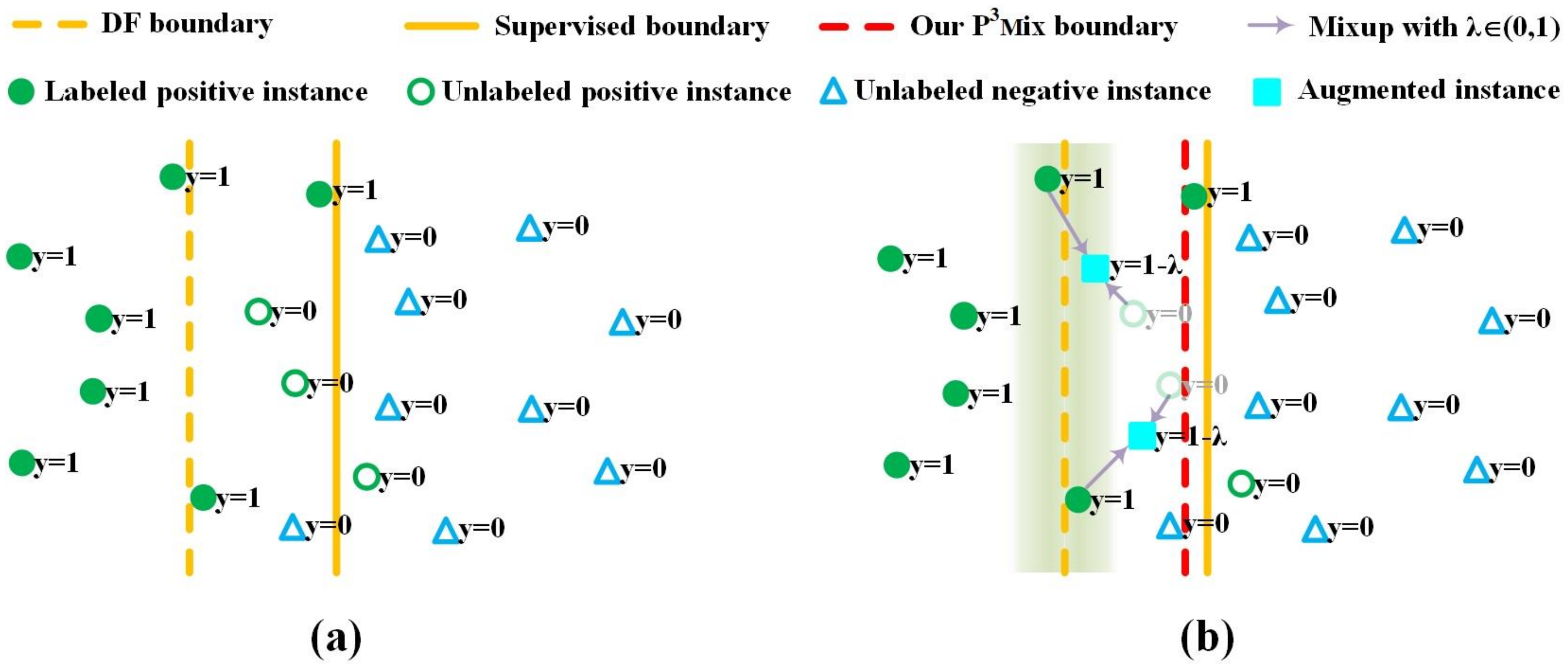
The number of training instances predicted as positive (PP rate) by the disambiguation-free classifier tends to be smaller than usual.



Motivation

- Decision boundary deviation phenomenon in PU learning

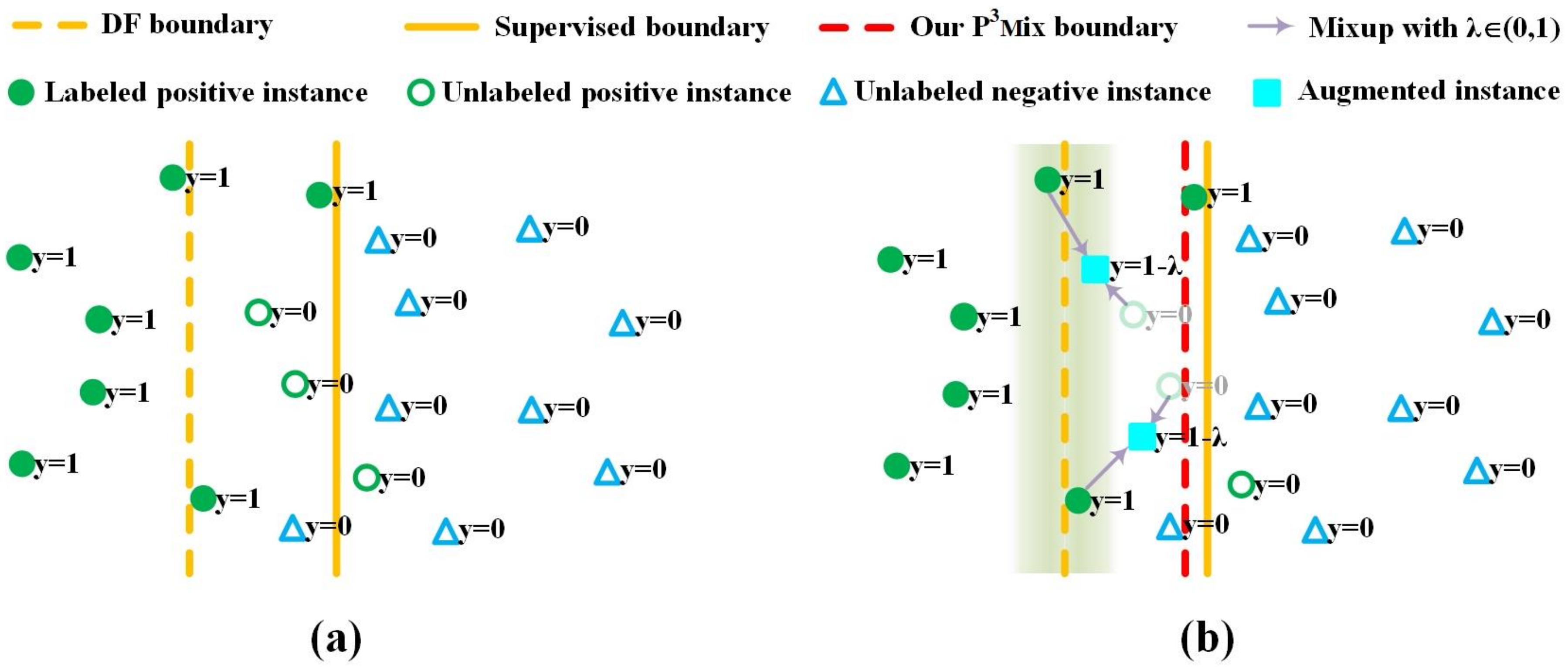
The disambiguation-free boundary tends to deviate from the fully supervised boundary towards the positive side.



Motivation

- Decision boundary deviation phenomenon in PU learning

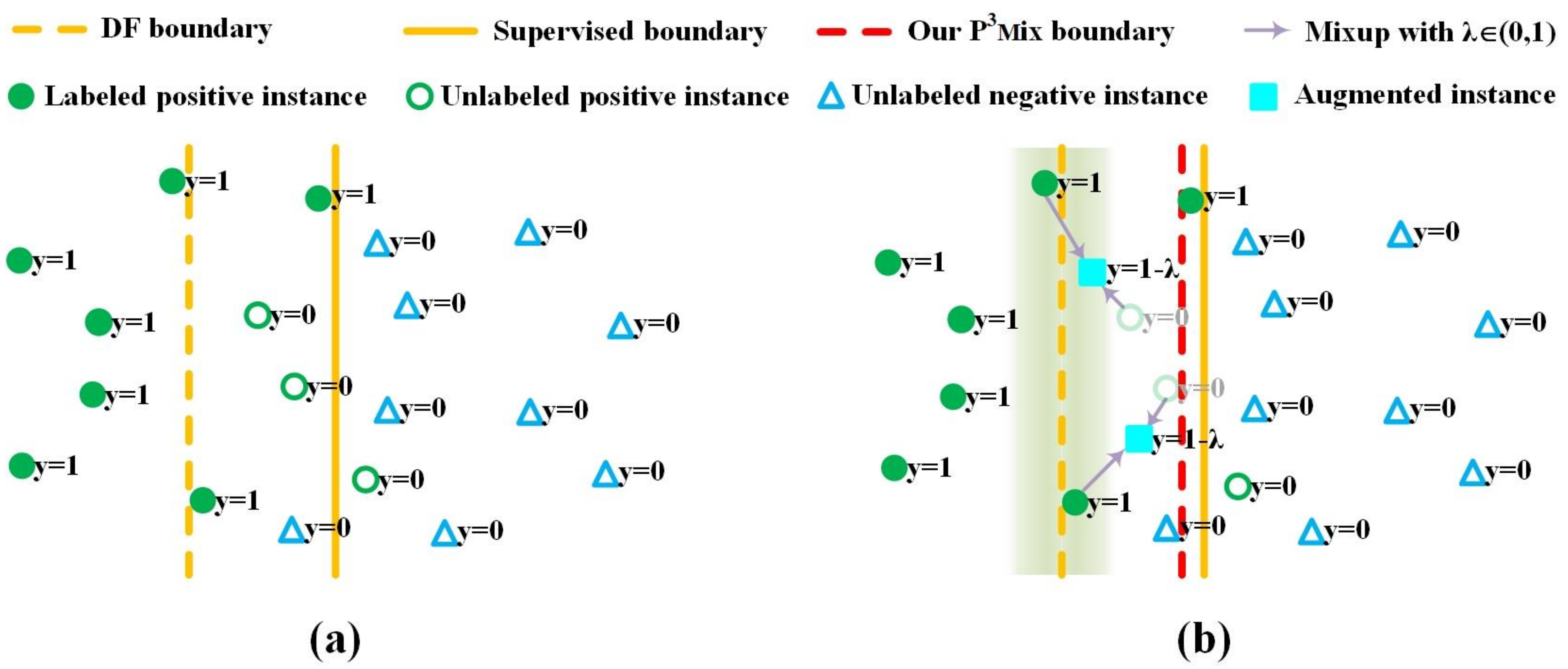
This is mainly caused by the marginal pseudo-negative instances, which lie between the two boundaries, and are more likely to be positive but actually annotated by negative.



Motivation

- Heuristic mixup for PU learning

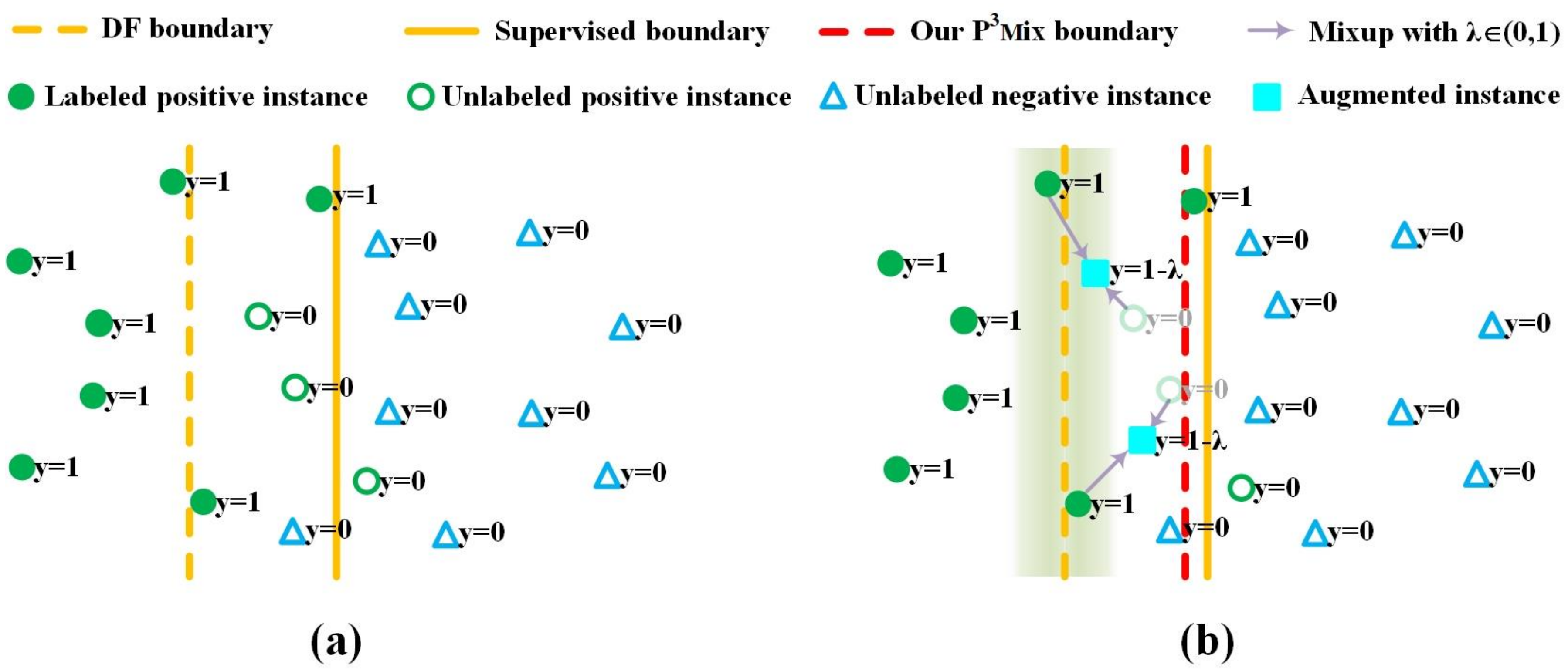
We extend mixup to a specific heuristic version for PU learning, enabling to achieve both data augmentation and supervision correction.



Motivation

- Heuristic mixup for PU learning

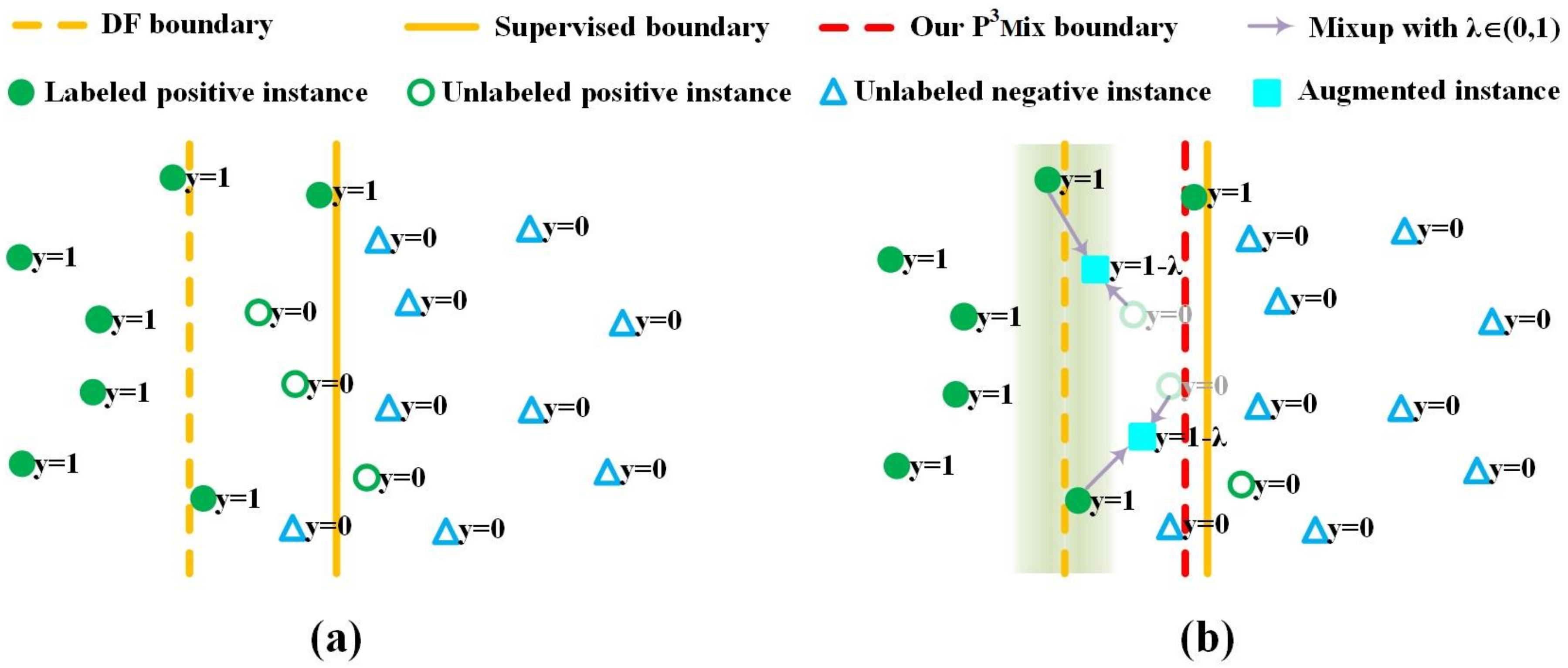
It transforms the marginal pseudo-negative instances into augmented instances which are partially positive and yet also lie between the two boundaries, so as to push the learned boundary towards the fully supervised one.



Motivation

- Heuristic mixup for PU learning

This can be achieved by selecting the mixup partners for marginal pseudo-negative instances from the positive instances that are around the learned boundary.



P^3Mix

- Objective of P^3Mix

P^3Mix transforms batches of positive instances $\mathcal{X}_p \subset \mathcal{P}$ and pseudo-negative ones $\mathcal{X}_u \subset \mathcal{U}$ into the batches of augmented instances $\mathcal{X}_p$ and $\mathcal{X}_u$ using the proposed heuristic mixup, and its objective is given by:

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

$$\widehat{\mathcal{X}}_p, \widehat{\mathcal{X}}_u = \mathrm{HeuristicMixup}(\mathcal{X}_p, \mathcal{X}_u, \alpha),$$

P³Mix

- Heuristic mixup

1. 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 $(\hat{\mathbf{x}}_i, \hat{y}_i)$ by using the modified mixup operator:

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

2. Select the mixup partner for each of marginal pseudo-negative instances $\mathcal{X}_{mpn} \subset \mathcal{X}_u$ from the candidate mixup pool $\mathcal{X}_{cnd} \subset \mathcal{P}$ of positive instances that are around the current learned boundary; Select the mixup partners for positive instances $\mathcal{X}_p$ and other pseudo-negative instances $\mathcal{X}_u \setminus \mathcal{X}_{mpn}$ from both positive and pseudo-negative ones $\mathcal{X}_p \cup \mathcal{X}_u$

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

P³Mix

- Marginal pseudo-negative instance estimation

We define the marginal pseudo-negative instances as the "unreliable" pseudo-negative instances measured by the predictive scores with thresholding $\gamma \in [0.5,1]$

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

- Candidate mixup pool

For each positive instance we compute its entropy value of the predictive score, and update the candidate mixup pool with the positive instances with the top-*k* maximum entropy values as follows:

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

Robustness of P^3Mix

- Early-learning regularization and P^3Mix-E

We employ the early-learning regularization to prevent the memorization of imprecise supervision.

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

- Pseudo-negative instance correction and P^3Mix-C

We focus on the pseudo-negative instances with high confidence to be positive:

$$\{(\mathbf{x}, y = 0) | (\mathbf{x}, y = 0) \in \mathcal{X}_u, f(\mathbf{x}; \mathbf{\Theta}) > \gamma\}$$

and directly revise their labels to positive before their corresponding mixup operators.

Experiment 1: Classification Accuracy

P^3Mix-E and P^3Mix-C consistently outperform all PU learning baselines on all benchmark datasets, indicating their superior performance.

Table 2: Results of classification accuracy (mean±std). The highest scores among PU learning 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
nnPU+mixup	91.4±0.3	88.2±0.7	87.2±0.6	85.8±1.2	79.8±0.8	82.2±0.9
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
MIXPUL	87.5±1.5	89.0±0.5	87.0±1.9	87.0±1.1	77.8±0.7	78.9±1.9
PULNS	90.7±0.5	87.9±0.5	87.2±0.6	83.7±2.9	80.2±0.8	83.6±0.7
P^3Mix-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^3Mix-C	**92.0±0.4**	**89.4±0.3**	**88.7±0.4**	**87.9±0.5**	**80.7±0.7**	**84.1±0.3**
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

Experiment 2: Ablation Study

1. The proposed heuristic mixup can significantly improve the classification performance.
2. Both early-learning regularization and pseudo-negative instance correction contribute to the improvement of the classification performance in all cases.

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

Dataset	F-MNIST-1	F-MNIST-2	CIFAR-10-1	CIFAR-10-2	STL-10-1	STL-10-2
DF	75.2±1.2	62.7±2.8	72.0±3.2	57.4±3.7	78.1±0.6	80.6±2.4
DF+mixup	78.4±1.7	72.4±1.4	79.2±3.0	67.4±2.5	78.9±0.3	80.7±1.9
P^3Mix	87.0±1.1	79.0±1.6	87.0±1.1	84.3±0.6	79.8±0.7	83.4±0.7
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^3Mix-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	84.7±1.1	80.2±3.0	82.7±2.6
DF-C+mixup	91.6±0.3	88.3±1.2	87.7±1.1	81.3±3.6	79.9±3.1	81.6±2.8
P^3Mix-C	**92.0±0.4**	**89.4±0.3**	**88.7±0.4**	**87.9±0.5**	**80.7±0.7**	**84.1±0.3**

Who Is Your Right Mixup Partner in Positive and Unlabeled Learning?

Thank you!

Changchun Li

changchunli93@gmail.com