Beyond Myopia: Learning from Positive and Unlabeled Data through Holistic Predictive Trends

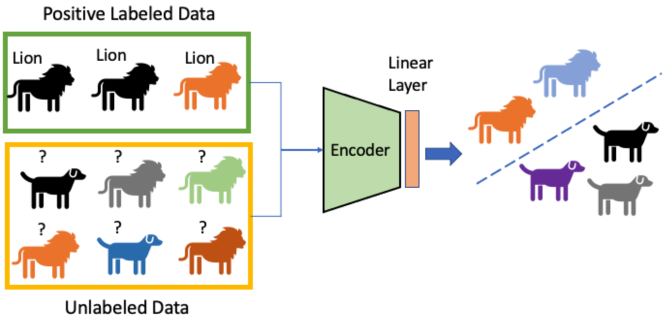
Xinrui Wang

2023/10/21

PAttern Recognition and NEural Computing

Background - Positive and Unlabeled Learning (PUL)

- Definition of PUL: A binary classification task with limited positive labeled data and a large amount of unlabeled data[1]
- Formalization: Positive set $\mathcal{X}_+ = \{x_i, y_i = 0\}_{i=1}^{n_+}$ & Unlabeled set $\mathcal{X}_u = \{x_i\}_{i=1}^{n_u}$

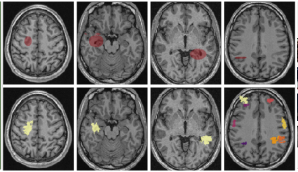


[1] Xiao-Li Li and Bing Liu. "Learning from positive and unlabeled examples with different data distributions". In: *European conference on machine learning*. Springer. 2005, pp. 218–229.

ParNeC

Background - Application

- Real-world applications: Matrix Completion[2], Deceptive Reviews Detection[3], Fraud Detection[4] & Medical Diagnosis[5].


[2] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. "PU learning for matrix completion". In: *International conference on machine learning*. PMLR. 2015, pp. 2445–2453.

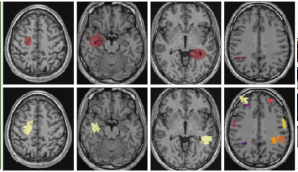
[3] Yafeng Ren, Donghong Ji, and Hongbin Zhang. "Positive unlabeled learning for deceptive reviews detection". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 488–498.

[4] Xiaoli Li, Bing Liu, and See-Kiong Ng. "Learning to Identify Unexpected Instances in the Test Set.". In: *IJCAI*. vol. 7. 2007, pp. 2802–2807.

[5] Peng Yang et al. "Positive-unlabeled learning for disease gene identification". In: *Bioinformatics* 28.20 (2012), pp. 2640–2647.

Background - Application

- Real-world applications: Matrix Completion[2], Deceptive Reviews Detection[3], Fraud Detection[4] & Medical Diagnosis[5].


- Serve as a basic component of more advanced ML problems.

[2] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. "PU learning for matrix completion". In: *International conference on machine learning*. PMLR. 2015, pp. 2445–2453.

[3] Yafeng Ren, Donghong Ji, and Hongbin Zhang. "Positive unlabeled learning for deceptive reviews detection". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 488–498.

[4] Xiaoli Li, Bing Liu, and See-Kiong Ng. "Learning to Identify Unexpected Instances in the Test Set.". In: *IJCAI*. vol. 7. 2007, pp. 2802–2807.

[5] Peng Yang et al. "Positive-unlabeled learning for disease gene identification". In: *Bioinformatics* 28.20 (2012), pp. 2640–2647.

Background - Method

- Starting from standard binary PN classification (ERM)

$$\hat{R}_{PN}(g) = \pi \hat{R}_p^+(g) + (1-\pi)\hat{R}_n^-(g) \tag{1}$$

where $\hat{R}_p^+(g) = \frac{1}{n_+} \sum_{x_i \in \mathcal{X}_+} l(g(x_i), +1), \hat{R}_n^-(g) = \frac{1}{n_-} \sum_{x_i \in \mathcal{X}_-} l(g(x_i), -1)$.

Background - Method

- Starting from standard binary PN classification (ERM)

$$\hat{R}_{PN}(g) = \pi \hat{R}_p^+(g) + (1-\pi)\hat{R}_n^-(g) \tag{1}$$

where $\hat{R}_p^+(g) = \frac{1}{n_+}\sum_{x_i \in \mathcal{X}_+} l(g(x_i), +1), \hat{R}_n^-(g) = \frac{1}{n_-}\sum_{x_i \in \mathcal{X}_-} l(g(x_i), -1)$.

- Since $R_u^-(g) = \pi R_p^-(g) + (1-\pi)R_n^-(g)$, an unbiased PU risk estimator:

$$\hat{R}_{PU}(g) = \pi \hat{R}_p^+(g) - \pi \hat{R}_p^-(g) + \hat{R}_u^-(g) \tag{2}$$

where $\hat{R}_p^-(g) = \frac{1}{n_+}\sum_{x_i \in \mathcal{X}_+} l(g(x_i), -1), \hat{R}_u^-(g) = \frac{1}{n_u}\sum_{x_i \in \mathcal{X}_u} l(g(x_i), -1)$.

Introduction

- Under certain assumptions on loss functions:

$$\hat{R}_{PU}(g) = 2\pi \hat{R}_p^{+}(g) + \hat{R}_u^{-}(g) - \pi \tag{3}$$

The unbiased risk estimator can be perceived as a reweighting or resampling based on positive prior π.

Introduction

- Assumption: The dataset consists of n i.i.d. samples from the following distributions:

$$\begin{aligned}\mathbb{P}(x|y=0) \sim \mathcal{N}(+v, \sigma^2 I_{p\times p}),\\ \mathbb{P}(x|y=1) \sim \mathcal{N}(-v, \sigma^2 I_{p\times p}).\end{aligned} \tag{4}$$

where v is an arbitrary unit vector in $\mathbb{R}^p$ and σ^2 is a small constant, radii $\sigma\sqrt{p} >> 2$ when $n, p \to \infty$ which makes this classification nontrivial.

Introduction

- Assumption: The dataset consists of n i.i.d. samples from the following distributions:
$$\begin{aligned} \mathbb{P}(x|y=0) &\sim \mathcal{N}(+v, \sigma^2 I_{p\times p}), \\ \mathbb{P}(x|y=1) &\sim \mathcal{N}(-v, \sigma^2 I_{p\times p}). \end{aligned} \tag{4}$$
where v is an arbitrary unit vector in $\mathbb{R}^p$ and σ^2 is a small constant, radii $\sigma\sqrt{p} >> 2$ when $n, p \to \infty$ which makes this classification nontrivial.
- Under the above Assumption, the Bayesian optimal decision hyperplane h_{pu} derived from an appropriate resampling strategy is equivalent to the Bayesian optimal decision hyperplane h^*_{pn}.
$$h_{pu} = h^*_{pn}. \tag{5}$$

Introduction

Typically, PUL methods can be divided into two main categories: cost-sensitive methods & sample-selection methods.

- The cost-sensitive methods rely on the negativity assumption[6], which may introduce estimation bias due to the mislabeling of positive examples as negative.
- The sample-selection methods struggle with distinguishing reliable negative examples[7], particularly during the initial stage, which also results in error accumulation during the training process.
- This bias can be accumulated and even worsen during later training stages, making its elimination challenging[8].

[6] Ryuichi Kiryo et al. "Positive-unlabeled learning with non-negative risk estimator". In: *Advances in neural information processing systems* 30 (2017).

[7] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. "PEBL: positive example based learning for web page classification using SVM". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 239–248.

[8] Daiki Tanaka, Daiki Ikami, and Kiyoharu Aizawa. *A Novel Perspective for Positive-Unlabeled Learning via Noisy Labels*. 2021. arXiv: 2103.04685.

Introduction

- To verify it, we make a simple pilot experiment:

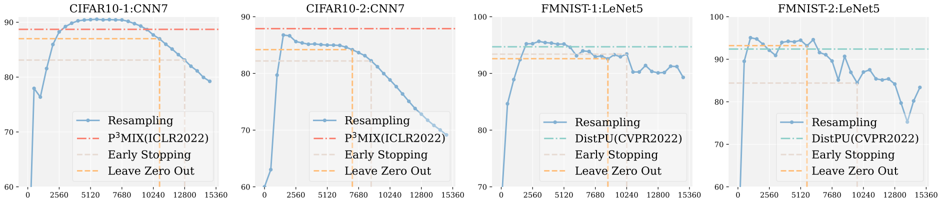
$$\mathcal{L} = \frac{1}{|\mathcal{X}_+|} \sum_{(x_i, y_i) \in \mathcal{X}_+} \ell(\hat{y}_i, y_i) + \frac{1}{|\mathcal{X}_u|} \sum_{x_i \in \mathcal{X}_u} \ell(\hat{y}_i, 1),\ \hat{y}_i = f(x_i). \tag{6}$$

Introduction

- To verify it, we make a simple pilot experiment:

$$\mathcal{L} = \frac{1}{|\mathcal{X}_+|} \sum_{(x_i, y_i) \in \mathcal{X}_+} \ell(\hat{y}_i, y_i) + \frac{1}{|\mathcal{X}_u|} \sum_{x_i \in \mathcal{X}_u} \ell(\hat{y}_i, 1), \ \hat{y}_i = f(x_i). \tag{6}$$

- Serious overfits occur:

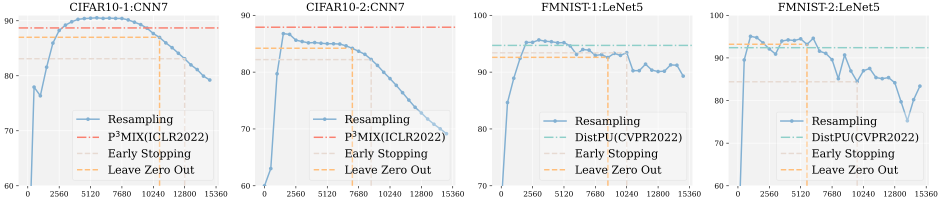


Introduction

- To verify it, we make a simple pilot experiment:

$$\mathcal{L} = \frac{1}{|\mathcal{X}_+|} \sum_{(x_i,y_i)\in\mathcal{X}_+} \ell(\hat{y}_i, y_i) + \frac{1}{|\mathcal{X}_u|} \sum_{x_i\in\mathcal{X}_u} \ell(\hat{y}_i, 1), \ \hat{y}_i = f(x_i). \tag{6}$$

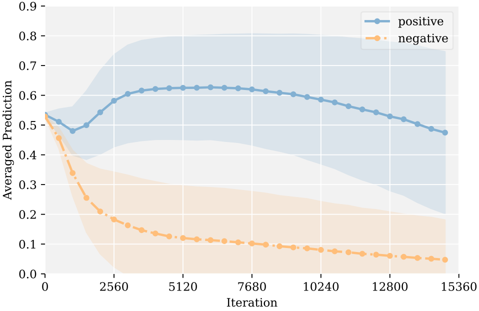
- Serious overfits occur:

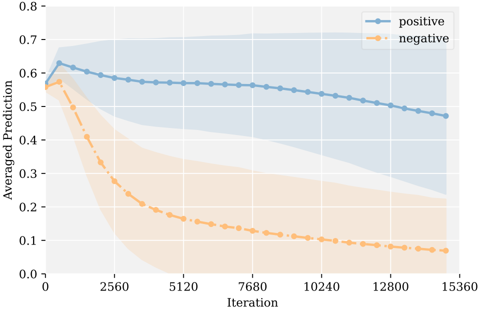


- From another perspective, as a basic component for various PUL methods, the resampling method shows its potential.

Introduction

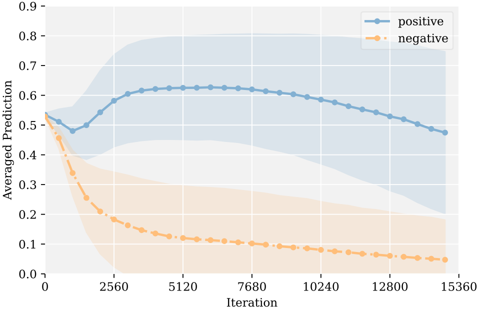
- Threshold selection on CIFAR10-1 & CIFAR10-2:

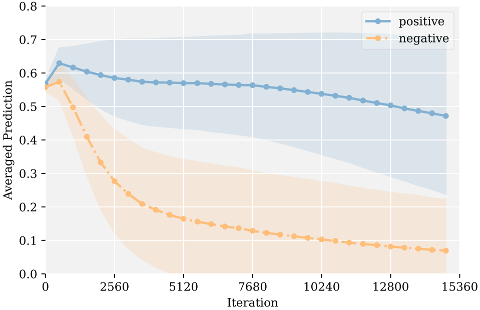




Introduction

- Threshold selection on CIFAR10-1 & CIFAR10-2:

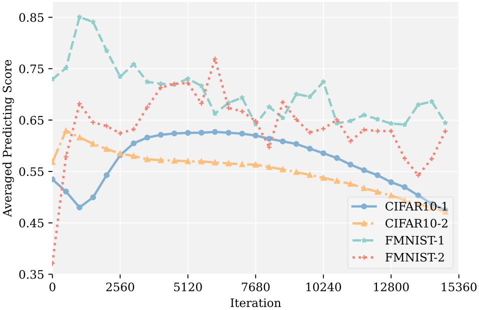


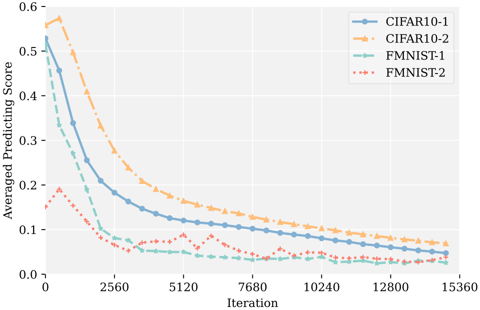


- Different from cost-sensitive methods & sample-selection methods relying on one single-step prediction that is prone to model uncertainty, we take a holistic view and examine the predictive trend of unlabeled data during the training process.

Observation

- Averaged predicting scores (output probability) of positive (left) and negative (right) examples in an unlabeled dataset during the first 30 epochs of training.

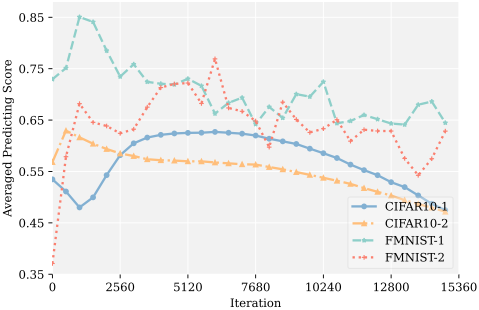


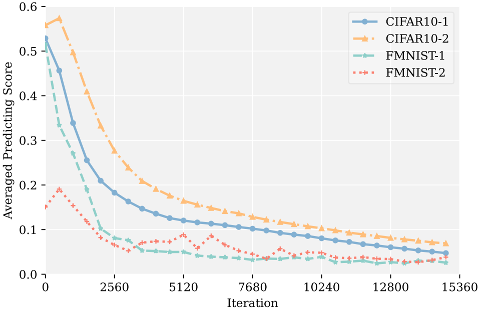


[9] Sheng Liu et al. "Early-learning regularization prevents memorization of noisy labels". In: *Advances in neural information processing systems* 33 (2020), pp. 20331–20342.

Observation

- Averaged predicting scores (output probability) of positive (left) and negative (right) examples in an unlabeled dataset during the first 30 epochs of training.





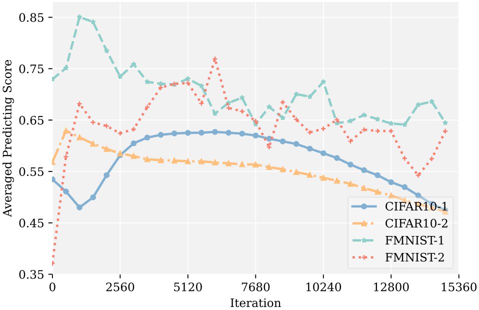
- The averaged predictive trends for different classes exhibit significant differences.

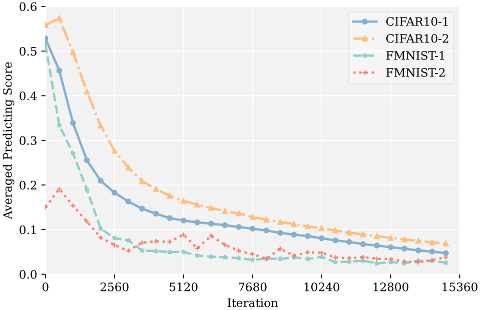
[9] Sheng Liu et al. "Early-learning regularization prevents memorization of noisy labels". In: *Advances in neural information processing systems* 33 (2020), pp. 20331–20342.

ParNeC

Observation

- Averaged predicting scores (output probability) of positive (left) and negative (right) examples in an unlabeled dataset during the first 30 epochs of training.





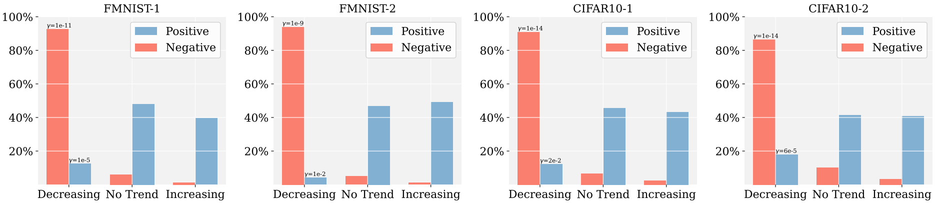
- The averaged predictive trends for different classes exhibit significant differences.
- Possible explanation: model's early focus on learning simpler patterns, which aligns with the early learning theory of noisy labels[9].

[9] Sheng Liu et al. "Early-learning regularization prevents memorization of noisy labels". In: *Advances in neural information processing systems* 33 (2020), pp. 20331–20342.

ParN_eC

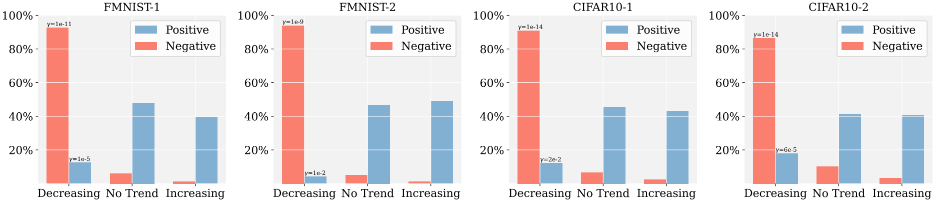
Identifying Predictive Trends

- We treat the prediction trend as a Temporal Point Process (TPP) and perform a Mann-Kendall Test to detect the predictive trends.



Identifying Predictive Trends

- We treat the prediction trend as a Temporal Point Process (TPP) and perform a Mann-Kendall Test to detect the predictive trends.



- Measure the differences between positive and negative examples through our proposed **trend score** S.

$$\hat{S} = \frac{2}{t(t-1)} \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} \psi(\alpha \Delta p_{ij}),\ \psi(\Delta p_{ij}) = sign(\Delta p_{ij}) \cdot log(1 + |\Delta p_{ij}| + \Delta p_{ij}^2/2) \tag{7}$$

Identifying Predictive Trends

Theorem: Let $P = \{p_{ij} | 1 \leq i \leq t-1, 2 \leq j \leq t, i < j\}$ be an observation set of changes in predictions in which $\mathbb{E}[\Delta p]$ is the expected values of the ordered difference in a temporal point process and σ^2 is the variance of P. By exploiting the non-decreasing influence function $\psi(\cdot)$, for any $\epsilon > 0$, we have the following bound with probability at least $1 - 2\epsilon$:

$$|\hat{S} - \alpha\mathbb{E}[\Delta p]| < \frac{2\alpha\sigma\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)}}}{1 - \sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)\alpha^2\sigma^2}}} = O\Big(\big(log(\epsilon^{-1})\big)^{\frac{1}{2}} t^{-1}\Big). \quad (8)$$

New Labeling Approach

- In the existing literature, threshold-based criteria and small loss criteria are the two primary approaches used for selecting reliable or clean examples.

New Labeling Approach

- In the existing literature, threshold-based criteria and small loss criteria are the two primary approaches used for selecting reliable or clean examples.
- Requiring extensive hyperparameter tuning efforts to choose appropriate thresholds or ratios for data selection.

New Labeling Approach

- In the existing literature, threshold-based criteria and small loss criteria are the two primary approaches used for selecting reliable or clean examples.
- Requiring extensive hyperparameter tuning efforts to choose appropriate thresholds or ratios for data selection.
- Clustering Unlabeled Data by the Fisher Criterion:

$$
\begin{aligned}
&\min_{C_1, C_2} \frac{\sum_{x \in C_1} (\hat{S}_x - \mu_1)^2}{|C_1|} + \frac{\sum_{x \in C_2} (\hat{S}_x - \mu_2)^2}{|C_2|} \\
&s.t.\ C_1 \cap C_2 = \emptyset,\ C_1 \cup C_2 = x_1, x_2, \ldots, x_N.
\end{aligned}
\tag{9}
$$

New Labeling Approach

- In the existing literature, threshold-based criteria and small loss criteria are the two primary approaches used for selecting reliable or clean examples.
- Requiring extensive hyperparameter tuning efforts to choose appropriate thresholds or ratios for data selection.
- Clustering Unlabeled Data by the Fisher Criterion:

$$\min_{C_1, C_2} \frac{\sum_{x \in C_1} (\hat{S}_x - \mu_1)^2}{|C_1|} + \frac{\sum_{x \in C_2} (\hat{S}_x - \mu_2)^2}{|C_2|} \tag{9}$$
$$s.t.\ C_1 \cap C_2 = \emptyset,\ C_1 \cup C_2 = x_1, x_2, \ldots, x_N.$$

- Once the unlabeled data is classified, the remaining task becomes a straightforward supervised learning problem.

Transductive Results

- Classification accuracy (Recall rate is reported on Credit Card):

Dataset	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	Credit Card	Alzheimer
nnPU	85.31	82.46	83.11	83.23	62.53	64.01
PGPU	92.02	90.17	85.67	88.38	42.12	75.09
Self-PU	94.04	91.59	84.06	83.77	71.00	70.05
P^3MIX-C	91.59	87.65	86.05	88.14	76.21	68.01
Ours	**95.41**	**96.00**	**91.42**	**91.17**	**98.90**	**75.13**

Transductive Results

- Classification accuracy (Recall rate is reported on Credit Card):

Dataset	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	Credit Card	Alzheimer
nnPU	85.31	82.46	83.11	83.23	62.53	64.01
PGPU	92.02	90.17	85.67	88.38	42.12	75.09
Self-PU	94.04	91.59	84.06	83.77	71.00	70.05
P^3MIX-C	91.59	87.65	86.05	88.14	76.21	68.01
Ours	**95.41**	**96.00**	**91.42**	**91.17**	**98.90**	**75.13**

- Positive prior estimation (Absolute error with the true positive prior):

Algorithm	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	STL10-1	STL10-2	Credit Card	Alzheimer
π	0.40	0.60	0.40	0.60	0.50	0.50	0.05	0.50
KM2	0.146	0.106	0.115	0.164	0.096	0.101	0.236	0.094
BBE*	0.082	0.073	0.034	0.059	0.046	0.064	0.112	0.026
$(\text{TED})^n$	0.026	0.020	0.042	0.044	0.024	0.021	0.018	0.014
Ours	**0.014**	**0.021**	**0.016**	**0.031**	**0.018**	**0.009**	**0.004**	**0.011**

Main Results

- Results of classification accuracy (%) on 3 generic datasets with 6 settings (mean±std):

Algorithm	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	STL10-1	STL10-2
uPU	81.6±1.2	85.7±2.6	76.5±2.5	71.6±1.4	76.7±3.8	78.2±4.1
nnPU	91.4±0.6	90.2±0.7	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	87.7±2.4	89.9±3.2	87.0±0.3	82.8±1.0	77.7±2.5	79.8±1.4
vPU	92.6±1.2	90.5±0.8	86.8±1.2	82.5±1.1	78.4±1.1	82.9±0.7
MIXPUL	90.4±1.2	89.6±1.2	87.0±1.9	87.0±1.1	77.8±0.7	78.9±1.9
PULNS	91.0±0.5	89.1±0.8	87.2±0.6	83.7±2.9	80.2±0.8	83.6±0.7
Dist-PU	94.7±0.4	92.4±0.4	86.8±0.7	87.2±0.9	79.8±0.6	82.9±0.4
P^3MIX-E	92.6±0.4	91.8±0.2	88.2±0.4	84.7±0.5	80.2±0.9	83.7±0.7
P^3MIX-C	92.8±0.6	90.4±0.1	88.7±0.4	87.9±0.5	80.7±0.7	84.1±0.3
Ours	**95.8±0.3**	**96.0±0.3**	**91.1±0.2**	**90.3±0.1**	**83.7±0.3**	**85.3±0.6**

Main Results

- Comparative results(%) on Credit Card Fraud dataset (mean±std):

Algorithm	F1 score	Recall	Accuracy	Precision	AUC
uPU	89.5±3.1	83.4±1.3	97.0±0.2	96.5±3.6	93.4±3.1
nnPU	89.9±1.0	83.4±1.3	98.4±0.1	97.4±1.1	94.2±0.9
nnPU+mixup	89.0±2.8	82.9±1.6	98.1±0.1	96.0±3.2	93.8±2.9
Self-PU	89.0±2.4	85.8±2.0	99.2±0.1	92.4±3.4	95.6±2.8
PAN	91.5±0.9	85.4±1.3	**99.1±0.1**	98.5±1.0	96.6±1.1
VPU	91.7±3.9	84.9±5.7	98.6±0.5	**99.7±0.6**	96.9±3.1
MIXPUL	82.9±2.8	86.6±1.3	98.4±0.3	79.2±3.5	91.3±0.7
PULNS	89.0±2.0	83.2±2.1	99.0±0.1	95.6±1.9	94.5±0.7
Dist-PU	87.9±3.4	80.2±4.1	98.8±0.4	97.2±1.6	96.5±2.7
P^3MIX-E	91.9±2.1	87.7±2.0	99.0±0.1	96.5±1.8	97.5±0.9
P^3MIX-C	90.2±1.4	86.5±1.8	98.8±0.1	94.1±1.2	97.3±1.2
Our Method	**99.1±0.2**	**99.0±0.2**	**99.1±0.1**	99.3±0.1	**99.7±0.1**

Main Results

- Comparative results(%) on Alzheimer dataset (mean±std):

Algorithm	F1 score	Recall	Accuracy	Precision	AUC
uPU	67.6±2.8	66.1±6.1	68.5±2.2	69.7±3.5	73.8±2.9
nnPU	68.6±3.2	69.5±7.2	68.3±2.1	68.0±2.3	72.9±2.8
RP	62.1±5.6	64.6±15.9	61.6±3.2	61.9±4.5	66.1±3.3
PUSB	69.2±2.4	69.3±2.4	69.2±2.4	69.2±2.4	74.4±2.4
PUbN	70.4±3.2	72.0±8.4	70.0±1.3	69.4±2.5	70.0±1.3
Self-PU	72.1±1.1	75.4±5.1	70.9±0.7	69.3±2.5	75.9±1.8
aPU	70.5±3.4	75.7±8.2	68.5±1.8	66.2±0.9	70.7±3.7
VPU	70.2±1.1	76.7±3.6	67.4±0.7	64.7±1.1	73.1±0.9
ImbPU	68.8±1.9	70.6±6.5	68.2±0.8	67.5±2.5	73.8±0.7
Dist-PU	73.7±1.6	**80.1±5.1**	71.6±0.6	68.5±1.2	**77.1±0.7**
Our Method	**74.5±2.4**	79.5±5.8	**72.8±0.9**	**70.2±1.6**	**77.1±2.3**

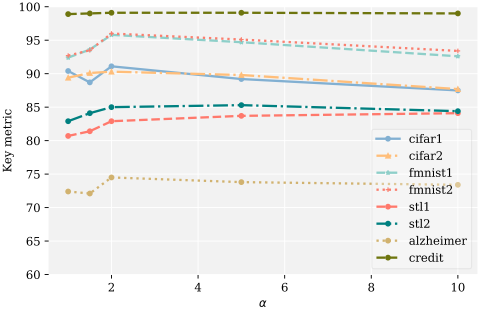
Ablation Study

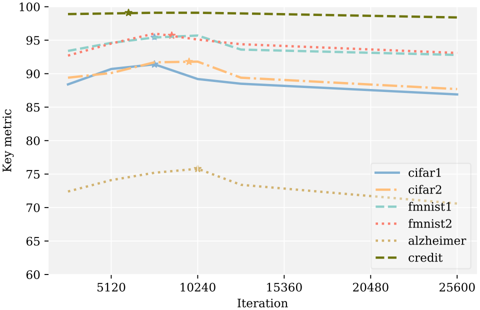
- Ablation results (%) on CIFAR-10 (acc), Credit Fraud (recall) and Alzheimer (f1 score). "✓" indicates the enabling of the corresponding components.

	Trend Measure			**Clustering**		**Dataset**		
Resampling	TS	Simplified TS	MK	Natural break	k-means	CIFAR10-1	Credit Fraud	Alzheimer
	✓			✓		84.1	88.6	69.2
✓	✓				✓	89.4	99.3	70.5
✓			✓	✓		90.2	99.0	69.7
✓		✓		✓		90.7	99.2	73.9
✓	✓			✓		91.1	99.1	74.5

Sensitivity Analysis

- Sensitivity analysis was performed on two parameters: α (left) and stopping iteration (right). The stopping iteration of LZO (also the one we use) is denoted by '$*$' on the right.





Future Works

- Similar concepts can be utilized to enhance out-of-distribution (OOD) data detection or semi-supervised learning.

Future Works

- Similar concepts can be utilized to enhance out-of-distribution (OOD) data detection or semi-supervised learning.
- When we look into this problem the majority of unlabeled data is positive or negative. It even makes PUL two completely different questions.

Method	$\pi = 0.124,\ \gamma = 1000$			$\pi = 0.712,\ \gamma = 10$			$\pi = 0.888,\ \gamma = 100$			$\pi = 0.960,\ \gamma = 1000$		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Resampling	92.05	96.41	91.45	74.13	82.32	42.10	70.40	79.45	35.31	67.24	71.90	14.11
ImbPU	**92.61**	97.12	92.51	83.22	**93.15**	86.11	74.12	84.58	77.25	71.27	80.31	65.47
Ours	92.52	96.60	**92.80**	**83.57**	90.84	**86.85**	**80.01**	**90.02**	**84.68**	**75.35**	**88.51**	**80.72**

Thank you!