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

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Learning binary classifiers from positive and unlabeled data (PUL) is vital in many real-world applications, especially when verifying negative examples is difficult. Despite the impressive empirical performance of recent PUL methods, challenges like accumulated errors and increased estimation bias persist due to the absence of negative labels. In this paper, we unveil an intriguing yet long-overlooked observation in PUL: *resampling the positive data in each training iteration to ensure a balanced distribution between positive and unlabeled examples results in strong early-stage performance. Furthermore, predictive trends for positive and negative classes display distinctly different patterns.* Specifically, the scores (output probability) of unlabeled negative examples consistently decrease, while those of unlabeled positive examples show largely chaotic trends. Instead of focusing on classification within individual time frames, we innovatively adopt a holistic approach, interpreting the scores of each example as a temporal point process (TPP). This reformulates the core problem of PUL as recognizing trends in these scores. We then propose a novel TPP-inspired measure for trend detection and prove its asymptotic unbiasedness in predicting changes. Notably, our method accomplishes PUL without requiring additional parameter tuning or prior assumptions, offering an alternative perspective for tackling this problem. Extensive experiments verify the superiority of our method, particularly in a highly imbalanced real-world setting, where it achieves improvements of up to 11.3% in key metrics.

1 Introduction

Positive and Unlabeled Learning (PUL) is a binary classification task that involves limited positive labeled data and a large amount of unlabeled data [36]. This learning scenario naturally arises in many real-world applications like matrix completion[25], deceptive reviews detection[45], fraud detection[35] and medical diagnosis[56]. It also serves as a key component of more complex machine learning problems, such as out-of-distribution detection[63] and adversarial training[18]. Two main categories of PUL methods are cost-sensitive methods and sample-selection methods. However, both approaches face their challenges. The cost-sensitive methods rely on the negativity assumption, which may introduce estimation bias due to the mislabeling of positive examples as negative[49]. This bias can be accumulated and even worsen during later training stages, making its elimination challenging. The sample-selection methods struggle with distinguishing reliable negative examples, particularly during the initial stage, which also results in error accumulation during the training process[23, 57].

As a basic component for various PUL methods, resampling the positive labeled data shows its potential in alleviating the bias brought by negative assumption [49, 52, 30, 33, 61]. For example, [30] resamples positive examples according to the given class prior and assumed label mechanism to achieve decent performance. In this paper, we dive deeper into this class of strategies. Instead of

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

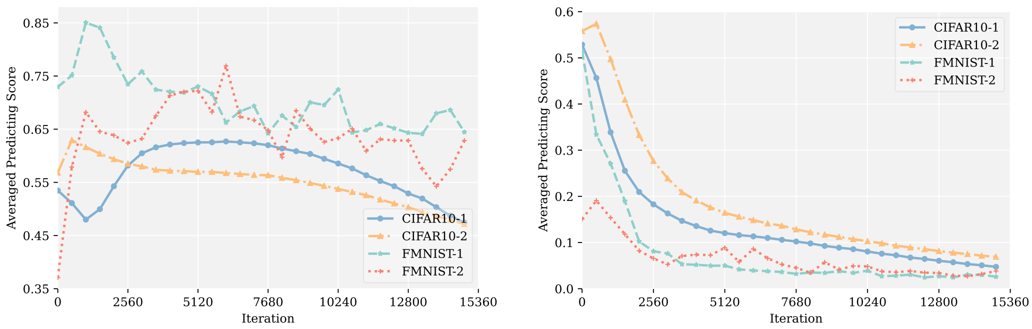

Figure 1: Averaged predicting scores (output probability) of positive (left) and negative (right) examples in an unlabeled dataset during the first 15,360 iterations of training (30 epochs).

relying on one single-step prediction which is prone to model uncertainty, we take a holistic view and examine the predictive trend of unlabeled data during the training process. Specifically, we treat the unlabeled data as negative. In each training epoch, we resample over the labeled positive data to ensure a balanced class distribution. We evaluate the model's performance on CIFAR10 and FMNIST datasets[32, 55] with 4 experimental settings. Our pilot experiments show that this resampling method achieves comparable or even state-of-the-art test performance at the outset, but underperforms soon after. Furthermore, the averaged predicting scores (output probability) of unlabeled negative examples exhibit a consistent decrease, whereas those of unlabeled positive examples display an initial increase before subsequent decreasing or oscillating. Conclusively, the averaged predictive trends for different classes exhibit significant differences, as depicted in Figure 1. One possible explanation for these observations is the model's early focus on learning simpler patterns, which aligns with the early learning theory of noisy labels [37]. Although the resampling strategy enjoys these advantages, selecting an appropriate model can be more challenging than the classification task itself due to the lack of a precise validation set.

To break the above limitation, we propose a novel approach that treats the predicting scores of each unlabeled training example as a temporal point process (TPP). It takes a holistic view and surpasses existing methods that focus on examining loss values or tuning confidence thresholds based on a limited history of predictions. By centering on the difference in trends of predicting scores, our approach provides a more comprehensive understanding of deep neural network training in PUL. To further investigate whether this difference in trends is prevalent in individual unlabeled examples, we apply the Mann-Kendall Test, a non-parametric statistical test used to detect trends in the temporal point process [20], to the continuously predicting scores of each example. These scores are classified into three types: *Decreasing*, *Increasing*, and *No Trend*. The statistical test reveals a clear distinction in the trends of predicted scores for each positive and negative example, supporting our observation. Our findings suggest that utilizing the model's classification ability in the early stages may be sufficient for successfully classifying unlabeled examples. This discovery offers us a new perspective on reformulating the problem of distinguishing positive and negative examples in the unlabeled set as identification of their corresponding predictive trends.

We then propose a novel TPP-inspired measure, called **trend score** to quantify the distinctions in predictive trends. It is obtained by applying a robust mean estimator [3] to the expected value of the ordered difference in a TPP (sequence of predicting scores for each example)[19]. Subsequently, we introduce a modified version of Fisher's Natural Break to distinguish these predictive trends, identifying a natural break point in the distribution of **trend score**. This approach divides examples into two groups: the group with **high trend score** represents positive examples, while the group with **low trend score** corresponds to negative examples. Our approach simplifies the training process by circumventing threshold selection when assigning pseudo-labels. Once the unlabeled data is classified, the remaining problem becomes a binary supervised learning task, and issues such as estimating class priors can be easily addressed. In summary, our main contributions are:

- We demonstrate the effectiveness of the proposed resampling strategy. It is also observed that predictive trends for each example can serve as an important metric for discriminating the categories of unlabeled data, providing a novel perspective for PUL.

- We propose a new measure, **trend score**, which is proved to be asymptotically unbiased in the change of predicting scores. We then introduce a modified version of Fisher's Natural Break with lower time complexity to identify statistically significant partitions. This process does not require additional tuning efforts and prior assumptions.
- We evaluate our proposed method with various state-of-the-art approaches to confirm its superiority. Our method also achieves a significant performance improvement in a highly imbalanced real-world setting.

2 Our Intuition and Method

2.1 Preliminary

We first revisit some important notations in PUL. Formally, let $x \in \mathbb{R}^d$ be the input data with d dimensions and $y \in \{0, 1\}$ be the corresponding label. Different from the traditional binary classification, PUL dataset is composed of a positive set $\mathcal{P} = \{x_i, y_i = 0\}_{i=1}^{n_p}$ and an unlabeled set $\mathcal{U} = \{x_i\}_{i=1}^{n_u}$, where the unlabeled set $\mathcal{U}$ contains both positive and negative data. Throughout the paper, we denote the positive class prior as $\pi = \mathbb{P}(y = 0)$.

2.2 Resampling Strategies for Positive and Unlabeled Learning

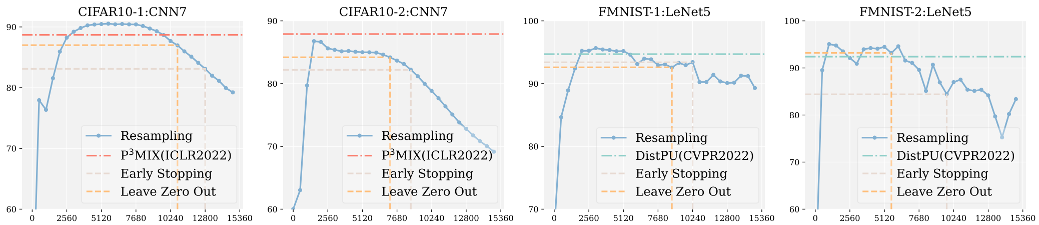

Figure 2: The accuracy of our resampling method (first 30 epochs). The horizontal line represents the accuracy of the state-of-the-art methods. Early stopping and Leave Zero Out represent different model selection strategies.

Resampling strategies have long been a baseline for dealing with imbalanced data or limited labels, which naturally fits PUL since its key challenge lies in limited labels and potentially imbalanced data distribution[5]. Different from popular resampling strategies applied in PUL[30], we follow the training scheme as [47, 58] to independently sample positive and unlabeled data as different data batches and the loss function is defined accordingly.

$$\mathcal{L} = \frac{1}{|\mathcal{B}p|} \sum_{(x_i, y_i) \in \mathcal{B}_p} \ell(\hat{y}_i, y_i) + \frac{1}{|\mathcal{B}u|} \sum_{x_i \in \mathcal{B}_u} \ell(\hat{y}_i, 1), \quad \hat{y}_i = f(x_i). \tag{1}$$

Here, we denote $f \in \mathcal{F}$ as a binary classifier, $\ell(\cdot, \cdot)$ as the loss function, $\mathcal{B}_p$ and $\mathcal{B}_u$ as the positive and unlabeled training batches respectively. We ensure that $|\mathcal{B}_p| = |\mathcal{B}_u|$ to achieve a balanced class prior during the training process. This approach emphasizes the labeled data and mitigates the imbalance of positive and pseudo-negative labels, which also provides a good theoretical explanation when dealing with high-dimensional data conforming to different Gaussian distributions. As shown in AppendixA.1, an optimal decision hyperplane can be attained when $|\mathcal{P}|/|\mathcal{U}|$ equals 1. Figure2 details the performance of our resampling baseline on two datasets under four different settings. It can be observed that the proposed method performs comparably or even better than state-of-the-art methods (P^{3}MIX[33] and DistPU[61]) in the early stages of training, as demonstrated by its test performance at certain epochs. However, the method's performance quickly degrades in all 4 settings as the estimation bias worsens during training due to the false negatives introduced by the negativity assumption. We also explore alternative model selection strategies, such as holding out a validation set from given labeled examples or using different versions of augmented data for model selection, as inspired by prior studies [34, 39]. In addition to the common practice of selecting the model from an additional positive validation set, we also implement LZO[34], which selects the model based on the mixup-induced validation set. As shown in Table1, the performance gap persists, especially when most of the unlabeled data belongs to the positive class.

Table 1: Classification accuracy (Recall rate is reported on Credit Card) on unlabeled training data. Resampling-P represents the model selected on an extra positive validation set. Resampling-LZO represents the model selected through LZO. Resampling* represents the best model selected on the test set which is an ideal case.

Dataset	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	STL10-1	STL10-2	Credit Card	Alzheimer
Resampling-P	89.93	84.29	81.06	72.93	-	-	60.75	70.09
Resampling-LZO	93.37	92.04	84.87	82.98	-	-	67.24	74.11
Resampling*	94.92	94.57	89.56	85.46	-	-	87.54	76.30
P^3MIX-C	91.59	87.65	86.05	88.14	-	-	76.21	68.01

To tackle the above issues, some denoising-based semi-supervised PUL methods, such as [8, 52, 49], have leveraged some threshold tuning or sample selection techniques to achieve acceptable empirical performance. These techniques have been criticized in [54] for relying solely on prediction scores or loss values, as they do not account for uncertainty in the selection process. This becomes even more problematic in PUL, where the noise ratio is typically higher when making a negativity assumption[2].

To break the above limitations, we record the whole predicting process of each unlabeled training example to take a holistic view of the training. It is evident that averaged model-predicting scores for positive and negative data display two distinct trends when implementing the above resampling strategy in the early training stages. Meanwhile, the standard deviation of predictions for positive examples increases rapidly during training, making it increasingly difficult to select an appropriate threshold for distinguishing between positive and negative examples. The appropriate threshold interval for discriminating positive and negative examples quickly shrinks as training progresses, indicating that existing denoising techniques cannot fundamentally alleviate the issues of accumulated errors and increased estimation bias. Therefore, a more robust evaluation measure is necessary beyond relying on raw model-predicted scores or loss values. Implementation details in model selection and visualizations of threshold tuning are provided in AppendixA.

2.3 Identifying Predictive Trends: A Key to Successful Classification

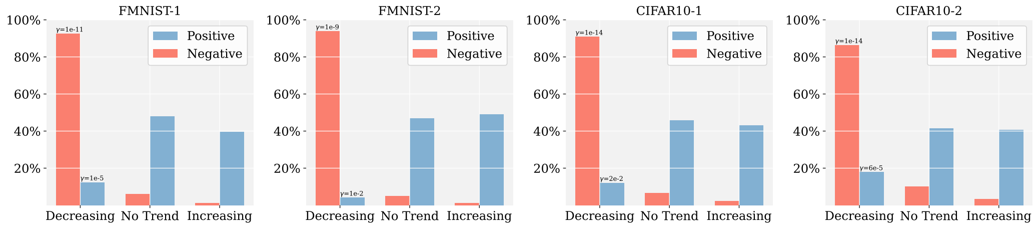

Figure 3: The Mann-Kendall Test is performed on 4 settings of CIFAR10 and FashionMnist datasets. The figure reports the fractions of positive and negative examples in an unlabeled dataset exhibiting different predictive trends during the early training stage (first 30 epochs).

While deep neural networks have strong learning capabilities, they are at risk of overfitting all provided labels, regardless of their correctness. This can result in all unlabeled examples being predicted as negative [1, 59]. We expect the predictive scores of negative examples in the unlabeled set to consistently decrease because all negative examples are given true negative labels by the negativity assumption. On the other hand, the predictive scores of positive examples in the unlabeled training set may not decrease initially because the resampled labeled examples are consistently emphasized from the start of training. To provide more evidence, we use the Mann-Kendall test to analyze the model-predicted scores of each example [20]. This test categorizes the prediction sequence into three situations: *Decreasing*, *Increasing*, and *No Trend*. The calculation process of the Mann-Kendall Test is detailed in AppendixB. Figure 3 shows a contrast between the trends of predicted scores for positive and negative examples. Even when certain positive and negative examples exhibit a similar trend of decreasing prediction scores during training, we observed significant differences in the significance index γ across different classes.

Our next objective is to measure the differences between positive and negative examples. To accomplish this, we require an evaluation measure that captures the significance of the observed trends in model-predicted scores. Before developing our own measure, an important notation in the TPP is first introduced, $\mathbb{E}[\Delta p]$, which represents the expected value of the ordered difference in a series of predicting scores.

$$\mathbb{E}[\Delta p] = \lim_{t\to\infty} \frac{2}{t(t-1)} \sum_{i<j}^{t} \Delta p_{ij},\ \Delta p_{ij} = p_j - p_i. \tag{2}$$

where p_i is the predicting score (output probability) at i-th epoch, t is the number of training epochs.

$$\tilde{S} = \frac{2}{t(t-1)} \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} \Delta p_{ij},\ \Delta p_{ij} = p_j - p_i. \tag{3}$$

While $\tilde{S}$ is the empirical mean and unbiased estimation of $\mathbb{E}[\Delta p]$, it can be unreliable for non-Gaussian examples and may not handle outliers or heavy-tailed data distributions well as illustrated in[3]. To address these issues, we propose a robust mean estimator inspired by[54, 20], called the **trend score** S, which measures the difference between each ordered pair of prediction scores:

$$\hat{S} = \frac{2}{t(t-1)} \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} \psi(\alpha\Delta p_{ij}),\ \Delta p_{ij} = p_j - p_i. \tag{4}$$

$$\psi(\Delta p_{ij}) = sign(\Delta p_{ij}) \cdot log(1 + |\Delta p_{ij}| + \Delta p_{ij}^2/2). \tag{5}$$

in which $\alpha > 0$ is a scaling parameter, and $sign()$ is the sign function that returns -1 if its argument is negative, 0 if its argument is zero, and 1 if its argument is positive. The function $\psi()$ can result in a more robust estimation by flattening the values of Δp_{ij} and reducing the influence of minority outlier points on the overall estimation. Besides, we also provide a simplified version as:

$$\dot{S} = \frac{1}{t-1} \sum_{i=1}^{t-1} \psi(\alpha\Delta p_{ij}),\ \Delta p_{ij} = p_j - p_i. \tag{6}$$

Notably, $\tilde{S}, \hat{S}, \dot{S}$ are all calculated on each example. Experiments show that both $\hat{S}, \dot{S}$ exhibit better empirical results than $\tilde{S}$ in Section3. For choosing the stopping epoch t, we implement the LZO[34] algorithm as described in Section2.2. We also derive a concentration inequality between our **trend score** $\hat{S}$ and the expected value of the ordered difference $\mathbb{E}[\Delta p]$.

Theorem 2.1. *Let $P = \{p_{ij} | 1 \le i \le t-1, 2 \le j \le t, i < j\}$ 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 $\psi()$, for any $\epsilon > 0$, we have the following bound with probability at least $1 - 2\epsilon$:*

$$|\hat{S} - \alpha\mathbb{E}[\Delta p]| < \frac{2\alpha\sigma\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)}}}{1 - \sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)\alpha^2\sigma^2}}} = O\Big(\big(log(\epsilon^{-1})\big)^{\frac{1}{2}} t^{-1}\Big). \tag{7}$$

It illustrates that the measure we propose is an asymptotically unbiased estimation with a linear weighting of $\mathbb{E}[\Delta p]$. The proof is provided in AppendixC. It is also proved in [3] that the deviations of this robust mean estimator can be of the same order as the deviations of the empirical mean computed from a Gaussian statistical sample, which further verifies the advantage of this estimator.

2.4 Clustering Unlabeled Data by the Fisher Criterion

The topic of accurately labeling unlabeled data is widely discussed in various fields, including PUL. In the existing literature, threshold-based criteria and small loss criteria are the two primary approaches used for selecting reliable or clean examples, as seen in studies such as [47, 58, 29, 49]. However, previous works generally select examples based solely on current predictions, ignoring the inherent uncertainty in training examples, leading to longer training times and poor generalization ability[54, 41]. Besides, they often require extensive hyperparameter tuning efforts to choose appropriate thresholds or ratios for data selection. In this section, we introduce a new labeling approach based on our proposed **trend score** tackling the above issues.

Our proposed **trend score** is the naturally comparable one-dimensional data and allows the Fisher Criterion to be a viable choice. It identifies a natural break point in the trend score distribution, which could be used to divide the data into two groups: one with high trend scores and one with low trend scores representing positive and negative examples respectively. Specifically, the objective function of finding this Fisher's natural break point can be formed as follows:

$$\begin{aligned} &\min_{C_1,C_2} \frac{\sum_{x\in C_1}(\hat{S}_x-\mu_1)^2}{|C_1|}+\frac{\sum_{x\in C_2}(\hat{S}_x-\mu_2)^2}{|C_2|} \\ &s.t.\ C_1\cap C_2=\emptyset,\ C_1\cup C_2=x_1,x_2,\ldots,x_N. \end{aligned} \tag{8}$$

where $\hat{S}_x$ is our derived **trend score** for example x, C_1 and C_2 are the two clusters, μ_i is the mean of cluster C_i, and N is the total number of data points. We utilize the Fisher natural break point method to automatically determine a threshold value that divided the trend score distribution into two distinct groups. Our implementation introduces an improved algorithm, which reduces the time complexity from $O(N^2)$ to $O(Nlog(N))$, as explained in AppendixD. This method eliminates the need for manual threshold selection or hyperparameter tuning, both of which can be time-consuming and error-prone. Furthermore, the data-driven approach we used optimizes the threshold value for the specific dataset under analysis, rather than relying on arbitrary or pre-defined values.

Once the unlabeled data is classified, the remaining task becomes a straightforward supervised learning problem. We directly train by a cross-entropy loss on the estimated labels given by Eq.8 on the backbone network given in Table4. Besides, issues such as estimating class priors can be addressed easily when unlabeled data are classified.

3 Experiments

3.1 Classification on Unlabeled Training Set

In this subsection, we first evaluate the performance of our method on the unlabeled training set compared with some state-of-the-art methods. As shown in Table2, our method demonstrates excellent classification performance on the unlabeled training data (the true labels of unlabeled data are not available in STL10). Moreover, a comparison with state-of-the-art prior estimation methods in PUL is conducted to further verify the effectiveness of our approach, and the results are presented in Table3.

Table 2: Classification accuracy (Recall rate is reported on Credit Card) on unlabeled training data.

Dataset	F-MNIST-1	F-MNIST-2	CIFAR10-1	CIFAR10-2	STL10-1	STL10-2	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^{3}MIX-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**

Table 3: Absolute estimation error with the true positive prior in the first row. We implement an oracle early stopping for the extant methods as defined in [15]. Our method significantly reduces estimation error when compared with existing methods.

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
(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**

3.2 Test Performance

We use three synthetic prevalent benchmark datasets including FashionMnist (F-MNIST) [55], CIFAR10 [32] and STL10 [10] and two real-world datasets on fraud detection[1] and Alzheimer diagnosis[2] as our test set. We provide the dataset description and corresponding backbones in Table4,

---

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[2] https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images

Table 4: Dataset description and corresponding backbones.

Dataset	#Trainset	#Testset	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
Alzheimer	5,890	1,279	3×224×224	ResNet-50
Credit Fraud	8,392	2098	30	6-Layer MLP

and the positive priors of each setting are given in Table3. More detailed description of benchmark datasets, dataset split and implementation details are given in AppendixF. For each dataset, we run our method for 5 times with different random seeds and report the averaged classification accuracy. We follow the settings in [52, 61] when making the comparison: randomly select 769 positive examples in Alzheimer dataset, 100 positive examples in Credit Fraud dataset and 1000 positive examples in others as the labeled set in training. Classification accuracy on test sets is reported as the main criterion. For highly imbalanced distributed (Credit Fraud) and biasedly selected (Alzheimer) datasets, we provide additional metrics such as Recall, F1 score and AUC on test sets for a more comprehensive comparison.

Table 5: 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**

Sythetic datasets. Our proposed method consistently outperforms all PUL baselines by 1% to 4% on all generic benchmark datasets and settings, as shown in Table 5, demonstrating its superior performance. Furthermore, many existing PUL methods rely on a given positive prior or make various assumptions that are not available in real-world settings, whereas our method does not require any of them. To avoid inherent challenges such as accumulated errors and estimation bias, we transform the above challenges into a much simpler task of discerning the trend of the model-predicting scores. Considering we can achieve outstanding classification accuracy in unlabeled data, it is natural to expect our method to outperform existing PUL methods. While using some tricks for label noise learning like Co-teaching[22] and large loss criterion[28] could possibly further improve the performance of our method, we believe that in most scenarios, our method can effectively solve existing PUL problems with simplicity.

Table 6: 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**

Table 7: 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**

Table 8: 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

Real-world datasets. This subsection presents experimental results on two real-world datasets, including one highly imbalanced Credit Fraud dataset. In fraud detection, recall is typically more important than precision or accuracy, as the consequences of missing a fraudulent transaction can be much more severe than flagging a legitimate transaction as fraudulent. As shown in Table 6, our proposed method achieves significantly higher recall rates and F1 scores, as well as comparable accuracy and precision, indicating its ability to better handle highly imbalanced scenarios. Our approach offers a novel perspective compared to traditional prediction-based methods, as the model's predictive trends are not affected by the positive prior, as long as the observation outlined in Section 2.3 holds. Furthermore, our method also demonstrates comparable performance on the Alzheimer dataset to the state-of-the-art method DistPU, which employs various regularization techniques and data augmentation strategies. In both two real-world settings, our method achieves a balanced good performance on all evaluation metrics which further illustrates its effectiveness.

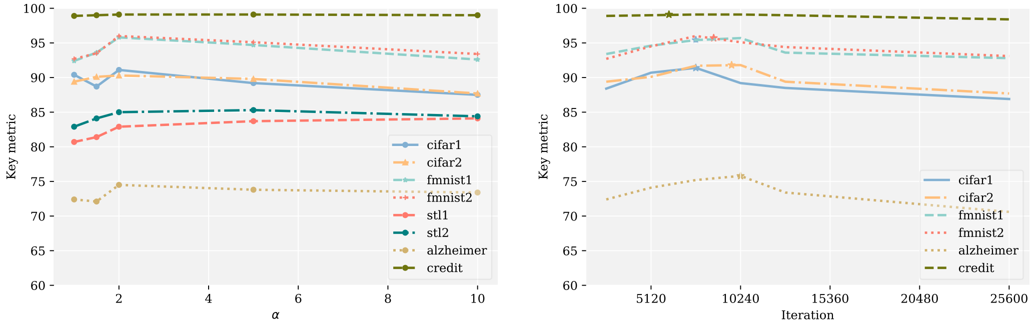

Figure 4: Sensitivity analysis was performed on two parameters: α (left) and stopping iteration (right). The stopping iteration of LZO (also the one we use) is denoted by '$*$' on the right.

Ablation Study. To investigate the specific effects of different components (Resampling, **trend score**, and Fisher Natural Break Partition) in our method, we conducted a series of ablation studies and compared them with some popular alternatives. From Table 8, we can draw several observations: (1) The resampling strategy plays a crucial role in our method as it maximizes the discrepancy of the trends in different classes of examples, particularly in the Credit Fraud dataset. It serves as an important factor in amplifying the model's early success, which is the foundation of our further approach towards achieving better performance. (2) Our proposed **trend score** provides a better evaluation metric than the statistic $\tilde{S}$ used in the standardized Mann-Kendall test, and the simplified **trend score** also shows competitive performance. (3) Fisher Natural Break Partition derives deterministic optimal partitions with better statistical properties and empirical performance compared to heuristic k-means. Moreover, it is unrelated to initialization and less time-consuming than the original version, as detailed in AppendixD.

Sensitivity Analysis. In this subsection, we investigate the impact of two hyperparameters, namely the scaling parameter α and the stopping iteration (we do not need to manually tune it), on the evaluation of predictive trends for each example. To facilitate comparisons, we set α to 2 and employ the LZO algorithm [34] discussed in Section 2.2 for selecting the stopping epoch in our experiments involving mixed labeled data. As depicted in Figure 4, our approach consistently delivers robust outcomes across diverse hyperparameter values. Moreover, the model tends to perform better when $\alpha > 1$ and demonstrates basically consistent performance. Figure 4 confirms the effectiveness of the LZO strategy which is free of manual intervention in the stopping epoch.

4 Related Works

For a long time, learning with limited supervision has been a striking task in the machine learning community and PUL is an emerging paradigm of weakly supervised learning [64, 17]. Despite its close relations with some similar concepts, the term PUL is generally accepted from [36, 12, 14]. Currently, the mainstream PUL methods cast this problem as a cost-sensitive classification task through importance reweighting, among which uPU [13] is the widely known one. Later, the authors of nnPU [31] suggest that uPU gets overfitting when using flexible and complex models such as Deep Neural Networks and thus propose a non-negative risk estimator. Some recent studies attempt to combine the cost-sensitive method with model's capability to calibrate and distill the labeled set with various techniques like denoise [49], self-paced curriculum [8] and heuristic mix up [33, 52].

Parallel with the cost-sensitive methods, another branch of PUL methods adopts a heuristic two-step method. The early trials of two-step methods mainly focus on the sample-selection task to form a reliable negative set and further yield the semi-supervised learning framework [57, 35, 23, 6, 27]. Other two-step methods are mainly derived from the large margin principle to correct the bias caused by unreliable negative data such as Loss Decomposition [46], Large margin based calibration and label disambiguation [16, 60]. Plus, different techniques have been employed to assign labels for unlabeled data in PUL like Graph-based models [4, 62], GAN [24, 27] and Reinforcement learning [38] in recent years. Plus, decision tree based PU methods are also investigated in [53].

Most PUL methods are oriented from a SCAR (selected completely at random) assumption or established on a given class prior. In this respect, there emerges some class prior estimation algorithms specially designed for PUL. PE attempts to minimize the Pearson divergence between the labeled and unlabeled distribution, PEN-L1 [9] and MPE [15] are then proposed to modify PE by using a simple Best Bin Estimation (BBE) technique. Unfortunately, most class prior estimation algorithms still rely on specific assumptions and the estimates will be unreliable otherwise[40]. Regarding the possibility of selection bias in the labeling process, the SCAR assumption is relaxed in [30]. VAE-PU is the first generative PUL model without a supposed labeling mechanism like SCAR assumption [42] and further investigated in [51]. For more details about PUL, readers are referred to a recent survey for a comprehensive understanding of this subject [2].

5 Conclusion

This study introduces a novel method for Positive-Unlabeled Learning (PUL) that takes a fresh perspective by identifying the unique characteristics of each example's predictive trend. Our approach is based on two key observations: Firstly, resampling positive examples to create a balanced training distribution can achieve comparable or even superior performance to existing state-of-the-art methods in the early stages of training. Secondly, the predicting scores of negative examples tend to exhibit a consistent decrease, while those of positive examples may initially increase before ultimately decreasing or oscillating. These insights lead us to reframe the central challenge of PUL as a task of discerning the trend of the model predicting scores. We also propose a novel labeling approach that uses statistical methods to identify significant partitions, circumventing the need for manual intervention in determining confidence thresholds or selecting ratios. Extensive empirical studies demonstrate the effectiveness of our method and its potential to contribute to related fields, such as learning from noisy labels and semi-supervised learning.

References

[1] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[2] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[3] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[4] S. Chaudhari and S. Shevade. Learning from positive and unlabelled examples using maximum margin clustering. In *International Conference on Neural Information Processing*, pages 465–473. Springer, 2012.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] H. Chen, F. Liu, Y. Wang, L. Zhao, and H. Wu. A variational approach for learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 33:14844–14854, 2020.

[7] P. Chen, X. Jin, X. Li, and L. Xu. A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2):5523–5544, 2021.

[8] X. Chen, W. Chen, T. Chen, Y. Yuan, C. Gong, K. Chen, and Z. Wang. Self-pu: Self boosted and calibrated positive-unlabeled training. In *International Conference on Machine Learning*, pages 1510–1519. PMLR, 2020.

[9] M. Christoffel, G. Niu, and M. Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pages 221–236. PMLR, 2016.

[10] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[11] O. Coudray, C. Keribin, P. Massart, and P. Pamphile. Risk bounds for positive-unlabeled learning under the selected at random assumption. *Journal of Machine Learning Research*, 24(107):1–31, 2023.

[12] F. Denis, R. Gilleron, and F. Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.

[13] M. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR, 2015.

[14] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[15] S. Garg, Y. Wu, A. J. Smola, S. Balakrishnan, and Z. Lipton. Mixture proportion estimation and pu learning: A modern approach. *Advances in Neural Information Processing Systems*, 34:8532–8544, 2021.

[16] C. Gong, T. Liu, J. Yang, and D. 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, 2019.

[17] C. Gong, J. Yang, J. You, and M. Sugiyama. Centroid estimation with guaranteed efficiency: A general framework for weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2841–2855, 2020.

[18] T. Guo, C. Xu, J. Huang, Y. Wang, B. Shi, C. Xu, and D. Tao. On positive-unlabeled classification in gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8385–8393, 2020.

[19] K. H. Hamed. Trend detection in hydrologic data: the mann–kendall trend test under the scaling hypothesis. *Journal of hydrology*, 349(3-4):350–363, 2008.

[20] K. H. Hamed and A. R. Rao. A modified mann-kendall trend test for autocorrelated data. *Journal of hydrology*, 204(1-4):182–196, 1998.

[21] Z. Hammoudeh and D. Lowd. Learning from positive and unlabeled data with arbitrary positive shift. *Advances in Neural Information Processing Systems*, 33:13088–13099, 2020.

[22] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[23] F. He, T. Liu, G. I. Webb, and D. Tao. Instance-dependent pu learning by bayesian optimal relabeling, 2018.

[24] M. Hou, B. Chaib-Draa, C. Li, and Q. Zhao. Generative adversarial positive-unlabelled learning, 2017.

[25] C.-J. Hsieh, N. Natarajan, and I. Dhillon. Pu learning for matrix completion. In *International conference on machine learning*, pages 2445–2453. PMLR, 2015.

[26] Y.-G. Hsieh, G. Niu, and M. Sugiyama. Classification from positive, unlabeled and biased negative data. In *International Conference on Machine Learning*, pages 2820–2829. PMLR, 2019.

[27] W. Hu, R. Le, B. Liu, F. Ji, J. Ma, D. Zhao, and R. Yan. Predictive adversarial learning from positive and unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7806–7814, 2021.

[28] J. Huang, L. Qu, R. Jia, and B. Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334, 2019.

[29] L. Junnan and S. C. Hoi. Dividemix: Learning with noisy labelsas semi-supervised learning. In *ICLR. International Conference on Learning Representations (ICLR)*, 2020.

[30] M. Kato, T. Teshima, and J. Honda. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*, 2019.

[31] R. Kiryo, G. Niu, M. C. Du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

[32] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[33] C. Li, X. Li, L. Feng, and J. Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *International Conference on Learning Representations*, 2022.

[34] W. Li, C. Geng, and S. Chen. Leave zero out: Towards a no-cross-validation approach for model selection, 2020.

[35] X. Li, B. Liu, and S.-K. Ng. Learning to identify unexpected instances in the test set. In *IJCAI*, volume 7, pages 2802–2807, 2007.

[36] X.-L. Li and B. Liu. Learning from positive and unlabeled examples with different data distributions. In *European conference on machine learning*, pages 218–229. Springer, 2005.

[37] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[38] C. Luo, P. Zhao, C. Chen, B. Qiao, C. Du, H. Zhang, W. Wu, S. Cai, B. He, S. Rajmohan, et al. Pulns: Positive-unlabeled learning with effective negative sample selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8784–8792, 2021.

[39] M. Mahsereci, L. Balles, C. Lassner, and P. Hennig. Early stopping without a validation set. *arXiv preprint arXiv:1703.09580*, 2017.

[40] A. Menon, B. Van Rooyen, C. S. Ong, and B. Williamson. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning*, pages 125–134. PMLR, 2015.

[41] D. Moore. Uncertainty. on the shoulders of giants: new approaches to numeracy. la steen, 1990.

[42] B. Na, H. Kim, K. Song, W. Joo, Y.-Y. Kim, and I.-C. Moon. Deep generative positive-unlabeled learning under selection bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1155–1164, 2020.

[43] C. G. Northcutt, T. Wu, and I. L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

[44] H. Ramaswamy, C. Scott, and A. Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR, 2016.

[45] Y. Ren, D. Ji, and H. Zhang. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 488–498, 2014.

[46] H. Shi, S. Pan, J. Yang, and C. Gong. Positive and unlabeled learning via loss decomposition and centroid estimation. In *IJCAI*, pages 2689–2695, 2018.

[47] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[48] G. Su, W. Chen, and M. Xu. Positive-unlabeled learning from imbalanced data. In *IJCAI*, pages 2995–3001, 2021.

[49] D. Tanaka, D. Ikami, and K. Aizawa. A novel perspective for positive-unlabeled learning via noisy labels, 2021.

[50] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[51] C. Wang, J. Pu, Z. Xu, and J. Zhang. Asymmetric loss for positive-unlabeled learning. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[52] T. Wei, F. Shi, H. Wang, W.-W. T. Li, et al. Mixpul: Consistency-based augmentation for positive and unlabeled learning, 2020.

[53] J. Wilton, A. Koay, R. Ko, M. Xu, and N. Ye. Positive-unlabeled learning using random forests via recursive greedy risk minimization. In *Advances in Neural Information Processing Systems*, 2022.

[54] X. Xia, T. Liu, B. Han, M. Gong, J. Yu, G. Niu, and M. Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445*, 2021.

[55] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[56] P. Yang, X.-L. Li, J.-P. Mei, C.-K. Kwoh, and S.-K. Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.

[57] H. Yu, J. Han, and K. C.-C. 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*, pages 239–248, 2002.

[58] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021.

[59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[60] C. Zhang, D. Ren, T. Liu, J. Yang, and C. Gong. Positive and unlabeled learning with label disambiguation. In *IJCAI*, pages 4250–4256, 2019.

[61] Y. Zhao, Q. Xu, Y. Jiang, P. Wen, and Q. Huang. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14461–14470, 2022.

[62] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

[63] Z. Zhou, L.-Z. Guo, Z. Cheng, Y.-F. Li, and S. Pu. Step: Out-of-distribution detection in the presence of limited in-distribution labeled data. *Advances in Neural Information Processing Systems*, 34:29168–29180, 2021.

[64] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

A Analysis for Resampling Method

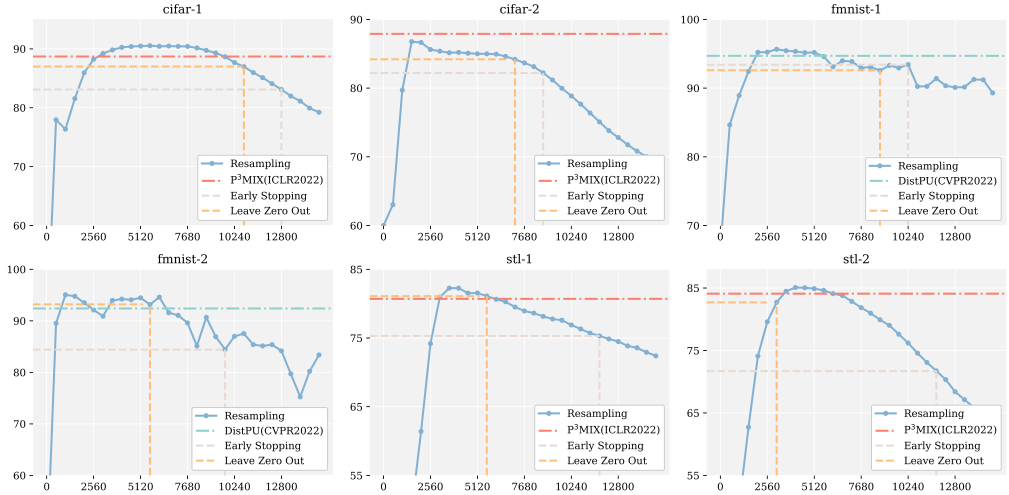

Figure 5: The accuracy of our resampling method on various settings across all 3 generic datasets. The horizontal line represents the accuracy of the state-of-the-art methods.

Empirical Results. Specifically, we employ the negativity assumption and resample positive data to achieve a balanced training distribution. Despite its simplicity, such a resampling approach achieves great empirical success as shown in Figure5, as it highlights the value of precious labels and mitigates the negative impact brought by false negatives and imbalanced label distribution. The outcomes suggest that the early predictive ability of the model could potentially facilitate our efforts in classification tasks. However, determining the optimal epoch to stop training and select the best model still remains a challenging task in PUL due to the absence of a precise validation set. For early stopping, we follow the settings in [52] and hold out 500 positive examples as a validation set. For LZO, we use an augmented validation set based on mix-up techniques following[34].

Assumption A.1. *We consider a naive situation where positive and negative data are drawn from a mixture of two Gaussians in $\mathbb{R}^p$ respectively and the dataset consists of n i.i.d. samples from the following distributions:*

$$\begin{aligned} \mathbb{P}(x|y=0) &\sim \mathcal{N}(+v, \sigma^2 I_{p\times p}), \\ \mathbb{P}(x|y=1) &\sim \mathcal{N}(-v, \sigma^2 I_{p\times p}). \end{aligned} \tag{9}$$

where v is an arbitrary unit vector in $\mathbb{R}^p$ and σ^2 is a small constant. Please keep in mind that the clusters are two spheres with radii $\sigma\sqrt{p} >> 2$ when $n, p \to \infty$ which makes this classification nontrivial. This binary classifier is trained by simply discriminating between positive and unlabeled data (i.i.d. sampled from the true distribution).

$$\mathbb{P}(x_u) \sim \pi\mathbb{P}(x|y=0) + (1-\pi)\mathbb{P}(x|y=1). \tag{10}$$

A.1 Bayesian Decision Hyperplane

Proposition A.1. *Under Assumption A.1, the Bayesian optimal decision hyperplane h_{pu} derived from the model using resampling strategy under a PU setting is equivalent to the Bayesian optimal decision hyperplane h^*_{pn} under a balanced PN binary classification setting.*

$$h_{pu} = h^*_{pn}. \tag{11}$$

Proof. We first discuss the decision hyperplane when both positive and negative data are available. By the virtue of Bayes' theorem, the score function g_{pn} and decision hyperplane h_{pn} separating each category at the same probability should be formulated as:

$$
\begin{aligned}
g_{pn}(x) &= g_p(x) - g_n(x) \\
&= ln[\mathbb{P}(x_p)\mathbb{P}(y=0)] - ln[\mathbb{P}(x_n)\mathbb{P}(y=1)] \\
&= ln\frac{\mathbb{P}(x|y=0)}{\mathbb{P}(x|y=1)} + ln\frac{\pi}{1-\pi} \\
&= ln\frac{N(+v, \sigma^2 I_{p\times p})}{N(-v, \sigma^2 I_{p\times p})} + ln\frac{\pi}{1-\pi} \\
&= \frac{2v^t x}{\sigma^2} + ln\frac{\pi}{1-\pi}.
\end{aligned}
\tag{12}
$$

$$
g_{pn}(x) = 0 \Rightarrow h_{pn} : 2v^t x + \sigma^2 ln\frac{\pi}{1-\pi} = 0. \tag{13}
$$

There exists an ideal decision hyperplane h^*_{pn} when the positive and negative data is balanced distributed ($\pi = 1 - \pi = 0.5$).

$$
h^*_{pn}(x) : 2v^t x = 0. \tag{14}
$$

When the distribution of negative data is unknown to us, we simply take the negativity assumption to make the classification by differentiating unlabeled data and positive data. Thus, the score function g_{pu} and decision hyperplane h_{pu} can be formulated as:

$$
\begin{aligned}
g_{pu}(x) &= g_p(x) - g_u(x) \\
&= ln[\mathbb{P}(x_p)\mathbb{P}(l)] - ln[\mathbb{P}(x_u)\mathbb{P}(u)] \\
&= ln\frac{\mathbb{P}(x|y=0)}{\pi\mathbb{P}(x|y=0) + (1-\pi)\mathbb{P}(x|y=1)} + ln\frac{\mathbb{P}(l)}{\mathbb{P}(u)} \\
&= ln\ N(+v, \sigma^2 I_{p\times p}) + ln\frac{|\mathcal{P}|}{|\mathcal{U}|} - ln[\pi N(+v, \sigma^2 I_{p\times p}) + (1-\pi)N(-v, \sigma^2 I_{p\times p}) \\
&= -ln[(1-\pi)exp(\frac{-2v^t x}{\sigma^2}) + \pi] + ln\frac{|\mathcal{P}|}{|\mathcal{U}|}.
\end{aligned}
\tag{15}
$$

$$
g_{pu}(x) = 0 \Rightarrow h_{pu} : 2v^t x + \sigma^2(ln(\frac{|\mathcal{P}|}{|\mathcal{U}|} - \pi) - ln(1-\pi)) = 0. \tag{16}
$$

When adopting a resampling strategy, the $|\mathcal{P}|/|\mathcal{U}|$ is set to 1, $h_{pu} = h^*_{pn}$. $\square$

We can also observe that when $\pi = 0$, Eq.16 degrades to Eq.13, which corresponds to the special case where the unlabeled set consists only of negative examples. However, it should be noted that in most cases, $|\mathcal{P}|/|\mathcal{U}|$ is less than π, making the decision hyperplane unlearnable. This underscores that label noise and data imbalance, introduced by the negativity assumption, are two key reasons for model degradation during the latter training phase. Therefore, we can consider $|\mathcal{P}|/|\mathcal{U}|$ as a flexible coefficient that controls the relative importance of data belonging to different classes. When we adopt a resampling strategy like our baseline, we aim to set this coefficient to 1, enabling us to derive an optimal decision hyperplane as shown in Eq.14.

A.2 Early Learning Phenomenon in PU Setting

To better illustrate model's early success when adoping the resampling strategy, we reformalize the theorem of Early Learning phenomenon given by [37] in a linear model and verify that this phenomenon also exists in PUL when taking cross entropy(CE) loss as the loss function. We first show that, for the first T iterations, the negative gradient has a constant correlation with v. (Note that, by contrast, a random vector in $\mathbb{R}^p$ typically has a negligible correlation with v.) Afterward, the false pseudo labels given by negativity assumption are memorized asymptotically.

Lemma A.1. *Under Assumption A.1, denote by* $\{S_t\}$ *the iterates of gradient descent with step size* η*. For any* $c \in (0, 1)$*, there exists a constant* σ_c *such that, if* $\sigma \leq \sigma_c$ *and* $p/n \in (1 - c/2, 1)$*, then with probability* $1 - o(1)$ *as* $n, p \to \infty$ *there exists a* $T = \Omega(1/\eta)$ *such that:*

- ***Early learning succeeds:*** *For $t < T$, $-\nabla \mathcal{L}_{CE}(S_t)$ is well correlated with the correct separator v, and at $t = T$ the classifier has higher accuracy on the wrongly labeled examples than at initialization.*
- ***Memorization occurs:*** *As $t \to \infty$, the classifier S_t memorizes all noisy labels.*

The only specialness of PUL setting is that the ratio of noise is given by negativity assumption controlled by the positive prior π. However, this only affects the constant c and corresponding σ_c. Readers are referred to [37] for detailed proof.

Combining the above empirical results and theoretical explanation, we better understand the capacity of resampling methods in the early stage of training.

A.3 Threshold Selection

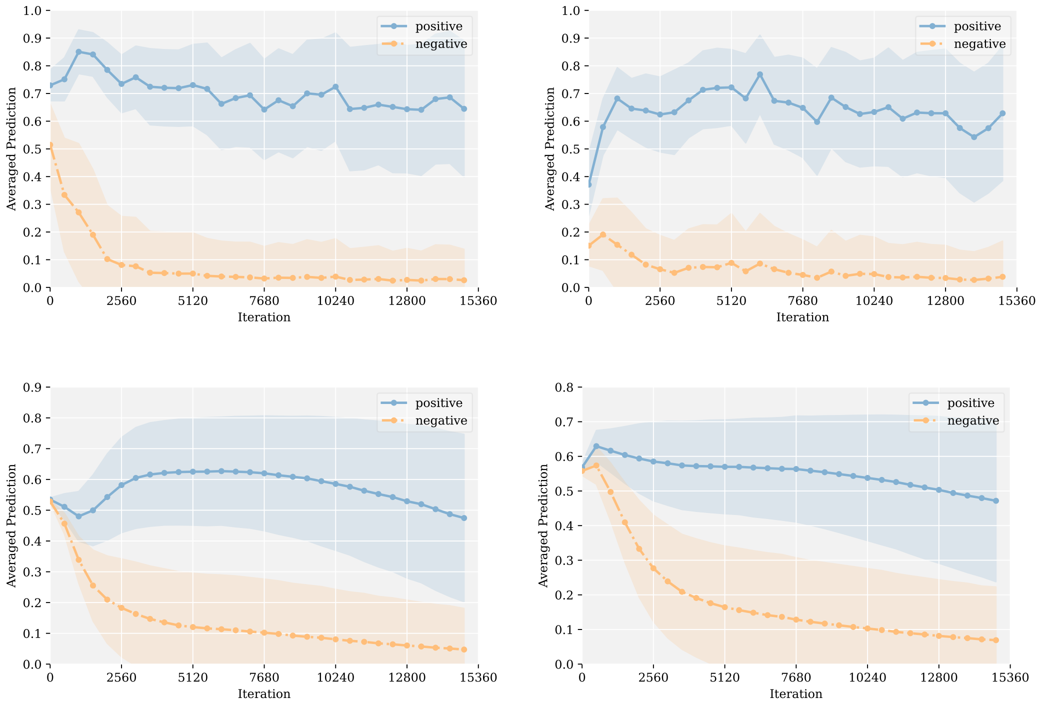

Figure 6: Averaged prediction confidence with a standard deviation of positive and negative examples on FMNIST1 (upper left), FMNIST2 (upper right), CIFAR10-1 (lower left) and CIFAR10-2 (lower right).

In this section, we present additional predictions and standard deviations obtained from four different settings utilizing CIFAR10 and FMNIST datasets. Notably, as illustrated in Figure 6, mislabeling errors of positive examples in the unlabeled set as negatives tend to increase with continued training when the threshold is set at 0.5. These results underscore the importance and challenge of accurately distinguishing between positive and negative examples in PUL tasks. Moreover, our findings indicate that differences between positive and negative examples are reflected in both the predictive trends and magnitudes of model-predicted scores. It also can be seen that, as the training progresses, the interval for an appropriate threshold shrinks.

B Mann-Kendall Test

The Mann-Kendall test is a non-parametric test used to determine if a time series has a trend over time. The test calculates the Mann-Kendall statistic S and the variance $Var(S)$. The test is performed by calculating:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(x_j - x_i). \tag{17}$$

where x is the time series data, n is the number of observations, and $sign()$ is the sign function that returns -1 if its argument is negative, 0 if its argument is zero, and 1 if its argument is positive. The variance of S is calculated as:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(t_p-1)(2t_p+5)}{18}. \tag{18}$$

where g is the number of tied groups, t_p is the number of tied values in the pth group. If the absolute value of S is greater than the critical value $(\alpha/2)$ times the standard error of $SE(S)$, where α is the significance level, then the null hypothesis of no trend is rejected. The standard error of S is calculated as:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{VAR(S)}}, S > 0 \\ \frac{S}{\sqrt{VAR(S)}}, S = 0 \\ \frac{S+1}{\sqrt{VAR(S)}}, S < 0 \end{cases} \tag{19}$$

To compute the significance of the Mann-Kendall test, we compare the absolute value of the Mann-Kendall statistic (Z_{MK}) to the critical value ($Z_{1-\alpha/2}$). The critical value depends on the level of significance (α) chosen and can be obtained from statistical tables or calculated using the software. If $|Z_{MK}| > Z_{1-\alpha/2}$, then the null hypothesis of no trend is rejected and we conclude that there is a significant trend present in the data.

The γ-value can also be calculated to determine the level of significance of the test. The γ-value is the probability of observing a Mann-Kendall statistic as extreme or more extreme than the observed value under the null hypothesis of no trend. If the γ-value is less than the chosen level of significance (α), then we reject the null hypothesis and conclude that there is a significant trend (either increasing or decreasing) in the data. If γ is bigger than α, we conclude there is no trend in this time series data.

To compute the γ-value, we first calculate the standardized test statistic (Z). Then, we calculate the probability of observing a Z value as extreme or more extreme than the observed value using a normal distribution table or software. The γ-value can be obtained by using the z-table.

C Proof of Theorem

Lemma C.1. *C_r-**inequlity:** For any $a, b \in \mathbb{R}$ and $p > 0$, we have:*

$$|a+b|^p \leq max\{2^{p-1}, 1\}(|a|^p + |b|^p), \tag{20}$$

and if $p > 1$, it is easy to verify:

$$|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p). \tag{21}$$

Before giving detailed proof, we first rewrite it as a reminder:

Theorem C.1. *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 $\tilde{S}$ is the statistic in the standardized Mann-Kendall test and σ^2 is the variance of P. By exploiting the non-decreasing influence function $\psi(x)$, for any $\epsilon > 0$, we have the following bound with probability at least $1 - 2\epsilon$:*

$$|\alpha\tilde{S} - \hat{S}| < \frac{2\alpha\sigma\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)}}}{1 - \sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)\alpha^2\sigma^2}}} = O\big((log(\epsilon^{-1}))^{\frac{1}{2}} t^{-1}\big). \tag{22}$$

Proof. We first specify some notions here for simplicity :

$$\alpha\tilde{S} = \frac{1}{t(t-1)}\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}\alpha\Delta p_{ij},\ \Delta p_{ij} = p_j - p_i,\ \alpha > 0. \tag{23}$$

$$S = \frac{2}{t(t-1)}\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}\psi(\alpha\Delta p_{ij}),\ \Delta p_{ij} = p_j - p_i,\ \alpha > 0. \tag{24}$$

$$\psi(x) = sign(x)\cdot log(1+|x|+x^2/2). \tag{25}$$

As suggested in [3], we can assume the upper and lower bounds of the proposed **trend score** S as S^- and S^+:

$$S^- \leq S \leq S^+. \tag{26}$$

Besides, although ψ is not derivative of some explicit error function, we will use it in the same way and consider it as an influence function. For some positive real parameter β, we will build our estimator $\hat{S}_\beta$ as the solution of the following equation:

$$\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}\psi[\beta(\alpha\Delta p_{ij} - \hat{S}_\beta)] = 0. \tag{27}$$

In fact, we choose the widest possible choice of the M estimator to derive a relatively stabilized empirical mean by making the smallest possible change that is closest to the empirical mean. Then we introduce the quantity and the exponential moment inequalities, from which deviation bounds will follow:

$$r(S) = \frac{2}{\beta t(t-1)}\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}\psi[\beta(\alpha\Delta p_{ij} - S)],\ S \in \mathbb{R}. \tag{28}$$

Simply following the assumptions and **Proposition 2.1** in [3], we can derive the following exponential moment inequalities through **LemmaC**:

$$\begin{aligned}
\mathbb{E}\big[e^{\frac{\beta t(t-1)r(S)}{2}}\big] &= \mathbb{E}[e^{\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}\psi\left(\beta(\alpha\Delta p_{ij}-S)\right)}] \\
&= \Big(\mathbb{E}\big[e^{\psi\left(\beta(\alpha\Delta p_{ij}-S)\right)}\big]\Big)^{\frac{t(t-1)}{2}} \\
&\leq \Big(\mathbb{E}\big[1+\beta(\alpha\Delta p_{ij}-S)+\frac{\beta^2}{2}(\alpha\Delta p_{ij}-S)^2\big]\Big)^{\frac{t(t-1)}{2}} \\
&\leq \Big(1+\beta(\alpha\tilde{S}-S)+\frac{\beta^2}{2}\mathbb{E}\big[(\alpha\Delta p_{ij}-S)^2\big]\Big)^{\frac{t(t-1)}{2}} \\
&\leq \Big(1+\beta(\alpha\tilde{S}-S)+\beta^2\big(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\big)\Big)^{\frac{t(t-1)}{2}} \\
&\leq e^{\frac{t(t-1)}{2}\beta(\alpha\tilde{S}-S)+\frac{t(t-1)}{2}\beta^2\left(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\right)}
\end{aligned} \tag{29}$$

Similarly, we have:

$$\mathbb{E}\big[e^{-\frac{\beta t(t-1)r(S)}{2}}\big] \leq e^{-\frac{t(t-1)}{2}\beta(\alpha\tilde{S}-S)+\frac{t(t-1)}{2}\beta^2\left(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\right)}. \tag{30}$$

According to Eq.29 and Eq.30, we have that for any $\epsilon \in (0, 1/2)$, there exists:

$$B_+(S) = \alpha\tilde{S} - S + \beta\big(\alpha^2\sigma^2 + (\alpha\tilde{S}-S)^2\big) + \frac{2log(\epsilon^{-1})}{t(t-1)\beta}. \tag{31}$$

$$B_-(S) = \alpha\tilde{S} - S - \beta\big(\alpha^2\sigma^2 + (\alpha\tilde{S}-S)^2\big) + \frac{2log(\epsilon^{-1})}{t(t-1)\beta}. \tag{32}$$

By Markov inequality and Eq.31 and Eq.32, we have:

$$\begin{aligned}\mathbb{P}\big(r(S) \geq B_+(S)\big) &= \mathbb{P}\big(e^{\frac{\beta t(t-1)r(S)}{2}} \geq e^{\frac{\beta t(t-1)B_+(S)}{2}}\big) \\ &\leq \frac{\mathbb{E}\big[e^{\frac{\beta t(t-1)r(S)}{2}}\big]}{e^{\frac{t(t-1)}{2}\beta(\alpha\tilde{S}-S)+\frac{t(t-1)}{2}\beta^2\left(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\right)+log(\epsilon^-1)}} \\ &\leq \frac{e^{\frac{t(t-1)}{2}\beta(\alpha\tilde{S}-S)+\frac{t(t-1)}{2}\beta^2\left(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\right)}}{e^{\frac{t(t-1)}{2}\beta(\alpha\tilde{S}-S)+\frac{t(t-1)}{2}\beta^2\left(\alpha^2\sigma^2+(\alpha\tilde{S}-S)^2\right)+log(\epsilon^-1)}} = \epsilon.\end{aligned} \tag{33}$$

Thus, we have:

$$\mathbb{P}\big(r(S) \leq B_+(S)\big) \geq 1-\epsilon. \tag{34}$$

Similarly,

$$\mathbb{P}\big(r(S) \geq B_-(S)\big) \geq 1-\epsilon. \tag{35}$$

Thus, we can claim:

$$\mathbb{P}\big(B_-(S) \leq r(S) \leq B_+(S)\big) \geq 1-2\epsilon. \tag{36}$$

According to **Lemma 2.3** in [7], we know that for positive real parameter β satisfying:

$$0 < \beta \leq \frac{\sqrt{\frac{1}{4}-\frac{2log(\epsilon^{-1})}{t(t-1)}}}{\alpha\sigma}. \tag{37}$$

there exists S_- and S_+ that $B_+(S_+) = 0$ and $B_-(S_+) = 0$, meanwhile, S_+ is the smallest solution and S_- is the largest solution. Then, it's easy to derive:

$$\mathbb{P}\big(S_- \leq \hat{S} \leq S_+\big) \geq 1-2\epsilon. \tag{38}$$

since our chosen $\psi(x)$ is a continuous function on x which also means that $r(S)$ is a continuous function on S. And we know from Eq.36 when $r(\hat{S}) = 0$ the following event holds with a probability of at least $1-2\epsilon$:

$$S_- \leq \hat{S} \leq S_+. \tag{39}$$

Following the **Theorem2.6** in [7], we denote $\beta = \frac{\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)}}}{\alpha\sigma}$, $n \geq (2\alpha^2\sigma^2+1)^2 log(\epsilon^{-1})/\alpha^2\sigma^2$. When the difference between S_- and S_+ is small we can derive the estimator can be localized in a small interval, which implies:

$$|\alpha\tilde{S}-\hat{S}| < \frac{2\alpha\sigma\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)}}}{1-\sqrt{\frac{2log(\epsilon^{-1})}{t(t-1)\alpha^2\sigma^2}}} = O\big((log(\epsilon^{-1}))^{\frac{1}{2}}t^{-1}\big). \tag{40}$$

holds with a probability of at least $1-2\epsilon$. ☐

After the theoretical analysis, we present a graph of our robust mean estimator, which sheds light on its underlying mechanism. As illustrated in Figure7, the estimator is less sensitive to outliers and deviations from normality when the input value x is too large or too small, as indicated by the flatter curve of $f(x)$ in its head and tail. Furthermore, the scaling parameter α enhances the flexibility of the estimator in handling extreme scenarios.

D Fisher-Jenks Natural Break Classification

In this section, we provide a specific training procedure for finding the Fisher Jenks Natural Break point in a binary scenario. As outlined in Algorithm1, the sorting process is simple and can be implemented using any sorting algorithm with a worst-case time complexity of $O\big(Nlog(N)\big)$. Then, we use a recursive approach to compute the mean and variance of the sequence in both ascending and descending orders. This enables us to obtain a chart for the sum of variances for every possible split, with a time complexity of $O\big(N\big)$. Therefore, the overall time complexity remains $O\big(Nlog(N)\big)$. Compared with the original algorithm of finding the Fisher Natural Break Point that asks for a time

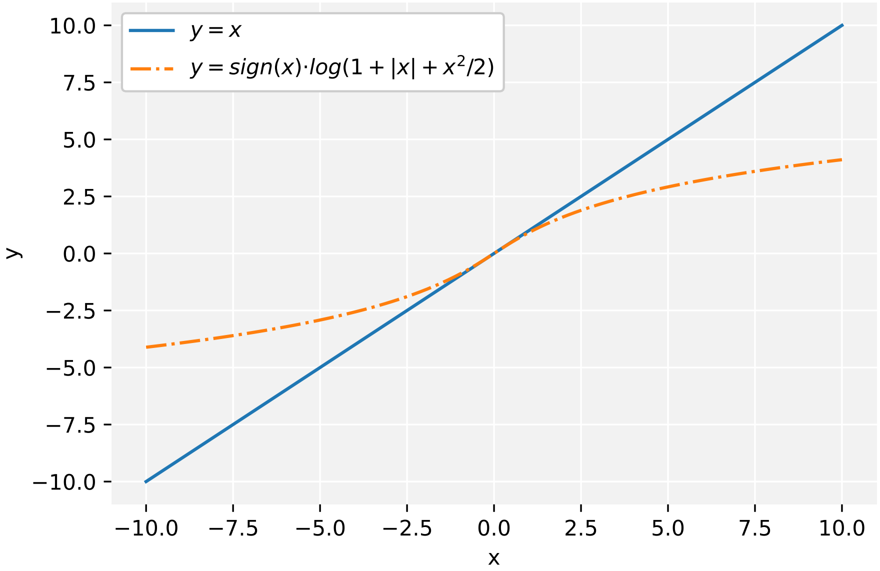

Figure 7: The illustration of our proposed robust mean estimator to assess the model's predictive trend.

Algorithm 1 Fisher (Jenks) Natural Break by Dynamic Programing

```
Input: Sequence of trend score values $x_i$ for $i \in 1, ..., N$
Output: Class-break index $b$
  sort $\mathcal{X} = \{x_i, 1 \le i \le N\}$ to a strictly increasing sequence.
  $\sigma_1^{2+} \leftarrow 0; \bar{X}_1^+ \leftarrow x_1; \sigma_N^{2-} - \leftarrow 0; \bar{X}_1^- \leftarrow x_n; b \leftarrow 0; s \leftarrow \infty$
  for $n = 2$ to $N$ do
    $\bar{X}_n^+ = \frac{1}{n} x_n + \frac{n-1}{n} \bar{X}_{n-1}^+$
    $\sigma_n^{2+} = \frac{n-2}{n-1} \sigma_{n-1}^{2+} + \frac{1}{n} (\bar{X}_n^+ - \bar{X}_{n-1}^+)^2$
  end for
  for $n = N - 1$ to $1$ do
    $\bar{X}_n^- = \frac{1}{n} x_n + \frac{n-1}{n} \bar{X}_{n-1}^-$
    $\sigma_n^{2-} = \frac{n-2}{n-1} \sigma_{n-1}^{2-} + \frac{1}{n} (\bar{X}_n^- - \bar{X}_{n-1}^-)^2$
  end for
  for $n = 1$ to $N - 1$ do
    if $\sigma_{n+1}^{2-} + \sigma_n^{2+} < s$ then
      $s = \sigma_{n+1}^{2-} + \sigma_n^{2+}; b = n$
    end if
  end for
  return $b$;
```

complexity of $O(N^2)$. Afterward, we provide a detailed derivation of our recursive method for computing the mean and variance. We take the variance σ_n^{2+} in ascending order as an example:

$$\bar{X}_n = \frac{1}{n} x_n + \frac{n-1}{n} \bar{X}_{n-1} \tag{41}$$

$$\sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[(x_i - \bar{X}_{n-1}) + (\bar{X}_{n-1} - \bar{X}_n) \right]^2. \tag{42}$$

where $\bar{X}_n$ is the averaged value of the first n values in the sequence. Then, we can have:

$$
\begin{aligned}
(n-1)\sigma_n^2 &= \sum_{i=1}^{n} \left[(x_i - \bar{X}_{n-1})^2 + (\bar{X}_{n-1} - \bar{X}_n)^2 + 2(x_i - \bar{X}_{n-1})(\bar{X}_{n-1} - \bar{X}_n)\right] \\
&= \sum_{i=1}^{n}(x_i - \bar{X}_{n-1})^2 + \sum_{i=1}^{n}(\bar{X}_{n-1} - \bar{X}_n)^2 + 2\sum_{i=1}^{n}(x_i - \bar{X}_{n-1})(\bar{X}_{n-1} - \bar{X}_n) \\
&= \sum_{i=1}^{n-1}(x_i - \bar{X}_{n-1})^2 + (x_n - \bar{X}_{n-1})^2 + n(\bar{X}_{n-1} - \bar{X}_n)^2 + \\
&\quad 2(\bar{X}_{n-1} - \bar{X}_n)\sum_{i=1}^{n}(x_i - \bar{X}_{n-1}) \\
&= (n-2)\sigma_{n-1}^2 + (x_n - \bar{X}_{n-1})^2 + n(\bar{X}_{n-1} - \bar{X}_n)^2 + \\
&\quad 2(\bar{X}_{n-1} - \bar{X}_n)\big[\sum_{i=1}^{n-1}(x_i - \bar{X}_{n-1}) + (x_n - \bar{X}_{n-1})\big] \\
&= (n-2)\sigma_{n-1}^2 + (x_n - \bar{X}_{n-1})^2 + n(\bar{X}_{n-1} - \bar{X}_n)^2 + \\
&\quad 2(\bar{X}_{n-1} - \bar{X}_n)(x_n - \bar{X}_{n-1}) \\
&= (n-2)\sigma_{n-1}^2 + (x_n - \bar{X}_{n-1})^2 + n(\bar{X}_{n-1} - \bar{X}_n)^2 - 2n(\bar{X}_{n-1} - \bar{X}_n)^2 \\
&= (n-2)\sigma_{n-1}^2 + (n^2 - n)(\bar{X}_{n-1} - \bar{X}_n)^2 \\
&= (n-2)\sigma_{n-1}^2 + \frac{n-1}{n}(x_n - \bar{X}_{n-1})^2.
\end{aligned}
\tag{43}
$$

Similarly, the variance σ_n^{2-} in descending order can be calculated in a similar way. Then, it's natural for us to have a chart for the sum of variances for every possible split from which the Fisher Natural break point is available.

E Additional Experiments

Here we discuss additional results in other practical settings and further demonstrate the robustness of our method. As mentioned in Section2.2, the model witnesses a dramatic performance degradation when positive data occupies a majority of the unlabeled set or the SCAR (selected completely at random) assumption is violated but such data scenarios are widespread in real-world applications. Moreover, we also make some brief comparisons with other methods under more complex backbones with a varying number of positive labels.

Table 9: Results of classification accuracy (%) on CIFAR10-1 wiht varying number of postive labels under different backbones (ResNet18 and CNN7 as the backbone model).

Backbone	Algorithm	n_p = 0.5k	n_p = 1k	n_p = 3k	n_p = 10k
CNN7	Resampling	86.29	90.02	92.64	93.41
	uPU	82.49	76.52	87.34	93.02
	nnPU	85.11	84.77	89.42	94.45
	vPU	83.05	86.74	90.54	**95.99**
	Dist-PU	85.15	87.25	91.76	95.07
	Ours	**87.21**	**90.58**	**91.80**	95.94
ResNet18	Resampling	84.27	88.32	90.21	93.88
	uPU	84.78	86.94	89.72	92.75
	nnPU	86.05	89.43	90.01	91.84
	vPU	71.40	86.85	88.54	89.89
	Dist-PU	92.15	92.94	93.47	**96.77**
	Ours	**93.21**	**94.58**	**95.77**	96.44

Based on Table9, the Trend-based PU framework performs better in scenarios where the number of positive labels is limited. This could be attributed to the fact that when there are 10,000 positive labels

Algorithm 2 Training procedure of the proposed method

```
**Input**: positive set $\mathcal{P}$, unlabeled set $\mathcal{U}$
**Parameter**: scaling parameter $\alpha$, evaluation step $q$
**Output**: model parameters $\Theta$
Initialize $\Theta$, $t = 0$ and translate the unlabeled set $\mathcal{U}$ into negative set by negativity assumption;
while $t \leq MaxEpoch$ do
    Shuffle $\mathcal{P} \cup \mathcal{U}$ into $I$ mini-batches and denote the $i$-th mini-batch as $(\mathcal{B}_p^i, \mathcal{B}_u^i)$;
    for $i = 1$ to $q$ do
        Compute the loss via Eq.1
        update model parameters $\Theta$ with Adam;
    end for
    Record the model's predictions on the unlabeled set $\mathcal{D}_t = \{p_1, p_2, \ldots, p_{|\mathcal{U}|}\}$
end while
for $i = 1$ to $|\mathcal{U}|$ do
    calculate the trend score $s_i$ on $\mathcal{D}$ through Eq.4 or Eq.6.
end for
Split the the unlabeled set $\mathcal{U}$ by Algorithm1 to get reformalized positive set $\mathcal{P}$ and negative set $\mathcal{N}$
Reinitialize $\Theta$ and train a binary model on the new positive set $\mathcal{P}$ and negative set $\mathcal{N}$
return model parameters $\Theta$
```

available, the estimation bias and prediction errors caused by the negative assumption are reduced due to the ample availability of supervised information. For imbalanced data, we give different imbalanced divisions compared with ImbalancedPU [48] by following the practice of long-tailed recognition. 10 different categories of CIFAR-10 are distributed under an exponential function with imbalance ratios γ in $\{10, 100, 1000\}$ (the ratio of most populated class to least populated) and we follow the division above in AppendixF to form the positive and negative set respectively. Thus, the positive prior π also gets fixed when the head class is determined as positive or negative. Compared with the division in ImbalancedPU that only choose one category as a positive class, our proposed one is more practical and challenging since it is common practice for a positive class to have different classes with an imbalanced number of data. Besides, in this case, the labeled data and positive data in the unlabeled set share different distributions which do not align with the common SCAR assumption. While our method also gets challenged when negative examples is rare, it still presents much better performance. Actually, when we look into this problem that the majority of unlabeled data is positive or negative. It even makes PUL two completely different questions,

Table 10: Results of classification accuracy(ACC), AUC and F1 score (%) on test set with same number of labels (1000) but varying positive prior.

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**

We compare our method with the resampling baseline and ImbalancedPU specially designed for imbalanced distributions based on popular nnPU and uPU. The results of accuracy, AUC and F1 score on the test set are given in Table 10. We denote the π as the positive prior of the whole dataset including the labeled data. It has illustrated that traditional cost-sensitive based methods can make competitive performance when the data distribution is balanced or positive class is rare. However, it witnesses a significant descent on all three metrics when the majority of unlabeled data belongs to the positive class and we argue that such a situation is quite common especially in the case when positive data is easy to obtain.

F Implementation details

The detailed description of these benchmark datasets is given in Table 4 and we denote the category labels with integers ranging from 0 to 9 following the default settings in torchvision. For each dataset, we split the dataset into two disjoint sets as positive and negative following the protocol of [6]. Specifically, the labels are defined as follows: F-MNIST-1: “0,2,4,7” vs “1,5,6,8,9”, F-MNIST-2: “1,5,6,8,9” vs “0,2,4,7”; 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”; Credit Fraud: "Fraud" vs "Non-Fraud; Alzheimer: "Demented" vs "Non-Demented".

For a fair comparison, we generally follow the experimental settings as [52, 61]. Specifically, we use the same data split as [52] in CIFAR-10-1, CIFAR-10-2, STL-10-1, STL-10-1 and Credit Card. For Alzheimer, F-MNIST-1 and F-MNIST-2, we follow the settings of [61]. To verify the effectiveness of our proposed method, We compare our method with several competitive PUL algorithms including uPU[13], nnPU[31], RP[43] nnPU with the mixup regularization term, Self-PU[8], PUSB[30], PUbN[26], aPU[21], vPU[6], MIXPUL[52], PAN[27], PULNS [38], Dist-PU[61] and P^3MIX [33]. For the methods requiring the positive prior, we provide them with an accurate prior except for STL since the true positive prior for STL is actually unknown considering it contains "real" unlabeled data. To this end, we estimate the positive prior of STL by KM2 method[44] before evaluating these methods. We report the results of these datasets under the backbones detailed in Table4 which is identical with [52]. It is worth mentioning that the true labels of unlabeled data in STL10 are not available and that’s the reason why we do not report any evaluation of the classification on the unlabeled training data in STL10. We run our method five times, following the procedure of [52], and report the average metrics and their standard deviations.

Furthermore, for the results presented in Table 2, we evaluate the key metrics of existing PUL methods based on their predictions on the unlabeled set, which can be considered as a transductive experimental setting. Specifically, we report the recall rate for the Credit Card dataset and the accuracy for the remaining datasets. For Table 3, we compare the estimated priors of our method with those of other state-of-the-art prior estimation methods. Although our method is not designed for prior estimation, the positive prior is naturally available when the classification of unlabeled data is performed.

In most cases, we perceive accuracy as the most important evaluation metric except for Credit Fraud dataset. In fraud detection, recall is often more important than precision or accuracy because the consequences of missing a fraudulent transaction can be much more severe than flagging a legitimate transaction as fraudulent. False negatives, which are fraudulent transactions that go undetected, can result in significant financial losses for both the individual and the company. On the other hand, false positives, which are legitimate transactions flagged as fraudulent, may cause temporary inconvenience but can usually be resolved through additional verification steps. Therefore, we emphasize more on recall rate and F1 score on the Credit Fraud dataset.

While existing Positive and Unlabeled Learning (PUL) methods mainly adopt an inductive learning paradigm, we have observed that some literature fails to report the hyperparameter tuning and model selection process. In traditional machine learning, researchers typically perform these tasks on an independent validation set, but this strategy may not be feasible in PUL due to the lack of negative data. While we can still use an extra positive set as a validation set, in real-world scenarios, the number of labeled data may be limited, especially for PUL paradigms. Furthermore, estimates made under such settings may be conservatively biased due to the limited number of data, particularly for small-scale validation sets. Instead of holding out data, we propose to perform model selection on an augmented validation set using mix-up techniques. Our approach yields comparable results to using an auxiliary positive validation set, as demonstrated in Table 2 and Table 1. In our comparison, we follow the settings in [52] and hold out 500 positive examples as a validation set. However, we use an augmented validation set based on mix-up techniques and the original labeled training set available to choose the stopping iteration to form our **trend score**.

For detailed experimental settings, we set the batch size to 64 and the evaluation step to 512 for all datasets and settings. The learning rate is set to 0.0015 for CIFAR10-1, CIFAR10-2, and STL10-1, 0.001 for STL10-2, and 0.002 for Credit Card and Alzheimer datasets. All experiments are implemented on RTX2080ti and RTX3080ti.

G Future Works

G.1 Risk Bound for PUL under SAR Assumption

In this subsection, we first review the upper and lower risk bound for PUL under the more general SAR assumption derived from [11]. Compared with the SCAR assumption that assumes the probability for a positive instance to be labeled is constant and thus independent from the covariates, a more general case is to assume the existence of a propensity function $e(x)$:

$$e(x) = \mathbb{P}(S = 1|Y = 1, X = x). \tag{44}$$

where $S = 1$ represents the labeled positive data. Besides, they also assume that the difficulty of the binary classification can be reflected by the *Massart margin* h derived from the regression function $\eta(x) = \mathbb{P}(Y = 1|X = x)$:

$$\exists h > 0, \forall x \in \mathbb{R}^d, |2\eta(x) - 1| \geq h. \tag{45}$$

Lemma G.1. *Let $\hat{g}$ be a minimizer of the unbiased empirical risk for PUL under the SAR assumption:$\hat{g} \in Argmin_{g \in \mathcal{G}} \hat{R}_n^{SAR}(g)$. Suppose that the separability and Massart margin hold, the propensity $e(.)$ is greater than $e_m > 0$. Then, we have the following upper bound on the excess risk:*

$$\mathbb{E}[\ell(\hat{g}, g^*)] \leq k_1 \Big[min\Big(\frac{V}{ne_m h}\big(1 + log(max(1, \frac{nh^2}{V})), \sqrt{\frac{V}{ne_m}}\Big)\Big]. \tag{46}$$

where $k_1 > 0$ is an absolute constant and V is the Vapnik-Chervonenkis dimension of $\mathcal{G}$[50].

Lemma G.2. *Suppose that $V \leq 2$ and $ne_m \geq V$. Let $h' = \sqrt{\frac{V}{ne_m}}$. Keep the assumptions hold in LemmaG.1, $\forall x \in \mathbb{R}^d$, there exists an absolute constant $k_2 > 0$ such that:*
if $h \geq h'$:

$$\mathcal{R}(\mathcal{G}, h) \geq k_2 \frac{V - 1}{hne_m}. \tag{47}$$

if $h \leq h'$:

$$\mathcal{R}(\mathcal{G}, h) \geq k_2 \sqrt{\frac{V - 1}{ne_m}}. \tag{48}$$

G.2 Limitation

It can be seen from LemmaG.1 and LemmaG.2 that both bounds depend on V, n, h and e_m. h evaluates the difficulty of the classification task and e_m represents the minimum of the propensity $e(.)$. When we recall the classification results in Table10 that evaluate the model's performance under various positive class priors. Both our method and the state-of-art PUL method special for imbalanced data witness a significant descent in all three metrics when the majority of unlabeled data belongs to the positive class. It may be explained by both the upper bounds and the lower bounds mentioned above. Specifically, when the majority of unlabeled data belongs to the positive class, e_m gets lower and both the upper and lower bounds in LemmaG.1 and LemmaG.2 get higher, making the classification more difficult. It asks for a more powerful model for PUL or a new perspective to tackle PUL. As argued in Section3, the predictive trends derived from the proposed resampling method can be a viable choice for such imbalanced scenarios. However, compared to the existing reweighting methods, the approach based on trend prediction still requires theoretical analysis. In addition, there are more possible methods worth exploring for additional resampling techniques, trend detection, and subsequent classification.