PUe: Biased Positive-Unlabeled Learning Enhancement by Causal Inference (suppmentary materials)

Xutao Wang[1], **Hanting Chen**[1], **Tianyu Guo**[1], **Yunhe Wang**[1*]
[1] Huawei Noah's Ark Lab.
{xutao.wang,chenhanting,tianyu.guo,yunhe.wang}@huawei.com,

1 Algorithm

Algorithm 1 PUe algorithm

Require: data χ_P,χ_U, size n,n_P,n_U, hyperparameter α_e,π.

Step 1:
Compute $\hat{e}(x)$ by minimizing $\frac{\pi}{n_P}\sum_{i=1}^{n_P} L(g(\boldsymbol{x}_i^P),+1) + \frac{1-\pi}{n_U}\sum_{i=1}^{n_U} L(g(\boldsymbol{x}_i^U),-1) + \alpha_e|\sum_{x_i\in\chi_P\cup\chi_U} e(x_i) - n_P|$;
Step 2:
Compute the weight of labeled samples: $w_i^P = \frac{\pi}{\hat{\tilde{e}}(x_i^P)}$
Step 3:
for $i = 1 \ldots$ **do**
    Shuffle (χ_P, χ_U) into M mini-batches
    for each mini-batch (χ_P^j, χ_U^j) **do**
        Compute the corresponding $\hat{R}_{\text{PUe}}(g)$
        Use $\mathcal{A}$ to update θ with the gradient information $\nabla_\theta \hat{R}_{\text{PUe}}(g)$
    end for
end for

2 Experiment Details

Table 1: Summary of used datasets and their corresponding models.

Dataset	Input Size	n_P	n_U	# Testing	π_P	Positive Class	true e(x)	Model
MNIST	28×28	2500	60,000	10,000	0.5	Even (0, 2, 4, 6 and 8)	[.65,.15,.1,.07,.03]	6-layer MLP
CIFAR-10	$3 \times 32 \times 32$	1,000	50,000	10,000	0.4	Vehicles (0, 1, 8 and 9)	[.72,.15,.1,.03]	13-layer CNN
Alzheimer	$3 \times 224 \times 224$	769	5,121	1,279	0.5	Alzheimer's Disease	unknow	ResNet-50

3 Complementary Experiment

LRe: Logistic regression estimation of propensity scores for PU learning.

According to paper [**?**], it cannot estimate identifiable PS without making certain assumptions about the data. But according to the formula we gave in the first question, it's approximate. This is not explained by the self-monitoring method. Results in the above table show that our scheme is better than self-PU in the case of biased label datasets.

*Corresponding Author.

Table 2: Supplemental Experiments on MINST

method	labeled distripution	ACC.(%)	Prec.(%)	Rec.(%)	F1.(%)	AUC.(%)	AP.(%)
LRe	[.65,.15,.10,.07,.03]	86.19(0.75)	92.94(0.64)	77.89(1.38)	84.75(0.93)	88.06(0.93)	88.72(1.09)
nnPUe	[.65,.15,.10,.07,.03]	92.45 (1.61)	90.45 (2.26)	94.73 (1.24)	92.53 (1.55)	92.48 (1.60)	88.29 (2.43)
nnPU without normalize	[.65,.15,.10,.07,.03]	90.95(1.61)	88.18(2.40)	94.38(2.74)	91.13(1.56)	91.00(1.61)	85.98(2.25)
Self-PU	[.65,.15,.10,.07,.03]	90.08(0.47)	90.08 (0.47)	89.35 (1.17)	90.70 (1.73)	90.00 (0.53)	85.61 (0.69)
anchor	[.65,.15,.10,.07,.03]	88.22(0.95)	94.66(1.41)	80.70(3.15)	87.06(1.37)	92.36(1.92)	93.37(2.33)

4 Proofs

4.1 error bound of bias

We may assume that the error of propensity scores estimated by the NN method is the same as that estimated by the linear method. (In fact, the NN methods are usually more general, which may produce results with less error.) That is, the estimate of the propensity score has a maximum error ratio of β, with $\beta e(x_i^L) \le \hat{e}(x_i^L) \le e(x_i^L)$. of the following shows that our regularization technique can yield a smaller error ratio with respect to sample weights. Obviously, the sample x_i^L has a sample weight of $\frac{1}{n\hat{e}(x_i^L)}$. in (Formula 1) with an error bound of $bias(\frac{1}{n\hat{e}(x_i^L)}) \le \frac{1}{ne(x_i^L)}(\frac{1}{\beta} - 1)$.

4.2 error ratio

In our approach, Sample x_i^L has a weight of $\pi\frac{\frac{1}{\hat{e}(x_i^L)}}{\sum_j \frac{1}{\hat{e}(x_j^L)}}$.$P(\gamma e(x_i^L) < \hat{e}(x_i^L) \le e(x_i^L)) = \alpha$ where the set of samples is S_1 .$P(\beta e(x_i^L) < \hat{e}(x_i^L) \le \gamma e(x_i^L)) = 1 - \alpha$ where the set of samples is S_2.$\beta < \gamma < 1$ and $\sum_{i\in S_1} \frac{1}{e(x_i^L)} = \sum_{i\in S_2} \frac{1}{e(x_i^L)} = B$. So that $\sum_j \frac{1}{e(x_j^L)} = 2B = N_p$. For $x_i^L \in S_1$, we have $\frac{1}{e(x_i^L)} \le \frac{1}{\hat{e}(x_i^L)} < \frac{1}{\gamma e(x_i^L)}$. For $x_i^L \in S_2$, we have $\frac{1}{\gamma e(x_i^L)} \le \frac{1}{\hat{e}(x_i^L)} < \frac{1}{\beta e(x_i^L)}$, so we can get $B(1 + \frac{1}{\gamma}) \le \sum_j \frac{1}{\hat{e}(x_j^L)} < B(\frac{1}{\gamma} + \frac{1}{\beta})$ and we have $bias(\pi\frac{\frac{1}{\hat{e}(x_i^L)}}{\sum_j \frac{1}{\hat{e}(x_j^L)}}) \le$max$[\frac{1}{ne(x_i^L)}(\frac{2}{(1+\gamma)} - 1), \frac{1}{ne(x_i^L)}(1 - \frac{2}{\frac{1}{\beta}+\frac{1}{\gamma}}), \frac{1}{ne(x_i^L)}(\frac{2}{(1+\frac{1}{\gamma})\beta} - 1), \frac{1}{ne(x_i^L)}(1 - \frac{2}{\frac{\gamma}{\beta}+1})] < \frac{1}{ne(x_i^L)}(\frac{1}{\beta} - 1)$ and obviously we have $\frac{2}{(1+\gamma)} < \frac{2}{(1+\frac{1}{\gamma})\beta} < \frac{1}{\beta}, 0 < 1 - \frac{2}{\frac{\gamma}{\beta}+1} < 1 - \frac{2}{\frac{1}{\beta}+\frac{1}{\gamma}} < 1 - \beta < \frac{1}{\beta} - 1$, which shows that our regularization technique has a smaller error ratio with respect to sample weights.

4.3 expectation

One understanding is that, according to the PS definition, each labeled sample x_j^P corresponds to $\frac{1}{e(x_j^P)}$ positive samples. So $\sum_{j=1}^{n_p} \frac{1}{e(x_j^P)} = N_p$ it's true. Because $P(x|s=1) = P(x, y=1|s=1)$, we have

$E_{P(x|s=1)} \frac{1}{P(s=1|x,y=1)}$

$= \sum P(x, y=1|s=1) \frac{1}{P(s=1|x,y=1)}$

$= \sum \frac{P(s=1|x,y=1)P(x,y=1)}{P(s=1)} \frac{1}{P(s=1|x,y=1)}$

$= \sum \frac{P(x,y=1)}{P(s=1)} = \frac{n}{n_P} \sum P(x, y=1)$

$= \frac{n}{n_P} \frac{N_P}{n} = \frac{N_P}{n_P}$.

It indicates that $\sum_{j=1}^{n_p} \frac{1}{e(x_j^P)} = N_p$.

4.4 PUbN

The PUbN formula is as follows:

Let $\sigma(x) = p(s=+1|x)$, however, the $\sigma(x)$ is actually unknown,we should replace $\sigma(x)$ by its estimate $\hat{\sigma}(x)$.We can get the classification risk of PUbN ($R_{PUbN}(g)$), as the following expression:

$R_{PUbN}(g) = \pi R_P(g,+1) + \rho R_{bN}(g,-1) + \bar{R}_{s=-1,\eta,\hat{\sigma}}(g)$

where $\bar{R}_{s=-1,\eta,\hat{\sigma}}(g) = \mathbb{E}_{x\sim p(x)}[\mathbb{1}_{\hat{\sigma}(x)\le\eta}L(-g(x))(1-\hat{\sigma}(x))] + \pi\mathbb{E}_{x\sim p_P(x)}[\mathbb{1}_{\hat{\sigma}(x)>\eta}L(-g(x))\frac{1-\hat{\sigma}(x)}{\hat{\sigma}(x)}] + \rho\mathbb{E}_{x\sim p_{bN}(x)}[\mathbb{1}_{\hat{\sigma}(x)>\eta}L(-g(x))\frac{1-\hat{\sigma}(x)}{\hat{\sigma}(x)}]$

Then $R_{bN}(g,-1)$ and $\bar{R}_{s=-1,\eta,\hat{\sigma}}(g)$ can also be approximated from data by $\hat{R}_{bN}(g,-1) = \frac{1}{n_{bN}}\sum_{i=1}^{n_{bN}} L(g(x_i^{bN}),-1)$ $\hat{\bar{R}}_{s=-1,\eta,\hat{\sigma}}(g) = \frac{1}{n_U}\sum_{i=1}^{n_U}[\mathbb{1}_{\hat{\sigma}(x_i^U)\le\eta}L(g(x_i^U),-1)(1-\hat{\sigma}(x_i^U))] + \frac{\pi}{n_P}\sum_{i=1}^{n_P}[\mathbb{1}_{\hat{\sigma}(x_i^P)>\eta}L(g(x_i^P),-1)\frac{1-\hat{\sigma}(x_i^P)}{\hat{\sigma}(x_i^P)}] + \frac{\rho}{n_b N}\sum_{i=1}^{n_{bN}}[\mathbb{1}_{\hat{\sigma}(x_i^{bN})>\eta}L(g(x_i^{bN}),-1)\frac{1-\hat{\sigma}(x_i^{bN})}{\hat{\sigma}(x_i^{bN})}]$

$\hat{R}_{PUbN,\eta,\hat{\sigma}}(g) = \pi\hat{R}_P(g,+1) + \rho\hat{R}_{bN}(g,-1) + \hat{\bar{R}}_{s=-1,\eta,\hat{\sigma}}(g)$

4.5 PUbNe

Our PUbNe formula is as follows:

$\hat{R}_{PUbN\hat{e},\eta,\hat{\sigma}}(g) = \pi\hat{R}_P^{\hat{e}}(g,+1) + \rho\hat{R}_{bN}^{\hat{e}}(g,-1) + \hat{\bar{R}}^{\hat{e}}_{s=-1,\eta,\hat{\sigma}}(g)$

,where $\hat{R}_{bN}^{\hat{e}}(g,-1) = \sum_{i=1}^{n_{bN}} \frac{1}{\tilde{\hat{e}}(x_i^{bN})}L(g(x_i^{bN}),-1)$ and

$\hat{\bar{R}}^{\hat{e}}_{s=-1,\eta,\hat{\sigma}}(g) = \frac{1}{n_U}\sum_{i=1}^{n_U}[\mathbb{1}_{\hat{\sigma}(x_i^U)\le\eta}L(g(x_i^U),-1)(1-\hat{\sigma}(x_i^U))] + \pi\sum_{i=1}^{n_P}[\frac{1}{\tilde{\hat{e}}(x_i^P)}\mathbb{1}_{\hat{\sigma}(x_i^P)>\eta}L(g(x_i^P),-1)\frac{1-\hat{\sigma}(x_i^P)}{\hat{\sigma}(x_i^P)}] + \rho\sum_{i=1}^{n_{bN}}[\frac{1}{\tilde{\hat{e}}(x_i^{bN})}\mathbb{1}_{\hat{\sigma}(x_i^{bN})>\eta}L(g(x_i^{bN}),-1)\frac{1-\hat{\sigma}(x_i^{bN})}{\hat{\sigma}(x_i^{bN})}]$

4.6 unbiased

$$
\begin{aligned}
&\mathbb{E}[\hat{R}_{PUe}(g)] \\
&\quad = \mathbb{E}[\pi\hat{R}_P^e(g,+1) + \hat{R}_U(g,-1) - \pi\hat{R}_P^e(g,-1)] \\
&\quad = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n_P}\frac{1}{e(\boldsymbol{x}_i^P)}\left(L(g(\boldsymbol{x}_i^P),+1) - L(g(\boldsymbol{x}_i^P),-1)\right) + \frac{1}{n}\sum_{i=1}^{n}L(g(\boldsymbol{x}_i),-1)] \\
&\quad = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n_P}\frac{1}{e(\boldsymbol{x}_i^P)}L(g(\boldsymbol{x}_i^P),+1) + \left(1-\frac{1}{e(\boldsymbol{x}_i^P)}\right)L(g(\boldsymbol{x}_i^P),-1) + \frac{1}{n}\sum_{i=1}^{n}(1-s_i)L(g(\boldsymbol{x}_i),-1)] \\
&\quad = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}s_i\frac{1}{e(\boldsymbol{x}_i^P)}L(g(\boldsymbol{x}_i^P),+1) + s_i\left(1-\frac{1}{e(\boldsymbol{x}_i^P)}\right)L(g(\boldsymbol{x}_i^P),-1) + (1-s_i)L(g(\boldsymbol{x}_i),-1)] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}y_ie_i\frac{1}{e(\boldsymbol{x}_i)}L(g(\boldsymbol{x}_i),+1) + y_ie_i\left(1-\frac{1}{e(\boldsymbol{x}_i)}\right)L(g(\boldsymbol{x}_i),-1) + (1-y_ie_i)L(g(\boldsymbol{x}_i),-1) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}y_iL(g(\boldsymbol{x}_i),+1) + y_i\left(e_i-1\right)L(g(\boldsymbol{x}_i),-1) + (1-y_ie_i)L(g(\boldsymbol{x}_i),-1) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}y_iL(g(\boldsymbol{x}_i),+1) + (1-y_i)L(g(\boldsymbol{x}_i),-1) \\
&\quad = R_{PN}(g|y).
\end{aligned}
\tag{1}
$$

The change of $\hat{R}_{PUe}(g)$ will be no more than $L_{\max}/n$ if some $x_i \in \boldsymbol{\chi}_P \cup \boldsymbol{\chi}_U$ is replaced, and McDiarmid's inequality gives us:

$$
\Pr\{|\hat{R}_{PUe}(g) - R_{PN}(g|y)| \ge \epsilon\} = \Pr\{|\hat{R}_{PUe}(g) - \mathbb{E}[\hat{R}_{PUe}(g)| \ge \epsilon\} \le 2\exp\left(-\frac{2\epsilon^2}{n(L_{\max}/n)^2}\right).
$$

And make the right side of the previous formula equal to η:

$$
\begin{aligned}
2\exp\left(-\frac{2\epsilon^2}{n(L_{\max}/n)^2}\right) = \eta & \\
& \iff \exp\left(\frac{2\epsilon^2}{n(L_{\max}/n)^2}\right) = \frac{2}{\eta} \\
& \iff \frac{2\epsilon^2}{L_{\max}^2/n} = \ln\left(\frac{2}{\eta}\right) \\
& \iff 2\epsilon^2 = \frac{L_{\max}^2 \ln\left(\frac{2}{\eta}\right)}{n} \\
& \iff \epsilon = \sqrt{\frac{L_{\max}^2 \ln\left(\frac{2}{\eta}\right)}{2n}}
\end{aligned}
$$

Equivalently, with probability at least $1-\eta$,

$$
|\hat{R}_{PUe}(g) - R_{PN}(g|y)| = |\hat{R}_{PUe}(g) - \mathbb{E}[\hat{R}_{PUe}(g)]| \leq \sqrt{\frac{L_{\max}^2 \ln \frac{2}{\eta}}{2n}}. \tag{2}
$$

And because we know the expressions of $\hat{R}_{PUe}(g)$ and $\hat{R}_{uPU}(g)$:

$$
\hat{R}_{PUe}(g) = \pi \hat{R}_P^e(g,+1) + \hat{R}_U(g,-1) - \pi \hat{R}_P^e(g,-1), \tag{3}
$$

$$
\hat{R}_{PU}(g) = \pi \hat{R}_P(g,+1) + \hat{R}_U(g,-1) - \pi \hat{R}_P(g,-1), \tag{4}
$$

Since we know that $\sum_{j=1}^{n_P} \frac{1}{e(\boldsymbol{x}_j^P)} = N_P$, we can get the following formula:

$$
\begin{aligned}
|\hat{R}_{PUe}(g) - \hat{R}_{PU}(g)| & \\
& = \pi|\hat{R}_P^e(g,+1) - \hat{R}_P^e(g,-1) - \Big((\hat{R}_P(g,+1) - \hat{R}_P(g,-1)\Big)| \\
& \leq \frac{1}{n}|\sum_{i=1}^{n_P}(\frac{1}{e(\boldsymbol{x}_i^P)} - \frac{N_P}{n_p})L(g(\boldsymbol{x}_i^P),+1)| + \frac{1}{n}|\sum_{i=1}^{n_P}(\frac{1}{e(\boldsymbol{x}_i^P)} - \frac{N_P}{n_p})L(g(\boldsymbol{x}_i^P),-1)| \\
& \leq \frac{2}{n} N_P L_{\max} = 2\pi L_{\max}
\end{aligned} \tag{5}
$$

Then, we can prove:

$$
\begin{aligned}
|R_{PN}(g|y) - \hat{R}_{PU}(g)| & \\
& \leq |R_{PN}(g|y) - \hat{R}_{PUe}(g)| + |\hat{R}_{PUe}(g) - \hat{R}_{PU}(g)| \\
& \leq 2\pi L_{\max} + \sqrt{\frac{L_{\max}^2 \ln \frac{2}{\eta}}{2n}}.
\end{aligned} \tag{6}
$$