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

Xutao Wang¹, Hanting Chen¹, Tianyu Guo¹, Yunhe Wang¹
¹Huawei Noah’s Ark Lab.
{xutao.wang,chenhanting,tianyu.guo,yunhe.wang}@huawei.com

1 Algorithm

Algorithm 1 PUe algorithm

Require: data $\chi_P, \chi_U$, size $n, n_P, n_U$, hyperparameter $\alpha_e, \pi$.

1: Step 1:
2: Compute $\hat{e}(x)$ by minimizing $\sum_{i=1}^{n_P} L(g(x^P_i), +1) + \frac{1-\pi}{n_U} \sum_{i=1}^{n_U} L(g(x^U_i), -1) + \alpha_e \left| \sum_{i \in \chi_P \cup \chi_U} e(x_i) - n_P \right|$

3: Step 2:
4: Compute the weight of labeled samples: $w^P_i = \frac{n}{\hat{e}(x^P_i)}$

5: Step 3:
6: for $i = 1 \ldots n$ do
7: Shuffle $\chi_P, \chi_U$ into $M$ mini-batches
8: for each mini-batch $(\chi^P_j, \chi^U_j)$ do
9: Compute the corresponding $\hat{R}_{PUe}(g)$
10: Use $A$ to update $\theta$ with the gradient information $\nabla_{\theta} \hat{R}_{PUe}(g)$
11: end for
12: end for

2 Experiment Details

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input Size</th>
<th>$n_P$</th>
<th>$n_U$</th>
<th># Testing $\pi_P$</th>
<th>Positive Class</th>
<th>true $e(x)$</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>28 $\times$ 28</td>
<td>25000</td>
<td>60,000</td>
<td>10,000</td>
<td>0.5 Even (0, 2, 4, 6 and 8) [.65, .15, .1, .07, .03]</td>
<td>6-layer MLP</td>
<td></td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>3 $\times$ 32 $\times$ 32</td>
<td>1,000</td>
<td>50,000</td>
<td>10,000</td>
<td>0.4 Vehicles (0, 1, 8 and 9) [.72, .15, .1, .03]</td>
<td>13-layer CNN</td>
<td></td>
</tr>
<tr>
<td>Alzheimer</td>
<td>3 $\times$ 224 $\times$ 224</td>
<td>769</td>
<td>5,121</td>
<td>1,279</td>
<td>0.5 Alzheimer’s Disease</td>
<td>unknow</td>
<td>ResNet-50</td>
</tr>
</tbody>
</table>

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.

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 $\beta$, with $\beta e(x^T_t) \leq \hat{e}(x^T_t) \leq e(x^T_t)$. Of the following shows that our regularization technique can yield a smaller error ratio with respect to sample weights. Obviously, the sample $x^T_t$ has a sample weight of $\frac{1}{ne(x^T_t)}$, in (Formula 1) with an error bound of $bias(\frac{1}{ne(x^T_t)}) \leq \frac{1}{ne(x^T_t)} (\beta - 1)$.

### 4.2 error ratio

In our approach, Sample $x^T_t$ has a weight of $\pi \frac{1}{\sum_{i} \gamma^E(x^T_t)} P(\gamma e(x^T_t) < \hat{e}(x^T_t) \leq e(x^T_t)) = \alpha$ where the set of samples is $S_1$. $P(\beta e(x^T_t) < \hat{e}(x^T_t) \leq \gamma e(x^T_t)) = 1 - \alpha$ where the set of samples is $S_2$. $\beta < \gamma < 1$ and $\sum_{i \in S_1} \frac{1}{\gamma^E(x^T_t)} = \sum_{i \in S_2} \frac{1}{\gamma^E(x^T_t)} = B$. So that $\sum_{j} \frac{1}{\gamma^E(x^T_t)} = 2B = N_p$. For $x^T_t \in S_1$, we have $\frac{1}{\gamma^E(x^T_t)} < \frac{1}{\hat{e}(x^T_t)} < \frac{1}{\gamma e(x^T_t)}$. For $x^T_t \in S_2$, we have $\frac{1}{\gamma e(x^T_t)} < \frac{1}{\hat{e}(x^T_t)} < \frac{1}{\beta e(x^T_t)}$, so we can get $B(1 + \frac{1}{\gamma}) \leq \sum_{j} \frac{1}{\gamma^E(x^T_t)} < B(\frac{1}{\beta} + \frac{1}{\gamma})$ and we have $bias(\pi \frac{\alpha^E(x^T_t)}{\sum_{j} \frac{1}{\gamma^E(x^T_t)}} \leq max[\frac{1}{ne(x^T_t)} (\frac{2}{1 + \frac{1}{\gamma}} - 1), \frac{2}{1 + \frac{1}{\gamma}}] < \frac{1}{ne(x^T_t)} (\frac{2}{\gamma} - 1)$ and obviously we have $\frac{1}{ne(x^T_t)} (\frac{2}{1 + \frac{1}{\gamma}}) < \frac{1}{ne(x^T_t)} (\frac{2}{\gamma} - 1)$. It 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 $x^P_t$ corresponds to $\frac{1}{\gamma^E(x^T_t)}$ positive samples. So $\sum_{j=1}^{N_p} \frac{1}{\gamma^E(x^T_t)} = N_p$ it’s true. Because $P(x|s=1) = P(x, y=1|s=1)$, we have

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

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

$= \frac{n_p N_p}{n_p N_p} = N_p$.

It indicates that $\sum_{i=1}^{N_p} \frac{1}{\gamma^E(x^T_t)} = N_p$.

### 4.4 PUbN

The PUbN formula is as follows:

<table>
<thead>
<tr>
<th>method</th>
<th>labeled distribution</th>
<th>ACC. (%)</th>
<th>Prec. (%)</th>
<th>Rec. (%)</th>
<th>F1. (%)</th>
<th>AUC. (%)</th>
<th>AP. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>[.65, .15, .07, .03]</td>
<td>86.19(0.75)</td>
<td>92.94(0.64)</td>
<td>77.89(1.38)</td>
<td>84.75(0.93)</td>
<td>88.06(0.93)</td>
<td>88.72(1.09)</td>
</tr>
<tr>
<td>nnPU</td>
<td>[.65, .15, .07, .03]</td>
<td>92.45 (1.61)</td>
<td>91.90 (2.26)</td>
<td>94.73 (1.24)</td>
<td>92.53 (1.55)</td>
<td>92.48 (1.60)</td>
<td>88.29 (2.43)</td>
</tr>
<tr>
<td>nnPU without normalize</td>
<td>[.65, .15, .07, .03]</td>
<td>90.95 (1.61)</td>
<td>88.18 (2.40)</td>
<td>94.38 (2.74)</td>
<td>91.13 (1.56)</td>
<td>91.00 (1.61)</td>
<td>85.98 (2.25)</td>
</tr>
<tr>
<td>Self-PU</td>
<td>[.65, .15, .07, .03]</td>
<td>90.08 (0.47)</td>
<td>90.08 (0.47)</td>
<td>89.35 (1.17)</td>
<td>90.70 (1.73)</td>
<td>90.00 (0.53)</td>
<td>85.61 (0.69)</td>
</tr>
<tr>
<td>anchor</td>
<td>[.65, .15, .07, .03]</td>
<td>88.22 (0.95)</td>
<td>94.66 (1.41)</td>
<td>80.70 (3.15)</td>
<td>87.06 (1.37)</td>
<td>92.36 (1.92)</td>
<td>93.37 (2.33)</td>
</tr>
</tbody>
</table>
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:

\[
R_{PUBN}(g) = \pi R_P(g, +1) + \rho R_{bN}(g, -1) + \hat{R}_{s=-1, \eta, \hat{\sigma}}(g)
\]

where \( \hat{R}_{s=-1, \eta, \hat{\sigma}}(g) \) is given by

\[
\hat{R}_{s=-1, \eta, \hat{\sigma}}(g) = \mathbb{E}_{x \sim \text{p}(x)} \left[ \mathbb{1}_{\hat{\sigma}(x) \leq \eta} L(-g(x)) (1 - \hat{\sigma}(x)) \right] + \pi \mathbb{E}_{x \sim \text{p}(x)} \left[ \mathbb{1}_{\hat{\sigma}(x) > \eta} L(-g(x)) (1 - \hat{\sigma}(x)) \right]
\]

Then \( R_{bN}(g, -1) \) and \( \hat{R}_{s=-1, \eta, \hat{\sigma}}(g) \) 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), -1) \hat{R}_{s=-1, \eta, \hat{\sigma}}(g) = \frac{1}{n_{bN}} \sum_{i=1}^{n_{bN}} \mathbb{1}_{\hat{\sigma}(x_i^{bN}) \leq \eta} L(g(x_i^{bN}, -1) (1 - \hat{\sigma}(x_i^{bN})) \]

\[
+ \frac{\pi}{n_{bN}} \sum_{i=1}^{n_{bN}} \mathbb{1}_{\hat{\sigma}(x_i^{bN}) > \eta} L(g(x_i^{bN}, -1) (1 - \hat{\sigma}(x_i^{bN})) \]

\[\hat{R}_{PUBN, \eta, \hat{\sigma}}(g) = \pi \hat{R}_P(g, +1) + \rho \hat{R}_{bN}(g, -1) + \hat{R}_{s=-1, \eta, \hat{\sigma}}(g) \]

### 4.5 PUbNe

Our PUbNe formula is as follows:

\[
\hat{R}_{PUBN, \eta, \hat{\sigma}}(g) = \pi \hat{R}_P(g, +1) + \rho \hat{R}_{bN}(g, -1) + \hat{R}_{s=-1, \eta, \hat{\sigma}}(g)
\]

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

\[
\hat{R}_{s=-1, \eta, \hat{\sigma}}(g) = \frac{1}{n_{bN}} \sum_{i=1}^{n_{bN}} \mathbb{1}_{\hat{\sigma}(x_i^{bN}) \leq \eta} L(g(x_i^{bN}, -1) (1 - \hat{\sigma}(x_i^{bN})) \]

\[
+ \pi \sum_{i=1}^{n_{bN}} \mathbb{1}_{\hat{\sigma}(x_i^{bN}) > \eta} L(g(x_i^{bN}, -1) (1 - \hat{\sigma}(x_i^{bN})) \]

### 4.6 unbiased

\[
\mathbb{E} [\hat{R}_{PU_{c}}(g)]
\]

\[
= \mathbb{E} [\pi \hat{R}_P(g, +1) + \hat{R}_U(g, -1) - \pi \hat{R}_P(g, -1)]
\]

\[
= \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{e(x_i^{P})} \left( L(g(x_i^{P}, +1) - L(g(x_i^{P}, -1)) \right) + \frac{1}{n} \sum_{i=1}^{n} L(g(x_i), -1) \right]
\]

\[
= \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{e(x_i^{P})} L(g(x_i^{P}, +1) + 1 - \frac{1}{e(x_i^{P})} L(g(x_i^{P}, -1) + \frac{1}{n} \sum_{i=1}^{n} (1 - s_i) L(g(x_i), -1) \right]
\]

\[
= \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} s_i \frac{1}{e(x_i^{P})} L(g(x_i^{P}, +1) + s_i \left( 1 - \frac{1}{e(x_i^{P})} \right) L(g(x_i^{P}, -1) + (1 - s_i) L(g(x_i), -1) \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} y_i e_i \frac{1}{e(x_i)} L(g(x_i), +1) + y_i e_i \left( 1 - \frac{1}{e(x_i)} \right) L(g(x_i), -1) + (1 - y_i e_i) L(g(x_i), -1) \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} y_i L(g(x_i), +1) + y_i (e_i - 1) L(g(x_i), -1) + (1 - y_i e_i) L(g(x_i), -1) \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} y_i L(g(x_i), +1) + (1 - y_i) L(g(x_i), -1) \right]
\]

\[
= R_{PN}(g)(y) \right]
\]

(1)

The change of \( \hat{R}_{PUD}(g) \) will be no more than \( L_{\text{max}}/n \) if some \( x_i \in \chi_P \cup \chi_U \) is replaced, and McDiarmid’s inequality gives us:

\[
\mathbb{P} \{|\hat{R}_{PUD}(g) - R_{PN}(g)(y)| \geq \epsilon\} = \mathbb{P} \{|\hat{R}_{PUD}(g) - \mathbb{E}[\hat{R}_{PUD}(g)]| \geq \epsilon\} \leq 2 \exp \left( -\frac{2\epsilon^2}{n(L_{\text{max}}/n)^2} \right).
\]
And make the right side of the previous formula equal to $\eta$:

$$2 \exp \left( - \frac{2 \epsilon^2}{n(L_{\text{max}}/n)^2} \right) = \eta$$

$$\iff \exp \left( \frac{2 \epsilon^2}{n(L_{\text{max}}/n)^2} \right) = \frac{2}{\eta}$$

$$\iff \frac{2 \epsilon^2}{L_{\text{max}}^2/n} = \ln \left( \frac{2}{\eta} \right)$$

$$\iff 2 \epsilon^2 = \frac{L_{\text{max}}^2 \ln \left( \frac{2}{\eta} \right)}{n}$$

$$\iff \epsilon = \sqrt{\frac{L_{\text{max}}^2 \ln \left( \frac{2}{\eta} \right)}{2n}}$$

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

$$|\hat{R}_{PUe}(g) - R_{PN}(g|y)| = |\hat{R}_{PUe}(g) - \mathbb{E}[\hat{R}_{PUe}(g)]| \leq \frac{L_{\text{max}}^2 \ln \frac{2}{\eta}}{2n}.$$  \hspace{1cm} (2)

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

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

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

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

$$|\hat{R}_{PUe}(g) - \hat{R}_{PU}(g)|$$

$$= \pi |\hat{R}_P(g, +1) - \hat{R}_P(g, -1) - \left( (\hat{R}_P(g, +1) - \hat{R}_P(g, -1)) \right)|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n_P} \left( \frac{1}{e(x_i^P)} - \frac{N_P}{n_P} \right) L(g(x_i^P), +1) \right) + \frac{1}{n} \sum_{i=1}^{n_P} \left( \frac{1}{e(x_i^P)} - \frac{N_P}{n_P} \right) L(g(x_i^P), -1)$$

$$\leq \frac{2}{n} N_P L_{\text{max}} = 2 \pi L_{\text{max}}$$  \hspace{1cm} (5)

Then, we can prove:

$$|R_{PN}(g|y) - \hat{R}_{PU}(g)|$$

$$\leq |R_{PN}(g|y) - \hat{R}_{PUe}(g)| + |\hat{R}_{PUe}(g) - \hat{R}_{PU}(g)|$$

$$\leq 2 \pi L_{\text{max}} + \sqrt{\frac{L_{\text{max}}^2 \ln \frac{2}{\eta}}{2n}}.$$  \hspace{1cm} (6)