Positive-Unlabeled Classification under Class Prior Shift and Asymmetric Error

THE UNIVERSITY OF TOKYO

Nontawat Charoenphakdee[1,2] Masashi Sugiyama[2,1]

1: The University of Tokyo 2: RIKEN AIP

RIKEN

Summary

Class prior shift heavily degrades the performance of positive-unlabeled classification **(PU classification)**.

We propose **two frameworks** for solving this problem:

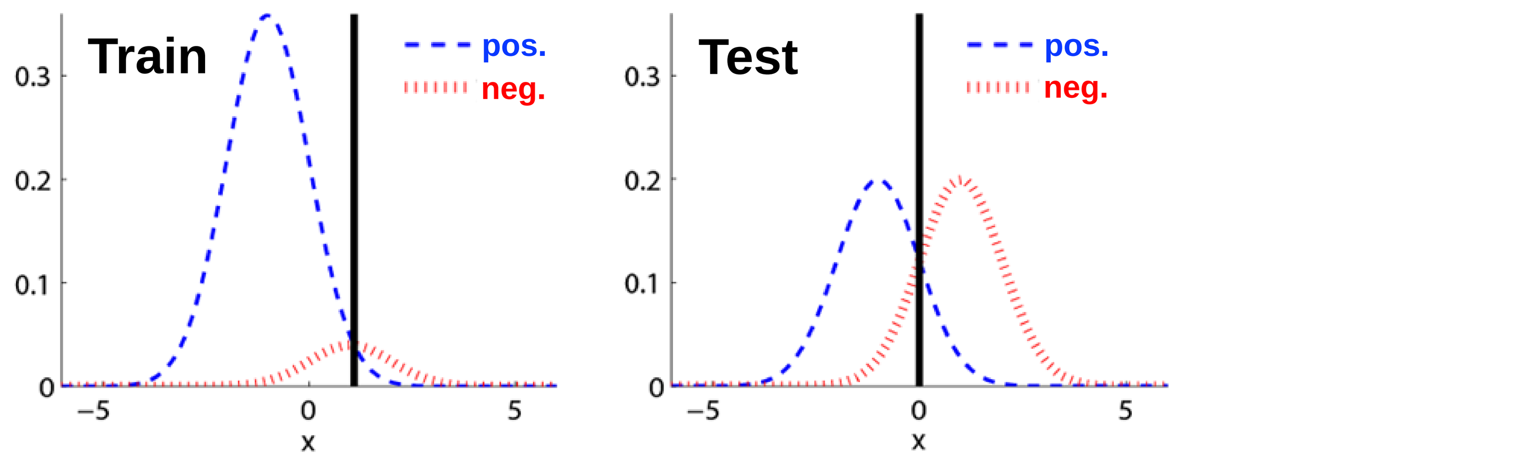
- Risk minimization framework
- Density ratio framework

We prove the **equivalence** of **class prior shift** and **asymmetric error** problems in **PU classification**.

Our methods are applicable for both problems!

Class prior shift

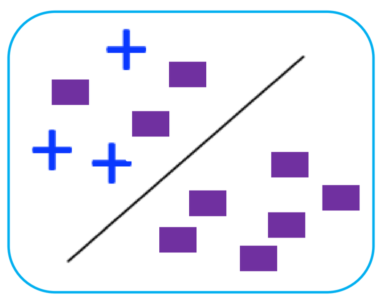
Positive-negative ratio in **training** and **test** data are different.



Decision boundary also shifts → **Lead to low accuracy!**

Practical examples: learn a classifier for **a specific user** from **the internet and many users' information**.

PU classification



Given: Two sets of data

Positive $\mathcal{X}_{\mathrm{P}} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} \mathrm{pos}(\boldsymbol{x})$

Unlabeled $\mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\,\mathrm{neg}(\boldsymbol{x})$

Class prior

$\pi : p(y=1)$
$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$
$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1)$
$\mathbb{E}_{\mathrm{P}}[\cdot] : \underset{\boldsymbol{x}\sim\mathrm{pos}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$
$\mathbb{E}_{\mathrm{N}}[\cdot] : \underset{\boldsymbol{x}\sim\mathrm{neg}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$

Goal: Minimize either

Class prior shift classification risk:

$$R_{\mathrm{Shift}}^{\ell_{0\text{-}1}}(g) = \pi_{\mathrm{te}}\,\mathbb{E}_{\mathrm{P}}\left[\ell_{0\text{-}1}(g(\boldsymbol{x}))\right] + (1-\pi_{\mathrm{te}})\,\mathbb{E}_{\mathrm{N}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right],$$ or

Asymmetric error classification risk:

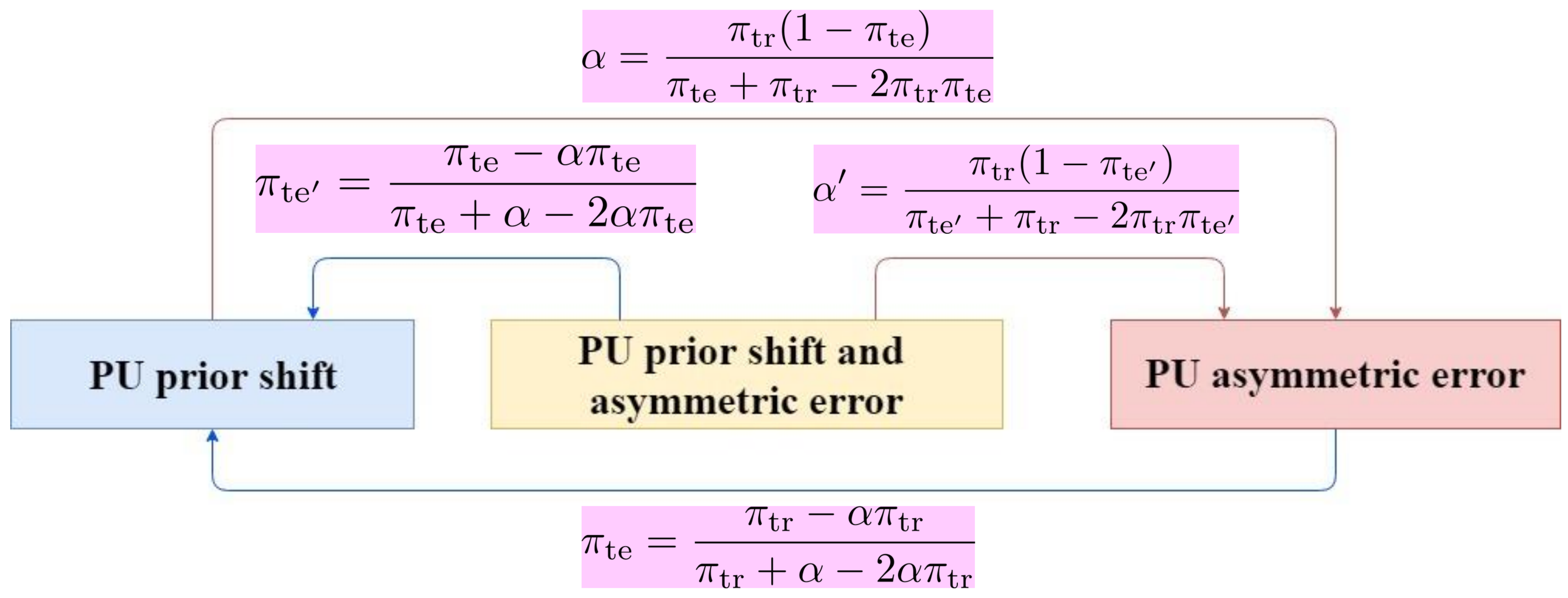
$$R_{\mathrm{Asym}}^{\ell}(g) = (1-\alpha)\pi_{\mathrm{tr}}\,\mathbb{E}_{\mathrm{P}}\left[\ell(g(\boldsymbol{x}))\right] + \alpha(1-\pi_{\mathrm{tr}})\,\mathbb{E}_{\mathrm{N}}\left[\ell(-g(\boldsymbol{x}))\right]$$

$\alpha \in (0,1)$: false negative error

Existing PU classification work: no class prior shift, no asymmetric error (du Plessis+, 2015, Kiryo+, 2017).

Existing class prior shift / asymmetric error work: require positive-negative data (Saerens, 2002, Scott+, 2012, du Plessis+, 2012).

Equivalence of class prior shift and asymmetric error in PU classification



We can relate these problems based on the analysis of Bayes-optimal classifier!

Risk minimization approach

Risk: $R_{\mathrm{Shift}}^{\ell_{0\text{-}1}}(g) = \pi_{\mathrm{te}}\,\mathbb{E}_{\mathrm{P}}\left[\ell_{0\text{-}1}(g(\boldsymbol{x}))\right] + (1-\pi_{\mathrm{te}})\,\mathbb{E}_{\mathrm{N}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right]$

Using the following identity: $\mathbb{E}_{\mathrm{u}}[\cdot] = \pi_{\mathrm{tr}}\,\mathbb{E}_{\mathrm{P}}[\cdot] + (1-\pi_{\mathrm{tr}})\mathbb{E}_{\mathrm{N}}[\cdot]$

We can equivalently express the risk as

$$R_{\mathrm{Shift}}^{\ell_{0\text{-}1}}(g) = \mathbb{E}_{\mathrm{P}}\left[\pi_{\mathrm{te}}\ell_{0\text{-}1}(g(\boldsymbol{x})) - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{1-\pi_{\mathrm{tr}}}\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right] + \frac{1-\pi_{\mathrm{te}}}{1-\pi_{\mathrm{tr}}}\mathbb{E}_{\mathrm{u}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right]$$

Coincides with the existing method (du Plessis+, 2015) when $\pi_{\mathrm{tr}} = \pi_{\mathrm{te}}$.

Minimize empirical risk with surrogate loss (Bartlett+, 2006).

$$\widehat{R}_{\mathrm{PU\text{-}shift}}^{\ell}(g) = \frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\left[\pi_{\mathrm{te}}\ell(g(\boldsymbol{x}_i^{\mathrm{P}})) - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{1-\pi_{\mathrm{tr}}}\ell(-g(\boldsymbol{x}_i^{\mathrm{P}}))\right] + \frac{1}{n_{\mathrm{U}}}\frac{1-\pi_{\mathrm{te}}}{1-\pi_{\mathrm{tr}}}\sum_{j=1}^{n_{\mathrm{U}}}\ell(-g(\boldsymbol{x}_j^{\mathrm{u}}))$$

(regularization can also be added.)

Density ratio approach

Bayes-optimal classifier: $\mathrm{unl}(\boldsymbol{x}) = \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\mathrm{neg}(\boldsymbol{x})$

$$f_{\mathrm{Bayes}}^{*}(\boldsymbol{x}) = \mathrm{sign}\left[p(y=+1|\boldsymbol{x}) - \frac{1}{2}\right]$$

can be equivalently expressed as

$$f_{\mathrm{Bayes}}^{*}(\boldsymbol{x}) = \mathrm{sign}\left[\pi_{\mathrm{tr}}\boxed{\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}} - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{\pi_{\mathrm{te}}+\pi_{\mathrm{tr}}-2\pi_{\mathrm{tr}}\pi_{\mathrm{te}}}\right]$$

Another **formulation:** **Density ratio!**

$$f_{\mathrm{Bayes}}^{*}(\boldsymbol{x}) = \mathrm{sign}\left[\frac{\pi_{\mathrm{te}}+\pi_{\mathrm{tr}}-2\pi_{\mathrm{tr}}\pi_{\mathrm{te}}}{(1-\pi_{\mathrm{te}})} - \boxed{\frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}}\right]$$

Q: Which formulation is preferable?

In general, density ratio is **unbounded**. ☹

In **PU classification**, density ratio $\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}$ is bounded.

$0 \le \frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})} \le \frac{1}{\pi_{\mathrm{tr}}}$ Lower and upper bounded ☺

$\pi_{\mathrm{tr}} \le \frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}$ Unbounded from above ☹

Naïve approach: estimate $\widehat{\mathrm{pos}}(\boldsymbol{x})$ and $\widehat{\mathrm{unl}}(\boldsymbol{x})$ separately then calculate $\frac{\widehat{\mathrm{pos}}(\boldsymbol{x})}{\widehat{\mathrm{unl}}(\boldsymbol{x})}$.

Division operation amplifies the estimation error!

More effective direct approach:

unconstrained Least-squares Important Fitting (uLSIF) (Kanamori+, 2012).

Experiments $\pi_{\mathrm{tr}} = 0.7, \pi_{\mathrm{te}} = 0.3$

Datasets: *banana, ijcnn1, MNIST, susy, cod-rna, magic*

Methods:

Density ratio ($\frac{p}{u}$uLSIF, $\frac{u}{p}$uLSIF)

Linear-in input model: Double hinge **(DH-Lin)** and squared **(Sq-Lin)** losses. Kernel model (Ker): Double hinge **(DH-Ker)** and squared **(Sq-Ker)** losses.

Parameter selection: (regularization, kernel width) 5-fold cross-validation.

Dataset	π^{g}	$\frac{u}{p}$uLSIF	$\frac{p}{u}$uLSIF	DH-Lin	DH-Ker	Sq-Lin	Sq-Ker
banana		83.0(1.0)	**86.4 (0.5)**	70.2(0.5)	78.3(1.0)	70.0(0.0)	83.4(0.4)
ijcnn1		70.8(0.6)	**74.2 (0.7)**	70.0(0.1)	69.8(0.2)	71.5(0.3)	69.2(0.5)
MNIST	π'	79.3(0.5)	**81.7 (0.5)**	74.0(1.1)	**82.4 (1.0)**	52.3(1.4)	**83.4 (0.9)**
susy	0.3	74.3(0.5)	**76.0 (0.3)**	72.7(0.6)	70.0(0.0)	**75.5 (1.4)**	74.7(0.7)
cod-rna		82.1(1.0)	82.8(0.8)	**87.3 (0.7)**	77.3(0.8)	**85.2 (1.1)**	80.2(1.0)
magic		71.5(0.7)	**75.8 (0.6)**	72.7(1.1)	70.8(0.4)	**75.0 (1.0)**	72.9(0.7)
banana		84.7(1.1)	**88.7 (0.7)**	54.9(1.4)	81.7(1.6)	53.6(1.2)	83.8(1.3)
ijcnn1		**64.9 (1.4)**	**66.6 (1.0)**	60.4(1.4)	51.6(3.0)	62.2(1.2)	48.2(2.8)
MNIST	0.5	81.9(0.4)	**84.1 (0.6)**	72.5(1.0)	82.5(0.7)	52.9(1.1)	81.9(0.9)
susy		**75.9 (1.1)**	**77.0 (0.6)**	67.5(1.4)	75.5(0.6)	71.6(1.0)	72.8(1.1)
cod-rna		**85.3 (0.7)**	**85.4 (0.5)**	**86.2 (0.7)**	80.1(1.1)	**86.5 (0.9)**	81.2(1.2)
magic		67.6(0.8)	**73.6 (0.9)**	**72.6 (0.7)**	62.4(1.9)	**71.8 (0.7)**	68.9(0.8)
banana		**80.6 (1.3)**	**82.1 (1.1)**	31.8(0.9)	48.9(1.5)	30.0(0.0)	69.9(1.1)
ijcnn1		35.2(1.4)	**42.4 (0.9)**	30.0(0.0)	30.0(0.0)	32.4(0.5)	30.9(0.4)
MNIST	π	**79.9 (0.7)**	72.6(0.6)	71.1(1.1)	64.8(1.1)	64.0(0.6)	74.2(1.0)
susy	0.7	35.6(3.1)	**44.2 (2.9)**	30.0(0.0)	30.0(0.0)	**42.0 (1.5)**	36.8(1.3)
cod-rna		**77.7 (2.2)**	**77.8 (2.1)**	**79.6 (0.7)**	67.8(0.8)	78.2(0.5)	68.3(1.0)
magic		51.6(0.3)	**60.3 (1.5)**	**56.2 (2.7)**	32.8(0.7)	**58.7 (1.4)**	50.1(1.6)

Correct test prior is given

Wrong test prior is given

Traditional PU

Results reported in mean and std. error of accuracy of 10 trials.
Dataset information and more experiments and can be found in the paper.

References

[1] Saerens, Marco, Patrice Latinne, and Christine Decaestecker. "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure." *Neural computation* 14.1 (2002): 21-41.
[2] Scott, Clayton. "Calibrated asymmetric surrogate losses." *Electronic Journal of Statistics* 6 (2012): 958-992.
[3] Du Plessis, Marthinus Christoffel, and Masashi Sugiyama. "Semi-supervised learning of class balance under class-prior change by distribution matching." *Neural Networks* 50 (2014): 110-119.
[4] Du Plessis, Marthinus C., Gang Niu, and Masashi Sugiyama. "Analysis of learning from positive and unlabeled data." *NeurIPS*. 2014.
[5] Du Plessis, Marthinus, Gang Niu, and Masashi Sugiyama. "Convex formulation for learning from positive and unlabeled data." *ICML*. 2015.
[6] Kiryo, Ryuichi, et al. "Positive-unlabeled learning with non-negative risk estimator." *NeurIPS*. 2017.
[7] Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds." *JASA* 101.473 (2006): 138-156.
[8] Kanamori, Takafumi, Taiji Suzuki, and Masashi Sugiyama. "Statistical analysis of kernel-based least-squares density-ratio estimation." *Machine Learning* 86.3 (2012): 335-367.