Accepted to ICML2017

Semi-Supervised Classification based on Classification from Positive and Unlabeled Data

Tomoya Sakai,[1,2] Marthinus Christoffel du Plessis,
Gang Niu,[1,2] and Masashi Sugiyama[2,1]

[1]The University of Tokyo, Japan
[2]RIKEN, Japan

Classification Problem

Identify class or category of data points

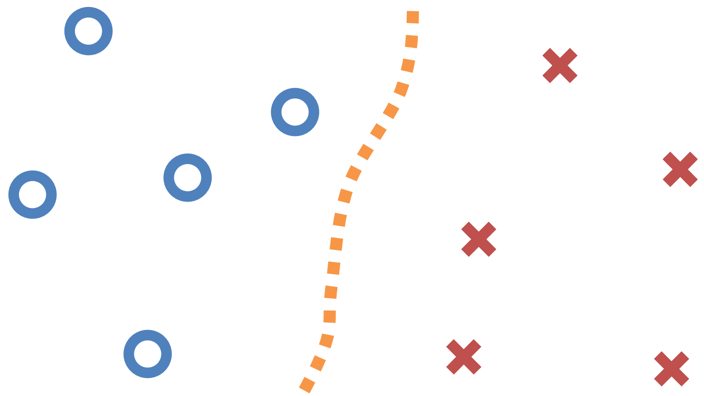
Examples

- Image is a cat (positive class) or not (negative class)

Class: +1 Decision boundary Class: -1

Supervised Learning (PN Learning)

Learn from **labeled** (positive and negative) data



Pattern: $\boldsymbol{x} \in \mathbb{R}^d$

Label: $y \in \{\pm 1\}$

○ : positive (P) data

$$\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x} \mid y = +1)$$

✖ : negative (N) data

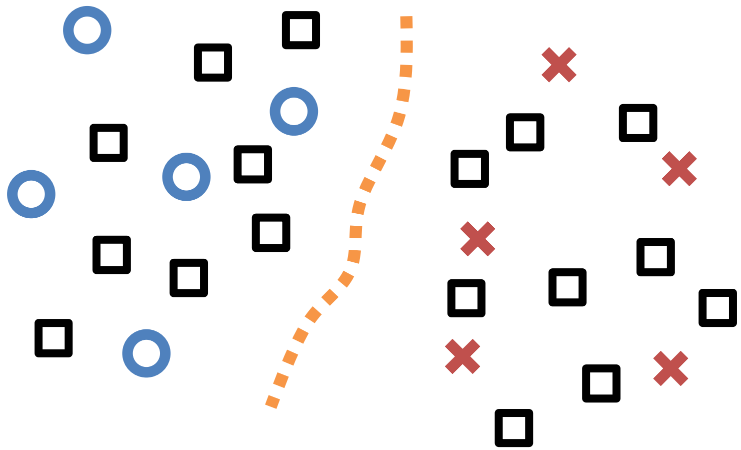
$$\{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x} \mid y = -1)$$

Better performance with many labeled data

Collecting labeled data is costly

Semi-Supervised Learning (SSL)

Learn from a small amount of **labeled** data
and a large amount of **unlabeled** data



: positive data

: negative data

: unlabeled data

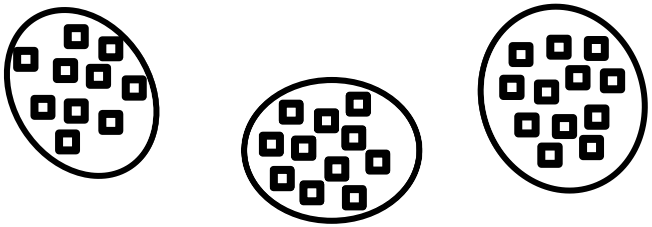
$\{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$

Labeling cost becomes cheap

This Work

Existing methods:

- Require **strong distributional assumptions** for utilizing unlabeled data
  - Ex. the cluster assumption requires the samples in the same cluster be likely to share the same label (Chapelle et al., NIPS, 2002)



  - If the distributional assumptions are not satisfied, the performance of the existing methods decreases

Propose method:

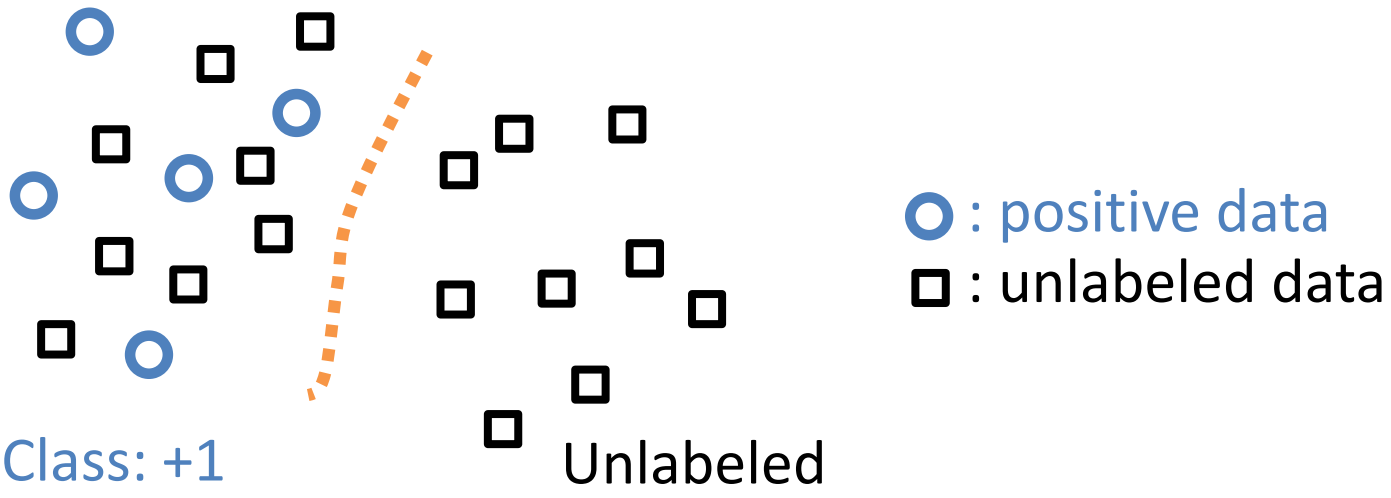
- NOT require the strong distributional assumptions

Outline

1. Introduction
2. Proposed Method
3. Experiment
4. Conclusions

Our Idea: Use of PU Learning

Learning from **positive (P)** and **unlabeled** (U) data

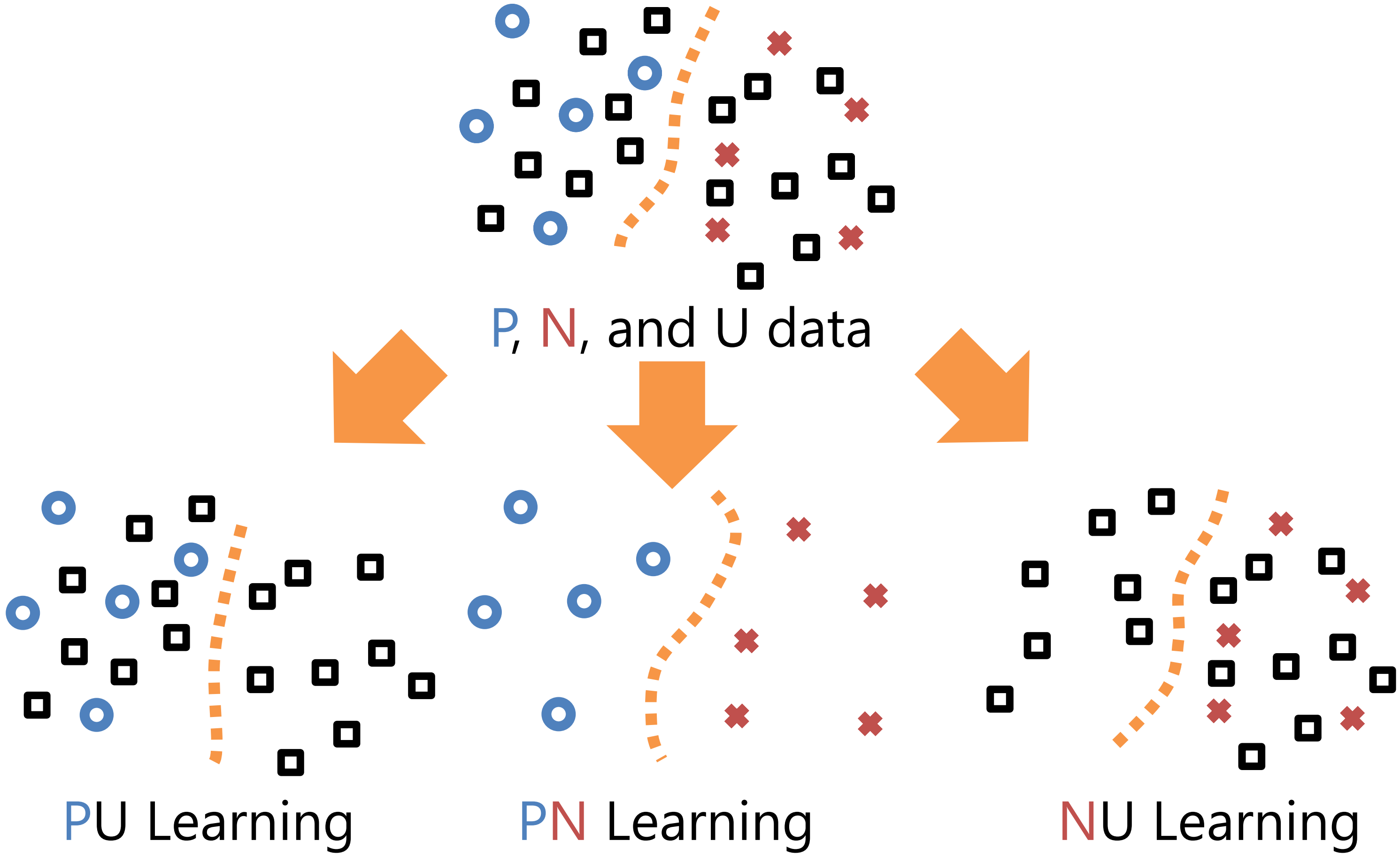



Utilize unlabeled data **without** the distributional assumption (du Plessis et al., NIPS, 2014)

How to use PU learning for semi-supervised learning?

Our Idea (cont.): SSL Decomposition

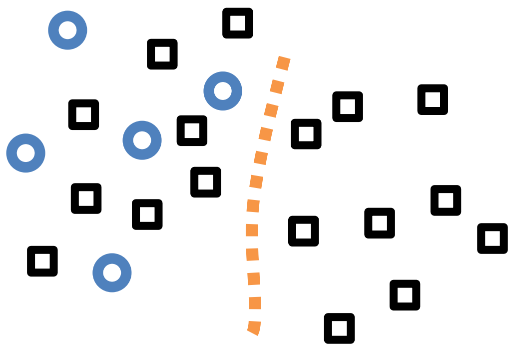
Decompose Semi-Supervised Learning (SSL)
into PU, PN, and NU learning



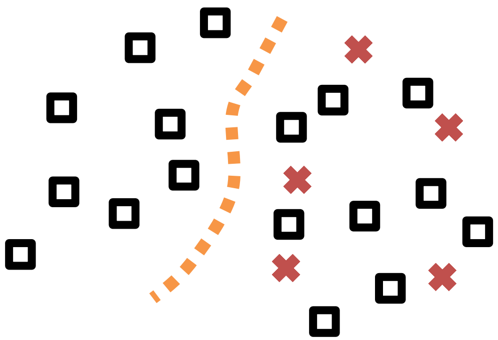
- Reconstruct SSL from PU, PN, and NU learning

PU+NU: PUNU Learning

Since PU and NU are symmetric,
it might be natural to combine **PU** and **NU** learning



PU Learning



NU Learning

Is this really good?

Is PUNU Learning Better?

From Niu et al. NIPS, (2016), we have

(Case I) The size of unlabeled data is sufficiently large

$$\begin{cases} \text{PU} > \text{PN} > \text{NU (classification accuracy)} \\ \text{or} \\ \text{NU} > \text{PN} > \text{PU} \end{cases}$$

- PU or NU is the best

(Case II) The size of unlabeled data is small

$$\begin{cases} \text{PN} > \text{PU} > \text{NU} \\ \text{or} \\ \text{PN} > \text{NU} > \text{PU} \end{cases}$$

- PN is always the best

- PUNU (=PU+NU) is not a good idea
  since it contains the worst one in the combination
- Combine PN with PU/NU would be promising
  (PN+PU or PN+NU is better)

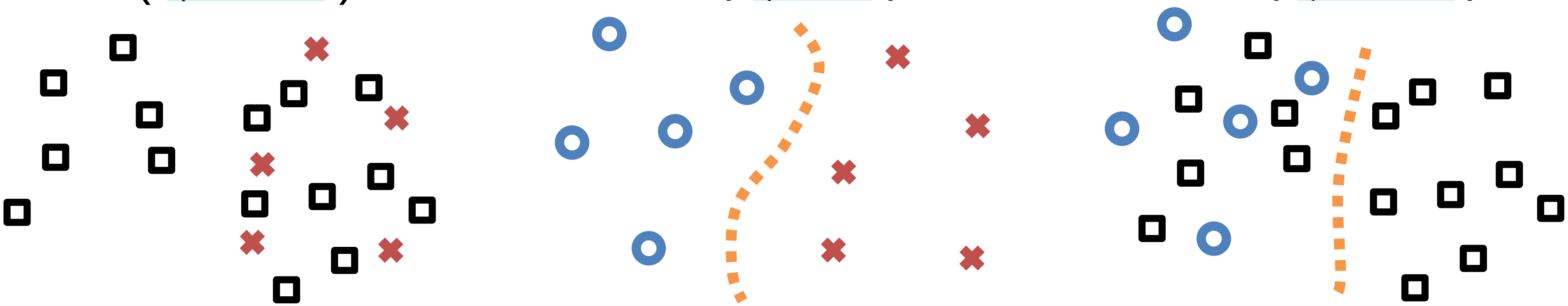
Proposed Method: PNU Learning

Combine PN with PU/NU learning (PN+PU or PN+NU)

NU Learning ($\eta = -1$)

PN Learning ($\eta = 0$)

PU Learning ($\eta = +1$)



N and U data

P and N data

P and U data

- The PNU risk:

$$R^{\eta}_{\mathrm{PNU}}(g) := \begin{cases} (1-\eta)R_{\mathrm{PN}}(g) + \eta R_{\mathrm{PU}}(g) & (\eta \geq 0) \\ (1+\eta)R_{\mathrm{PN}}(g) - \eta R_{\mathrm{NU}}(g) & (\eta < 0) \end{cases}$$

- Obtain the decision rule by minimizing the PNU risk

$\eta \in [-1, 1]$ g: classifier

R_{PN}, R_{PU}, R_{NU}: risks in PN, PU, and NU learning

Comparison with Existing Methods

Existing approach:

- Design regularizer based on the distributional assumption
- Use unlabeled data for regularization

$$\widehat{g} = \underset{g}{\operatorname{argmin}} \ \widehat{R}_{\mathrm{PN}}(g) + \lambda \widehat{W}(g)$$

Entropy regularizer

Manifold regularizer

⋮

Unlabeled data are used

Our approach:

R_{PN} : risk in supervised learning

- Not require the strong distributional assumptions
- Utilize unlabeled data for risk evaluation

$$\widehat{g} = \underset{g}{\operatorname{argmin}} \ \widehat{R}^{\eta}_{\mathrm{PNU}}(g)$$

Labeled and unlabeled data are used

Theoretical Analyses

Without the distributional assumptions, we prove the follows:

- **Generalization error bound**

$$\mathrm{E}_{p(\boldsymbol{x},y)}[\ell_{0\text{-}1}(yg(\boldsymbol{x}))] \leq 2\widehat{R}^{\eta}_{\mathrm{PNU}}(g) + \mathcal{O}_p\left(\frac{1}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{N}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right) \forall g \in \mathcal{G}$$

  - Unlabeled data help reduce the bound
  - Optimal parametric convergence rate

- **Variance reduction**

$\mathrm{Var}[\widehat{R}^{\eta}_{\mathrm{PNU}}(g)] < \mathrm{Var}[\widehat{R}_{\mathrm{PN}}(g)]$ for some η if n_{U} is sufficiently large

  - The PNU risk is stable in terms of the variance
  - More stable cross-validation

$\mathcal{G} = \{g(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x})\rangle \mid \|\boldsymbol{w}\| \leq C_w, \|\boldsymbol{\phi}(\boldsymbol{x})\| \leq C_\phi\}$

R_{PN} : risk in supervised learning $\quad$ $\ell_{0\text{-}1}(m)$: 0-1 loss

Outline

1. Introduction
2. Proposed Method
3. Experiment
4. Conclusions

Image Classification

(smaller is better)

Average with the standard error of misclassification rate $n_L = 100$

Data set	n_U	θ_P	$\widehat{\theta}_P$	PNU	ER	LapSVM	SMIR	WellSVM
Arts	1000	0.50	0.49 (0.01)	**27.4 (1.3)**	**26.6 (0.5)**	**26.1 (0.7)**	40.1 (3.9)	**27.5 (0.5)**
	5000	0.50	0.50 (0.01)	**24.8 (0.6)**	26.1 (0.5)	26.1 (0.4)	30.1 (1.6)	N/A
	10000	0.50	0.52 (0.01)	**25.6 (0.7)**	**25.4 (0.5)**	**25.5 (0.6)**	N/A	N/A
Deserts	1000	0.73	0.67 (0.01)	**13.0 (0.5)**	15.3 (0.6)	16.7 (0.8)	17.2 (0.8)	18.2 (0.7)
	5000	0.73	0.67 (0.01)	**13.4 (0.4)**	**13.3 (0.5)**	16.6 (0.6)	24.4 (0.6)	N/A
	10000	0.73	0.68 (0.01)	**13.3 (0.5)**	**13.7 (0.6)**	16.8 (0.8)	N/A	N/A
Fields	1000	0.65	0.57 (0.01)	**22.4 (1.0)**	26.2 (1.0)	26.6 (1.3)	28.2 (1.1)	26.6 (0.8)
	5000	0.65	0.57 (0.01)	**20.6 (0.5)**	22.6 (0.6)	24.7 (0.8)	29.6 (1.2)	N/A
	10000	0.65	0.57 (0.01)	**21.6 (0.6)**	**22.5 (0.6)**	25.0 (0.9)	N/A	N/A
Stadiums	1000	0.50	0.50 (0.01)	**11.4 (0.4)**	**11.5 (0.5)**	12.5 (0.5)	**17.4 (3.6)**	**11.7 (0.4)**
	5000	0.50	0.50 (0.01)	**11.0 (0.5)**	**10.9 (0.3)**	**11.1 (0.3)**	13.4 (0.7)	N/A
	10000	0.50	0.51 (0.00)	**10.7 (0.3)**	**10.9 (0.3)**	**11.2 (0.2)**	N/A	N/A
Platforms	1000	0.27	0.33 (0.01)	**21.8 (0.5)**	23.9 (0.6)	24.1 (0.5)	30.1 (2.3)	26.2 (0.8)
	5000	0.27	0.34 (0.01)	**23.3 (0.8)**	**24.4 (0.7)**	**24.9 (0.7)**	26.6 (0.3)	N/A
	10000	0.27	0.34 (0.01)	**21.4 (0.5)**	24.3 (0.6)	24.8 (0.5)	N/A	N/A

➢ PNU learning outperforms the existing methods

Class-prior is estimated by
the energy distance minimization method (Kawakubo et al., IEICE-ED, 2016)

* Colored cells indicate the best and comparable method in terms of t-test (sig. lev. 5%)
* The methods taking 2 hours were omitted and indicated as "N/A"

Outline

1. Introduction
2. Proposed Method
3. Experiment
4. Conclusions

World of SSL based on PU Learning

	PN	PU
MR Minimization	Vapnik, Springer, 2000.	du Plessis et al., NIPS, 2014.

MR: misclassification rate

- Obtain classification rules based on MR minimization

SSL: Semi-Supervised Learning

World of SSL based on PU Learning

	PN	PU	Semi-Supervised
MR Minimization	Vapnik, Springer, 2000.	du Plessis et al., NIPS, 2014.	This talk!

MR: misclassification rate

- Obtain classification rules based on MR minimization

SSL: Semi-Supervised Learning

World of SSL based on PU Learning

	PN	PU	Semi-Supervised
MR Minimization	Vapnik, Springer, 2000.	du Plessis et al., NIPS, 2014.	This talk!
AUC Maximization	Herschtal and Raskutti, ICML, 2004.	Tomorrow poster session! Sakai et al., MLJ, 2017.	

AUC: area under the receiver operating characteristic curve

- Useful for imbalanced classification when $n_{\mathrm{P}} \ll n_{\mathrm{N}}$

$$\{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x} \mid y = +1)$$

$$\{\boldsymbol{x}_i^{\mathrm{N}}\}_{i=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x} \mid y = -1)$$

World of SSL based on PU Learning

	PN	PU	Semi-Supervised
MR Minimization	Vapnik, Springer, 2000.	du Plessis et al., NIPS, 2014.	This talk!
AUC Maximization	Herschtal and Raskutti, ICML, 2004.	Tomorrow poster session! Sakai et al., MLJ, 2017.	
SMI Estimation	Suzuki et al., BMC bioinform, 2009.	Sakai et al., arXiv, 2017.	To be explored

SMI: squared-loss mutual information

- Statistical dependency measure
- Useful for dimension reduction, feature selection, independence test, and object matching

World of SSL based on PU Learning

	PN	PU	Semi-Supervised
MR Minimization	Vapnik, Springer, 2000.	du Plessis et al., NIPS, 2014.	This talk!
AUC Maximization	Herschtal and Raskutti, ICML, 2004.	Tomorrow poster session! Sakai et al., MLJ, 2017.	
SMI Estimation	Suzuki et al., BMC bioinform, 2009.	Sakai et al., arXiv, 2017.	To be explored

Code available at

https://github.com/t-sakai-kure/PNU