Classification from Pairwise Similarity and Unlabeled Data

Han Bao[†‡] Gang Niu[‡] Masashi Sugiyama[‡†]

[†] The University of Tokyo
[‡] RIKEN AIP

Abstract

Supervised classification requires vast amount of labeled data, which is costly. In order to mitigate this bottleneck, we proposed a classification problem where only pairwise similarity (two examples belong to the same class) and unlabeled data points are needed.

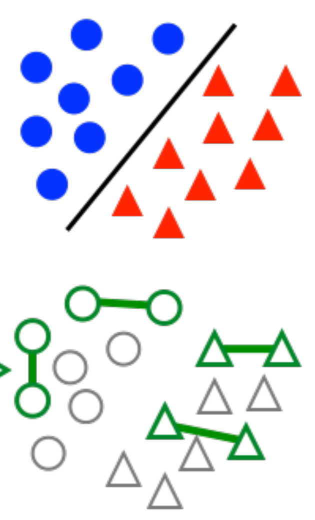
Introduction

Pairwise Data in Classification

Supervised classification:
Explicit labels might be difficult to obtain...

Instead,

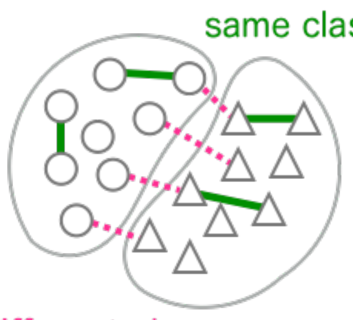
two people share the same property



Example: classification of sensitive matters
(e.g., politics, religion, opinion on racial issue)
⇒ "Which person do you share the same belief as?"

Related Work: Semi-supervised Clustering[1]

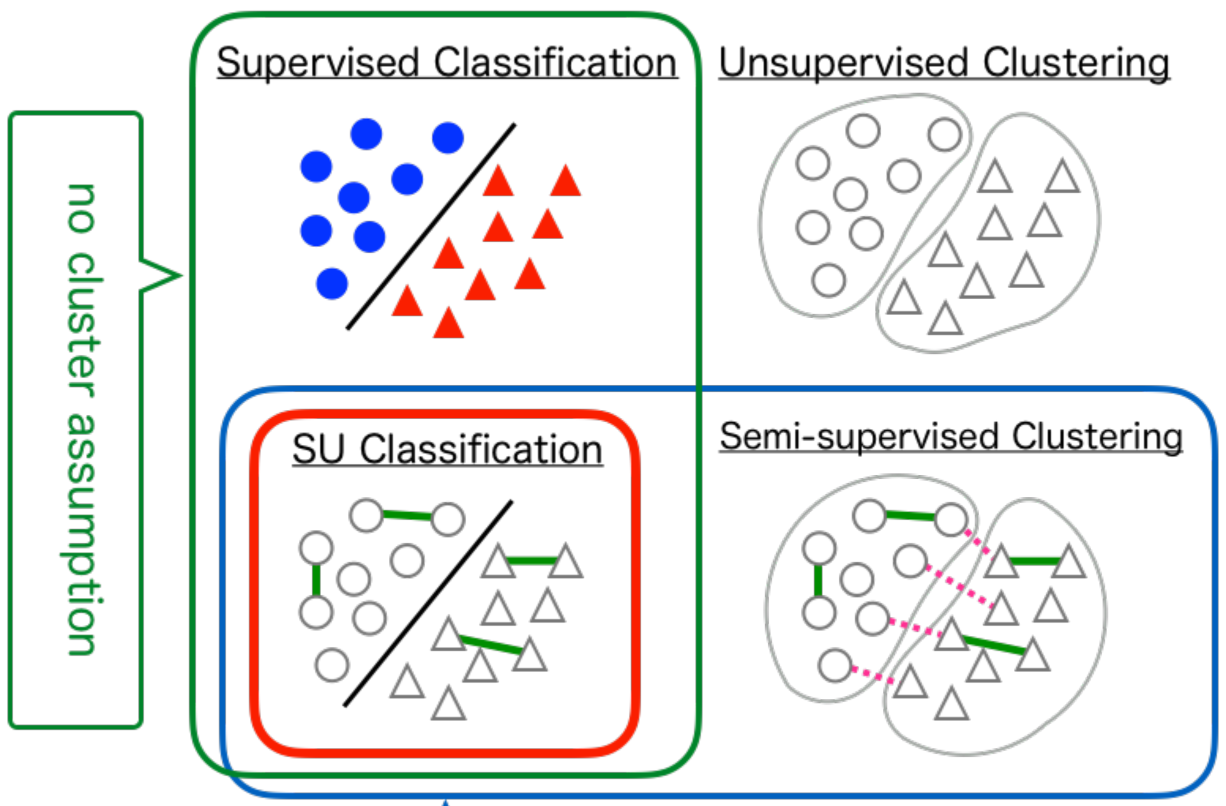
Clustering with
- Similar pair (must-link)
- Dissimilar pair (cannot-link)



Problem

Strong assumption on datasets
e.g., cluster assumption, manifold assumption

Goal of This Work



★Classify data using pairwise information w/o strong assumption

Empirical Risk Minimization

Supervised classification = Minimize misclassification rate

Classification Risk $R_{\mathrm{PN}}(f) = \mathbb{E}\left[\ell_{0\text{-}1}(yf(\boldsymbol{x}))\right]$

↑ unbiased

Empirical Risk $\widehat{R}_{\mathrm{PN}}(f) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell_{0\text{-}1}(y_i f(\boldsymbol{x}_i))}_{\text{Risk of labeled data}}$

Similar and Unlabeled Data

S(imilar) Data Pairs

$$\{(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i})\}_{i=1}^{n_{\mathrm{S}}} \sim p(\boldsymbol{x}, \boldsymbol{x}' \mid y = y' = +1 \vee y = y' = -1)$$

U(nlabeled) Data Points

$$\{\boldsymbol{x}_{\mathrm{U},i}\}_{i=1}^{n_{\mathrm{U}}} \sim p(\boldsymbol{x})$$

Unbiased Risk Estimator

$R_{\mathrm{PN},\ell}(f) = \mathbb{E}\left[\ell(yf(\boldsymbol{x}))\right]$ misclassification rate of f

↑ unbiased

$$\widehat{R}_{\mathrm{SU},\ell}(f) = \underbrace{\frac{\pi_{\mathrm{S}}}{n_{\mathrm{S}}}\sum_{i=1}^{n_{\mathrm{S}}}\frac{\mathcal{L}_{\mathrm{S},\ell}(f(\boldsymbol{x}_{\mathrm{S},i})) + \mathcal{L}_{\mathrm{S},\ell}(f(\boldsymbol{x}'_{\mathrm{S},i}))}{2}}_{\text{Risk of similar data}} + \underbrace{\frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}_{\mathrm{U},\ell}(f(\boldsymbol{x}_{\mathrm{U},i}))}_{\text{Risk of unlabeled data}}$$

$\pi_+ \triangleq p(y=+1)$ ℓ : loss function

$\pi_- \triangleq p(y=-1)$ $\pi_{\mathrm{S}} \triangleq \pi_+^2 + \pi_-^2$

$\mathcal{L}_{\mathrm{S},\ell}(z) \triangleq \frac{\ell(z) - \ell(-z)}{2\pi_+ - 1}$ $\mathcal{L}_{\mathrm{U},\ell}(z) \triangleq \frac{-\pi_-\ell(z) + \pi_+\ell(-z)}{2\pi_+ - 1}$

★Empirical risk can be minimized without explicitly labeled data

Prop. 1: Error → 0 asymptotically

Estimation Error Bound

$$\underbrace{R(\hat{f}) - R(f^*)}_{\text{Estimation error of risk of empirical minimizer } \hat{f}} = \mathcal{O}_p\left(\underbrace{\frac{1}{\sqrt{2n_{\mathrm{S}}}}}_{\#\text{S data}} + \underbrace{\frac{1}{\sqrt{n_{\mathrm{U}}}}}_{\#\text{U data}}\right)$$

Def. (Rademacher complexity): For function class $\mathcal{H}$,

$$\mathfrak{R}(\mathcal{H}; n, \mu) \triangleq \mathbb{E}_{Z_1,\dots,Z_n \sim \mu}\mathbb{E}_{\sigma}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(Z_i)\right]$$

where $\sigma = (\sigma_1, \dots, \sigma_n)$ are Rademacher variables

Assumption:

$\mathcal{F}$ satisfies the following assumption:

$\exists R \in \mathbb{R}_{>0}$ s.t. $\|f\|_\infty \le R$ $(\forall f \in \mathcal{F})$

$\mathfrak{R}(\mathcal{F}; n, \mu) \le \frac{C_{\mathcal{F}}}{\sqrt{n}}$ $(\forall \mu)$ ($C_{\mathcal{F}}$: constant)

where ρ is Lipschitz constant of ℓ

★Estimation error converges to 0 in the optimal rate[2] w/o strong assumption

Prop. 2: Reduced to convex optimization

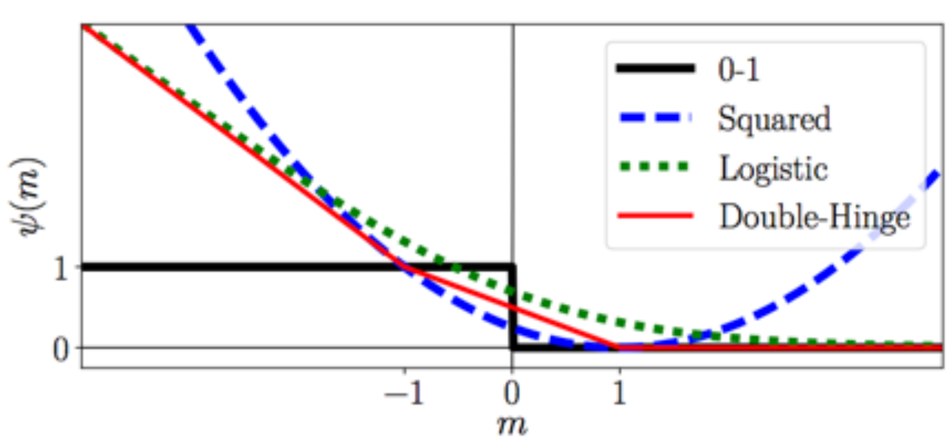
Theorem If ℓ satisfies $\ell(z) - \ell(-z) = -z$

then the optimization of $\widehat{R}_{\mathrm{SU},\ell}(f)$ with the classifier

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) + b$$

and the L2 regularization becomes convex.

Example Squared loss, Double-hinge loss



★Computationally efficient to optimize / Unique global optimum

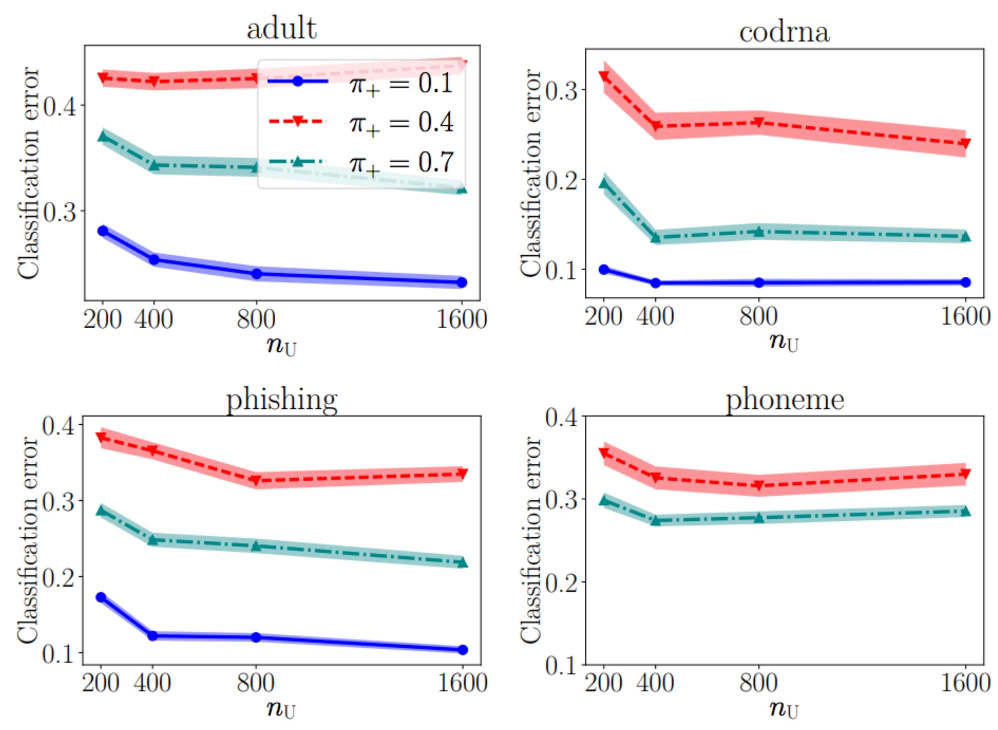
Experiments

Relationship between # of data and error

Setting

- linear-in-input model with squared loss
- regularization parameter ∈{10^{-1}, 10^{-4}, 10^{-7}} (chosen by cross-validation)
- # of S data = 200
- # of U data ∈ {200, 400, 800, 1600}
- true class-prior ∈ {0.1, 0.4, 0.7}

Result



Comparison with Baselines

Setting

- loss: squared / double-hinge loss
- # of S data = # of U data = 500
- true class-prior = 0.7
- (model, regularization are the same with the above)

Baselines

- SERAPH[3]
  clustering with semi-supervised metric learning
- 3SMIC[4]
  semi-supervised clustering with information maximization
- DirtyIMC[5]
  clustering by matrix completion

Results

	SU(proposed)		Baselines		
Dataset	Squared	Double-hinge	SERAPH	3SMIC	DIMC
adult	66.3 (1.2)	**84.5 (0.8)**	66.5 (1.7)	58.5 (1.3)	63.7 (1.2)
banana	64.1 (1.7)	**68.2 (1.2)**	55.0 (1.1)	61.9 (1.2)	64.3 (1.0)
cod-rna	**82.5 (1.1)**	71.0 (0.9)	62.5 (1.4)	56.6 (1.2)	63.8 (1.1)
higgs	54.9 (1.6)	**69.3 (0.9)**	63.4 (1.1)	57.0 (0.9)	65.0 (1.1)
ijcnn1	68.2 (1.3)	**73.6 (0.9)**	59.8 (1.2)	58.9 (1.3)	66.2 (2.2)
magic	**65.9 (1.5)**	**69.0 (1.3)**	55.0 (0.9)	59.1 (1.4)	63.1 (1.1)
phishing	75.2 (1.3)	**91.3 (0.6)**	62.4 (1.1)	60.1 (1.3)	64.8 (1.4)
phoneme	**68.0 (1.4)**	**70.8 (1.0)**	**69.1 (1.4)**	61.3 (1.1)	64.5 (1.2)
spambase	69.5 (1.3)	**85.5 (0.5)**	65.4 (1.8)	61.5 (1.3)	63.6 (1.3)
susy	60.7 (1.0)	**74.8 (1.2)**	58.0 (1.0)	57.1 (1.2)	65.2 (1.0)
w8a	60.5 (1.2)	**86.5 (0.6)**	N/A	60.5 (1.5)	65.0 (2.0)
waveform	78.6 (1.6)	**87.0 (0.5)**	56.5 (0.9)	56.5 (0.9)	65.0 (0.9)

Classification accuracies for each dataset (with standard errors). Bold-faces are outperforming methods chosen by 5% t-test.

References

[1] Klein, D., Kamvar, S. D., and Manning, C. D. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*, 2002.

[2] Mendelson, S. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 2008.

[3] Niu, G., Dai, B., Yamada, M., and Sugiyama, M. Information-theoretic semi-supervised metric learning via entropy regularization. In *ICML*, 2012.

[4] Calandriello, D., Niu, G., and Sugiyama, M. Semi-supervised information-maximization clustering. *Neural Networks*, 2014.

[5] Chiang, K.-Y., Hsieh, C.-J., and Dhillon, I. S. Matrix completion with noisy side information. In *NIPS*, 2015.

東京大学 THE UNIVERSITY OF TOKYO

RIKEN