Classification from Pairwise Similarity and Unlabeled Data

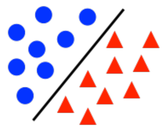
Han Bao[1,2] Gang Niu[2] Masashi Sugiyama[2,1]
[1]The University of Tokyo, Japan / [2]RIKEN, Japan

arXiv: 1802.04381

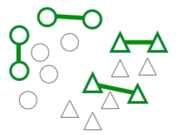
Introduction

Pairwise Data in Classification

Supervised Classification
Explicit labels might be difficult to obtain...
(e.g. politics, religion, racial issue)



Pairwise Information
mitigate data collection cost
⇒ "Which person do you share the same property?" (easier to answer)
cf. randomized response technique



two people share the same property

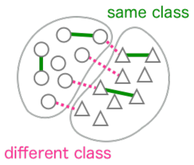
Training Data in Hand

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

Related Work: Semi-supervised Clustering

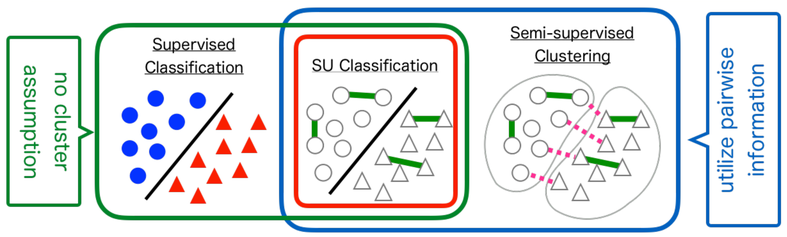
Clustering with
- Similar pair (must-link) = same class
- Dissimilar pair (cannot-link) = different class



Problem

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

Our Target



Do classification from weak supervision (here pairwise information) without strong assumptions

Empirical Risk and Unbiased Estimator

Empirical Risk Minimization

Supervised classification = Minimize misclassification rate

Classification Risk $R_{\mathrm{PN}}(f) = \mathbb{E}\left[\ell_{0\text{-}1}(yf(\boldsymbol{x}))\right]$ ◀····· $\widehat{R}_{\mathrm{PN}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_{0\text{-}1}(y_i f(\boldsymbol{x}_i))$ Empirical Risk

unbiased

Risk of labeled data

SU Unbiased Risk Estimator

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

unbiased

Empirical risk can be accessed via SU data (without explicitly labeled data!)

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

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

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

Property 1: Estimation Error → 0 asymptotically

Estimation Error Bound

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

#S data #U data

Assumptions:
$\mathcal{F}$: model class satisfying $\exists R \in \mathbb{R}_{>0}$ s.t. $\|f\|_\infty \leq R$ $(\forall f \in \mathcal{F})$ and
Rademacher complexity $\mathfrak{R}(\mathcal{F}; n, \mu) \leq \frac{C_\mathcal{F}}{\sqrt{n}}$ $(\forall \mu)$ for a constant $C_\mathcal{F}$
ℓ is ρ-Lipschitz

Estimation error converges to zero in the optimal parametric rate w/o strong assumption

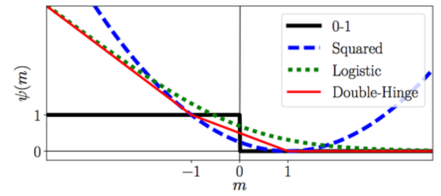
Property 2: Reduced to convex optimization

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, Logistic loss



SU classification can be made computationally efficient with unique solution

Experiments

Setting
- linear-in-input model + {squared, double-hinge} loss
- regularization parameter ∈{10^{-1}, 10^{-4}, 10^{-7}} (chosen by cross-validation)
- # of S data = # of U data = 500
- true class-prior = 0.7

Baselines
- ITML, SERAPH
  clustering w/ semi-supervised metric learning
- 3SMIC
  semi-supervised clustering with information maximization
- DirtyIMC
  clustering by matrix completion
- IMSAT(linear)
  information maximization + regularization making similar data close (implemented with logit model)

Result

		SU(proposed)		Baselines					
Dataset	Dim	Squared	Double-hinge	KM	ITML	SERAPH	3SMIC	DIMC	IMSAT(linear)
adult	123	64.5 (1.2)	**84.5 (0.8)**	58.1 (1.1)	57.9 (1.1)	66.5 (1.7)	58.5 (1.3)	63.7 (1.2)	69.8 (0.9)
banana	2	**67.5 (1.2)**	**68.2 (1.2)**	54.3 (0.7)	54.8 (0.7)	55.0 (1.1)	61.9 (1.2)	64.3 (1.0)	**69.8 (0.9)**
cod-rna	8	**82.8 (1.3)**	71.0 (0.9)	63.1 (1.1)	62.8 (1.0)	62.5 (1.4)	56.6 (1.2)	63.8 (1.1)	69.1 (0.9)
higgs	28	55.1 (1.1)	**69.3 (0.9)**	**66.4 (1.6)**	**66.6 (1.3)**	63.4 (1.1)	57.0 (0.9)	65.0 (1.1)	**69.7 (1.4)**
ijcnn1	22	65.5 (1.3)	**73.6 (0.9)**	54.6 (0.9)	55.8 (0.7)	59.8 (1.2)	58.9 (1.3)	66.2 (2.2)	68.5 (1.1)
magic	10	66.0 (2.0)	**69.0 (1.3)**	53.9 (0.6)	54.5 (0.7)	55.0 (0.9)	59.1 (1.4)	63.1 (1.1)	**70.0 (1.1)**
phishing	68	75.0 (1.4)	**91.3 (0.6)**	64.4 (1.0)	61.9 (1.1)	62.4 (1.1)	60.1 (1.3)	64.8 (1.4)	69.4 (0.8)
phoneme	5	**67.8 (1.5)**	**70.8 (1.0)**	65.2 (0.9)	66.7 (1.4)	**69.1 (1.4)**	61.3 (1.1)	64.5 (1.2)	**69.2 (1.1)**
spambase	57	69.7 (1.4)	**85.5 (0.5)**	60.1 (1.8)	54.4 (1.1)	65.4 (1.8)	61.5 (1.3)	63.6 (1.3)	70.5 (1.1)
susy	18	59.8 (1.3)	**74.8 (1.2)**	55.6 (0.7)	55.4 (0.9)	58.0 (1.0)	57.1 (1.2)	65.2 (1.0)	70.4 (1.2)
w8a	300	62.1 (1.5)	**86.5 (0.6)**	71.0 (0.8)	69.5 (1.5)	0.0 (0.0)	60.5 (1.5)	65.0 (2.0)	70.2 (1.2)
waveform	21	77.8 (1.3)	**87.0 (0.5)**	56.1 (0.8)	54.8 (0.7)	56.5 (0.9)	56.5 (0.9)	65.0 (0.9)	69.7 (1.1)