

ICML2017

正例とラベルなしデータからの分類に 基づく半教師付き分類

(Semi-Supervised Classification Based on
Classification from Positive and Unlabeled Data)

坂井智哉^{1,2} Marthinus C. du Plessis,
Gang Niu¹ 杉山将^{2,1}

¹東京大学 ²理化学研究所

コードが利用可能:

<http://www.ms.k.u-tokyo.ac.jp/software.html#PNU>

分類問題

データ点のクラス（カテゴリー）を明らかにする

具体例

- ウェブサイトがフィッシングサイトか否か
- 画像がネコか否か

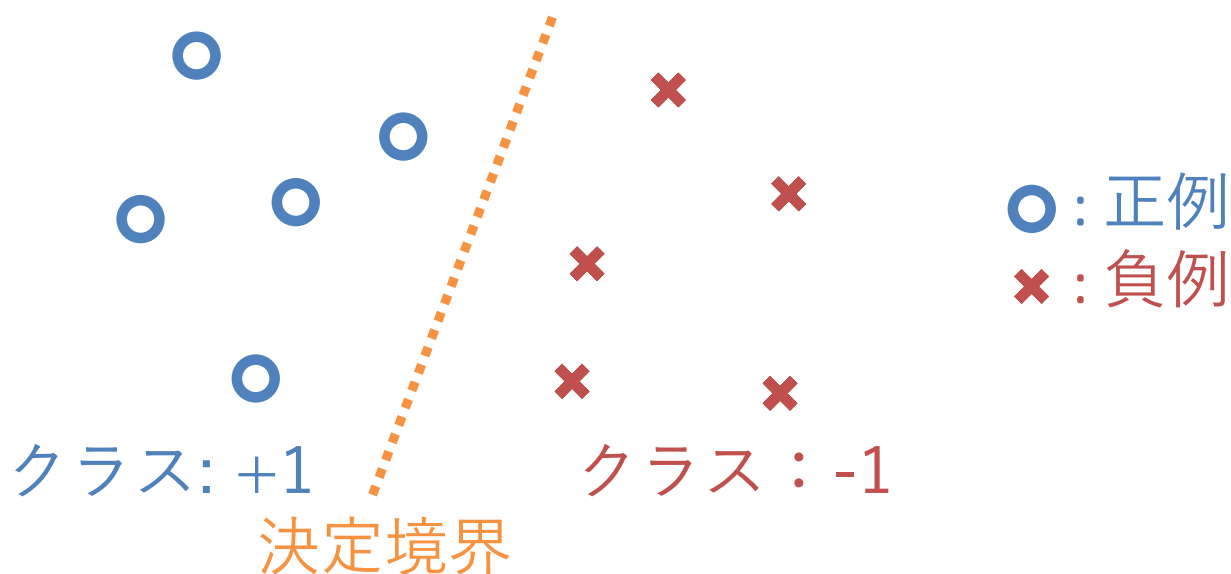


大熊猫

決定境界

教師付き分類

ラベル付きデータ (正例と負例) からの分類

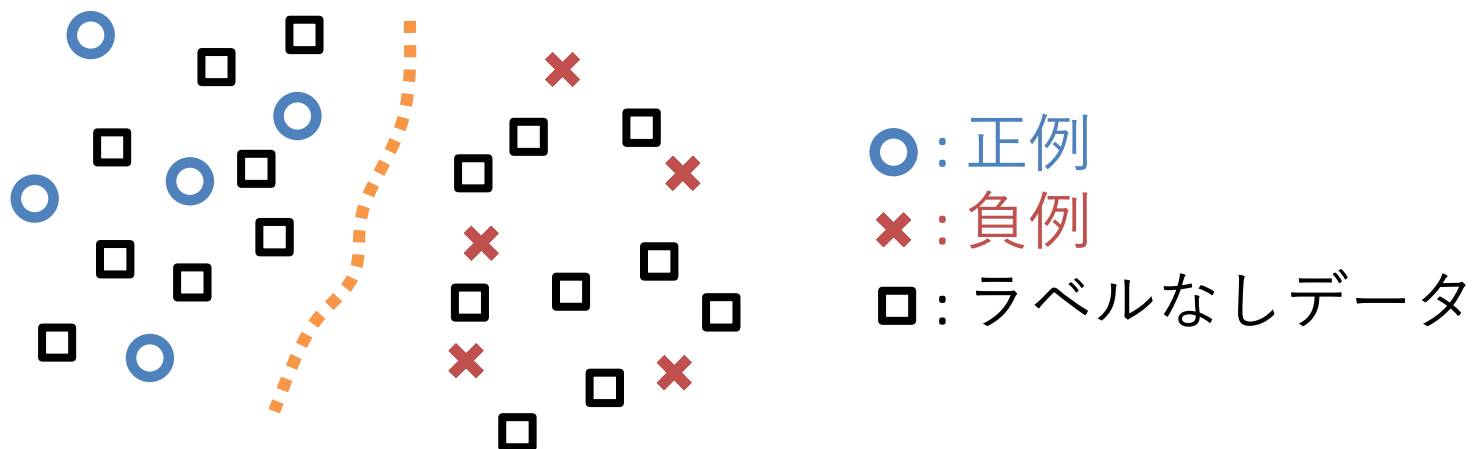


😊 **たくさん**ラベル付きデータがあれば高い分類精度

😞 ラベル付きデータは**高価**

半教師付き分類

ラベル付きデータとラベルなしデータからの分類



😊 ラベル付けの費用が減る

😞 従来法は**分布に対する強い仮定**を必要とする

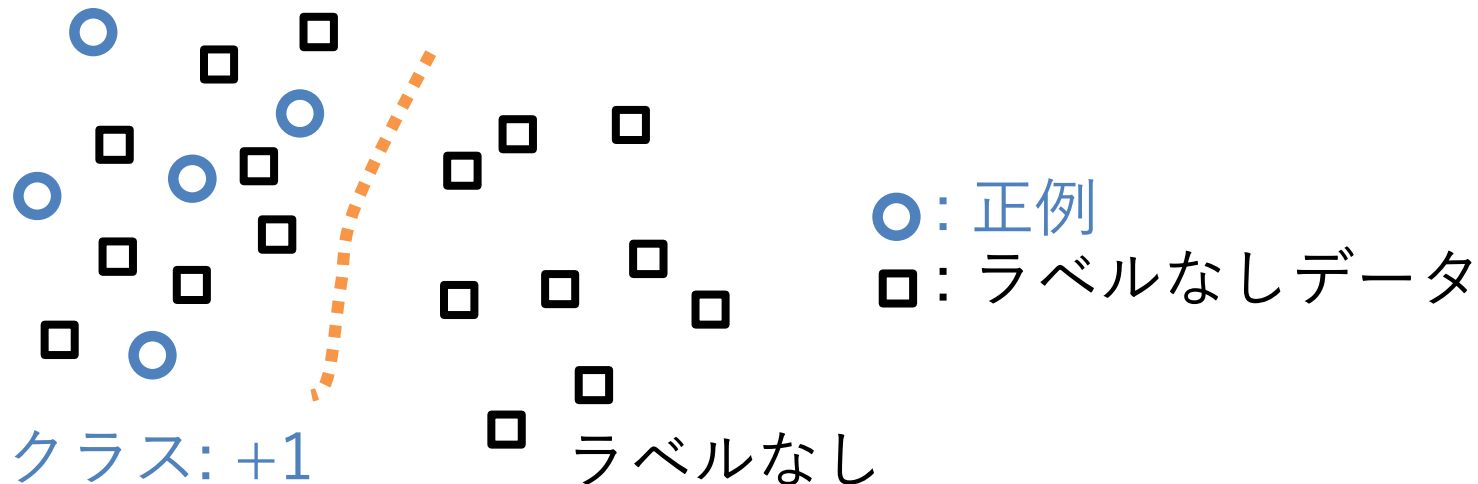
◆例) クラスタ仮定: 同じクラスタに属するデータ点は同じラベルを持つことを要求 (Chapelle et al., NIPS, 2002)

➤ 従来法は、仮定が成り立たないデータで性能が出ない

正例とラベルなしデータからの分類

Positive-Unlabeled Classification

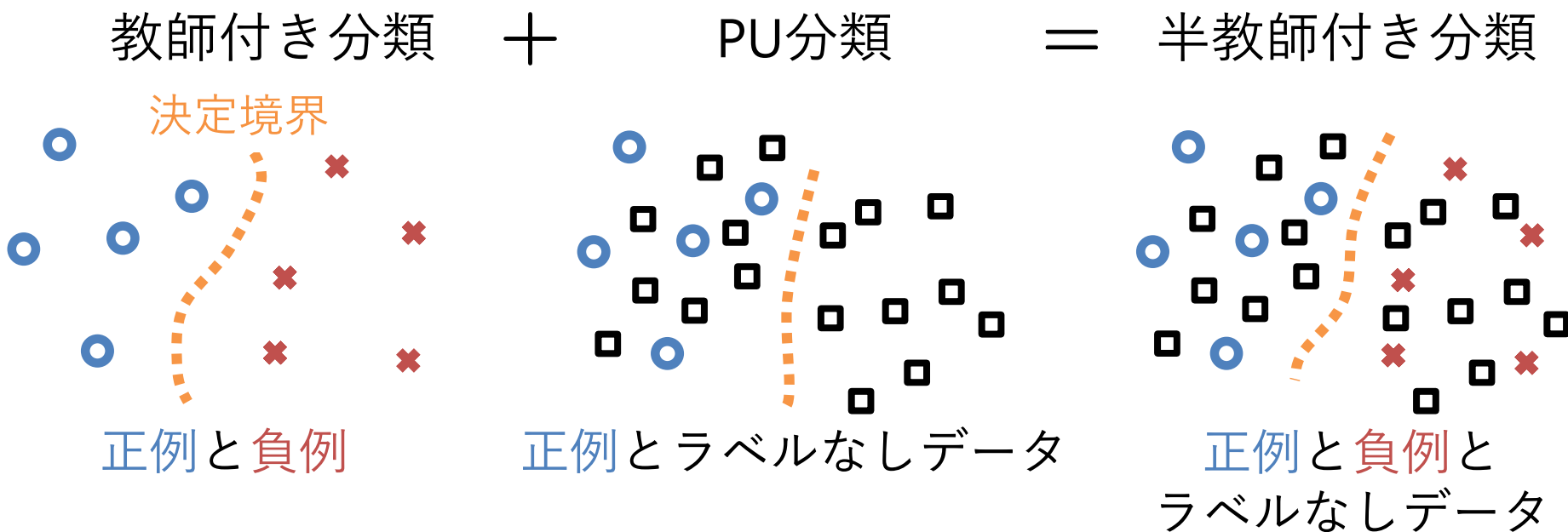
負例がない代わりにラベルなしデータが与えられる



😊 データ分布に対する強い仮定なしでラベルなしデータを利用可能 (du Plessis et al., NIPS, 2014)


提案する方法の概要

教師付き分類とPU分類を組み合わせる



😊 データ分布に対する強い仮定なしにラベルなしデータを利用可能

発表の流れ

1. 背景
2. 問題設定と従来法 
3. 提案法
4. 解析
5. 実験
6. まとめ

問題設定

正例 (P), 負例 (N), ラベルなしデータ (U):

$$\{\mathbf{x}_i^P\}_{i=1}^{n_P} \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x} \mid y = +1)$$

$$\{\mathbf{x}_i^N\}_{i=1}^{n_N} \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x} \mid y = -1)$$

$$\{\mathbf{x}_i^U\}_{i=1}^{n_U} \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}) := \theta_P p(\mathbf{x} \mid y = +1) + \theta_N p(\mathbf{x} \mid y = -1)$$

$$\begin{aligned} \mathbf{x} &\in \mathbb{R}^d \\ y &\in \{\pm 1\} \end{aligned}$$

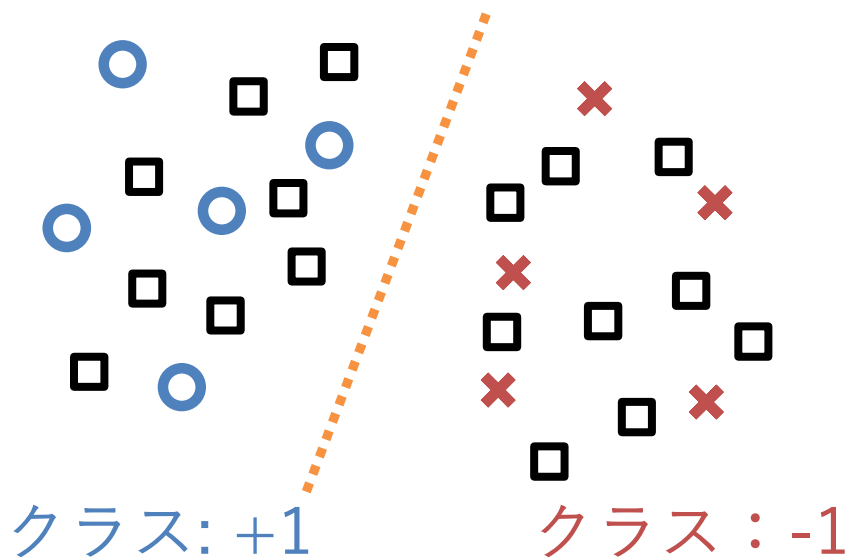
分類器: $g: \mathbb{R}^d \rightarrow \mathbb{R}$ (例 $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$)

$$\theta_P := p(y = +1)$$

$$\theta_N := p(y = -1)$$

$$\theta_P + \theta_N = 1$$

損失関数: $\ell: \mathbb{R} \rightarrow \mathbb{R}$ (例 $\ell(m) = (1 - m)^2$)

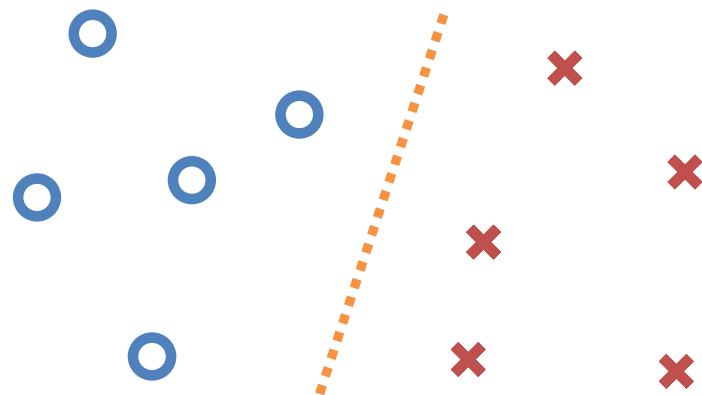


PN分類 (教師付き分類)

正例と負例から計算されるPNリスクを最小化するように分類器を訓練

$$R(g) = E_{p(\mathbf{x}, y)}[\ell(yg(\mathbf{x}))]$$

$$= \theta_P E_P[\ell(g(\mathbf{x}))] + \theta_N E_N[\ell(-g(\mathbf{x}))] := R_{PN}(g)$$



$$E_P[\cdot] := E_{p(\mathbf{x}|y=+1)}[\cdot]$$

$$E_N[\cdot] := E_{p(\mathbf{x}|y=-1)}[\cdot]$$

$$\theta_P = p(y = +1)$$

$$\theta_N = p(y = -1)$$

➤ 実用上は, 標本平均による経験リスクを利用

$$\hat{R}_{PN}(g) = \frac{\theta_P}{n_P} \sum_{i=1}^{n_P} \ell(g(\mathbf{x}_i^P)) + \frac{\theta_N}{n_N} \sum_{j=1}^{n_N} \ell(-g(\mathbf{x}_j^N))$$

$$\{\mathbf{x}_i^P\}_{i=1}^{n_P} \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x} | y = +1) \quad \{\mathbf{x}_i^N\}_{i=1}^{n_N} \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x} | y = -1)$$

PU分類

正例とラベルなしデータから計算されるPUリスクを最小化するように分類器を訓練

$$R_{PN}(g) = \theta_P E_P[\ell(g(\mathbf{x}))] + \theta_N E_N[\ell(-g(\mathbf{x}))]$$

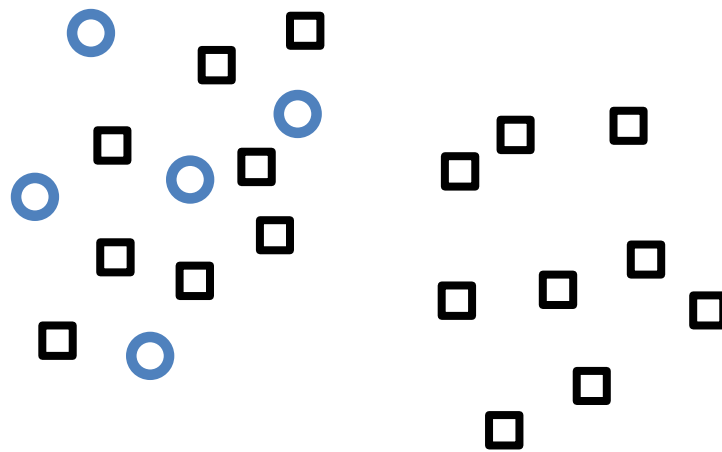
$$E_U[\ell(-g(\mathbf{x}))] = \theta_P E_P[\ell(-g(\mathbf{x}))] + \theta_N E_N[\ell(-g(\mathbf{x}))]$$

PUリスク:

$$R_{PU}(g) := \theta_P E_P[\ell(g(\mathbf{x}))] + E_U[\ell(-g(\mathbf{x}))] - \theta_P E_P[\ell(-g(\mathbf{x}))]$$

$$E_P[\cdot] := E_{p(\mathbf{x}|y=+1)}[\cdot]$$

$$E_U[\cdot] := E_{p(\mathbf{x})}[\cdot]$$



○: 正例

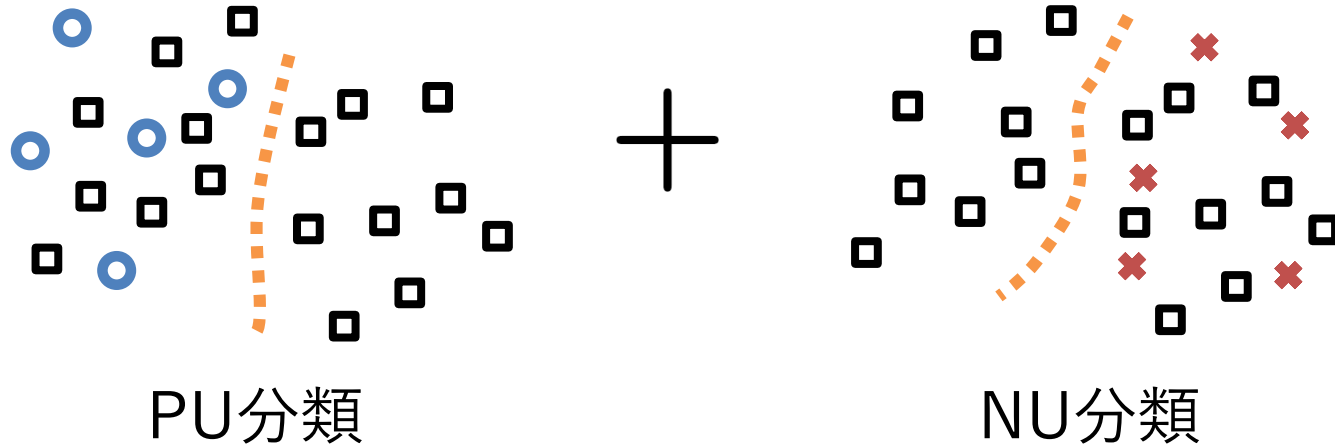
□: ラベルなしデータ

発表の流れ

1. 背景
2. 問題設定と従来法
3. 提案法 
4. 解析
5. 実験
6. まとめ

PUNU分類 = PU+NU

PU分類と対称なNU分類を組み合わせる



- PUNU分類におけるリスク (PUNUリスク):
- $$R_{\text{PUNU}}^{\gamma}(g) := (1 - \gamma)R_{\text{PU}}(g) + \gamma R_{\text{NU}}(g) \quad \gamma \in [0, 1]$$

PNU分類 = PN+PU & PN+NU

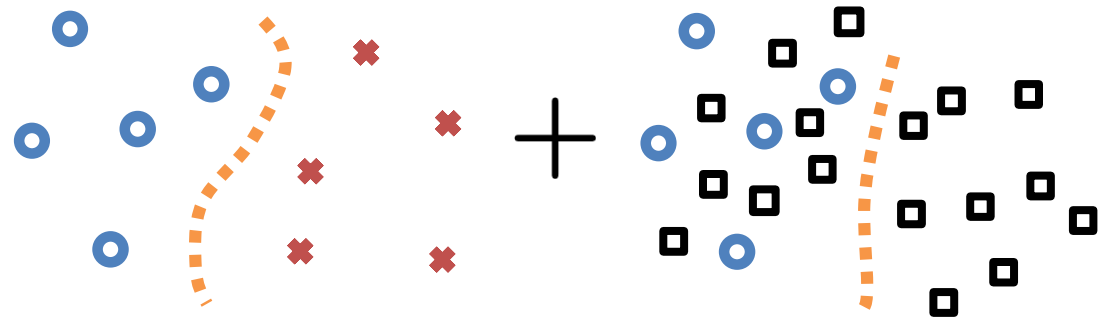
あるいは、**PN**分類と**PU**または**NU**分類を組み合わせる

➤ PNUリスク:

$$R_{\text{PNU}}^{\eta}(g) := \begin{cases} R_{\text{PNPU}}^{\eta}(g) & (\eta \geq 0) \\ R_{\text{PNNU}}^{-\eta}(g) & (\eta < 0) \end{cases} \quad \eta \in [-1, 1]$$

◆ PNPUリスク:

$$R_{\text{PNPU}}^{\gamma}(g) := (1 - \gamma)R_{\text{PN}}(g) + \gamma R_{\text{PU}}(g) \quad \gamma \in [0, 1]$$



◆ PNNUリスク:

$$R_{\text{PNNU}}^{\gamma}(g) := (1 - \gamma)R_{\text{PN}}(g) + \gamma R_{\text{NU}}(g) \quad \gamma \in [0, 1]$$

PUNU vs. PNU分類

理論解析 (Niu et al., 2016) によると,

(場合I)

$$\left\{ \begin{array}{l} R(\hat{g}_{\text{PU}}) < R(\hat{g}_{\text{PN}}) < R(\hat{g}_{\text{NU}}) \\ R(\hat{g}_{\text{NU}}) < R(\hat{g}_{\text{PN}}) < R(\hat{g}_{\text{PU}}) \end{array} \right. \quad n_{\text{U}} \text{ が十分大きいとき}$$

➤ PUまたはNU分類がベスト

(場合II)

$$\left\{ \begin{array}{l} R(\hat{g}_{\text{PN}}) < R(\hat{g}_{\text{NU}}) < R(\hat{g}_{\text{PU}}) \\ R(\hat{g}_{\text{PN}}) < R(\hat{g}_{\text{PU}}) < R(\hat{g}_{\text{NU}}) \end{array} \right. \quad n_{\text{U}} \text{ が小さいとき}$$

➤ PN分類が常にベスト

➤ PNU分類は場合Iと場合IIにおいてベストなもの
の組み合わせであるが、PUNU分類はそうではない

➤ **PNU**分類がPUNU分類よりも有望

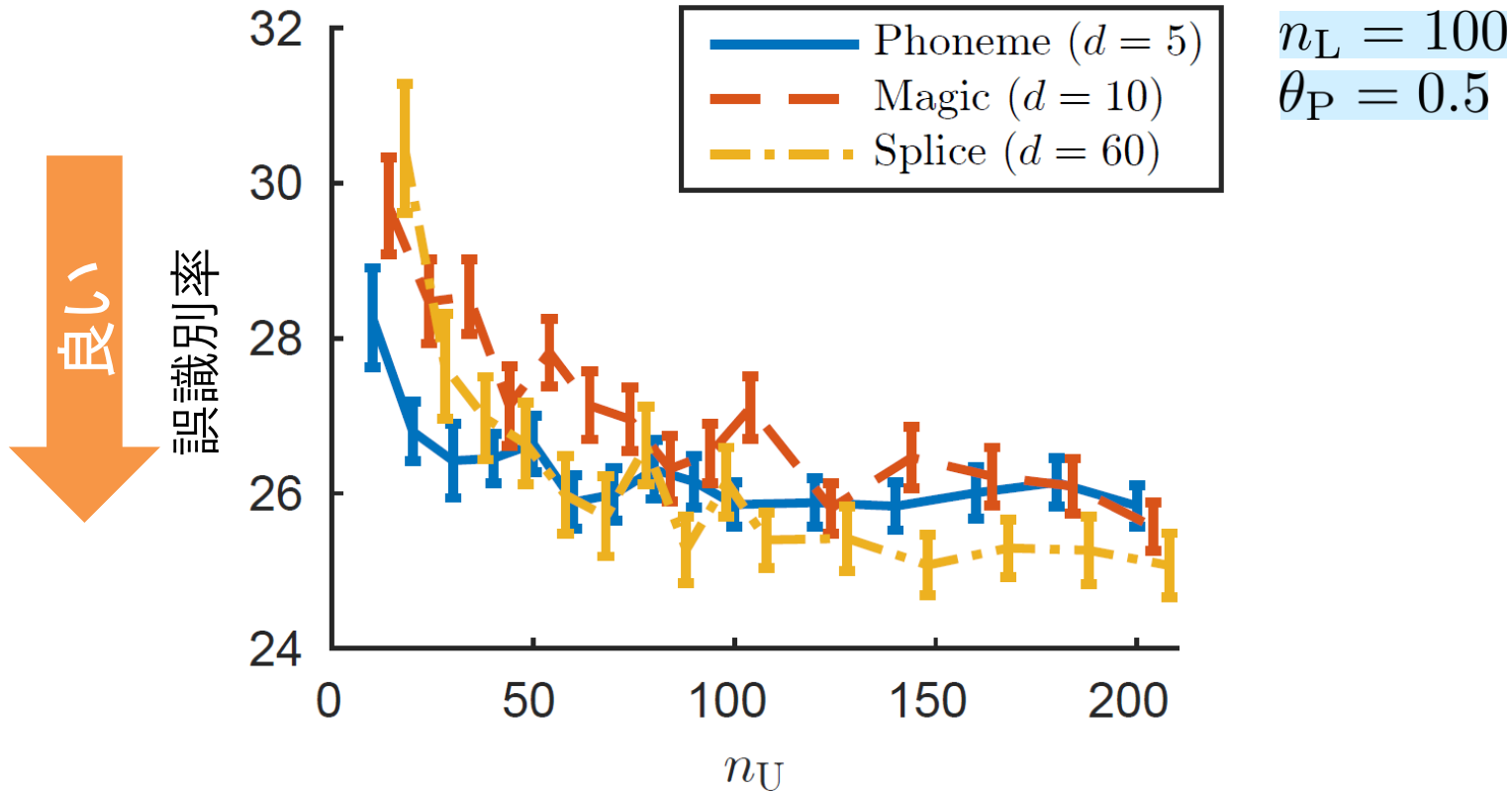
$$\hat{g}_{\text{PN}} := \operatorname{argmin}_{g \in \mathcal{G}} \hat{R}_{\text{PN}}(g) \quad \hat{g}_{\text{PU}} := \operatorname{argmin}_{g \in \mathcal{G}} \hat{R}_{\text{PU}}(g) \quad \hat{g}_{\text{NU}} := \operatorname{argmin}_{g \in \mathcal{G}} \hat{R}_{\text{NU}}(g)$$

発表の流れ

1. 背景
2. 問題設定と従来法
3. 提案法
4. 解析 
5. 実験
6. まとめ

ラベルなしデータの役割

Q. どのようにラベルなしデータが役立つか？



A. n_U が増加するにつれて誤識別率が減少

汎化誤差上界

汎化誤差:

$$I(g) := \mathbb{E}_{p(\mathbf{x}, y)}[\ell_{0-1}(yg(\mathbf{x}))]$$

$$0-1 \text{ 損失: } \ell_{0-1}(m) = \frac{1 - \text{sign}(m)}{2}$$

分布に対する強い仮定なしに, 汎化誤差上界を証明:

どのような $\delta > 0$ に対しても,

$$I(g) \leq 2\hat{R}_{\text{PNPU}}(g) + C_{w, \phi, \delta} \left(\frac{(1 + \gamma)\theta_P}{\sqrt{n_P}} + \frac{(1 - \gamma)\theta_N}{\sqrt{n_N}} + \frac{\gamma}{\sqrt{n_U}} \right)$$

が確率 $1 - \delta$ 以上で, すべての $g \in \mathcal{G}$ に対して成り立つ

➤ ラベルなしデータが汎化誤差上界の減少に役立つ

PNNUとPUNUともに同様の性質を持つ

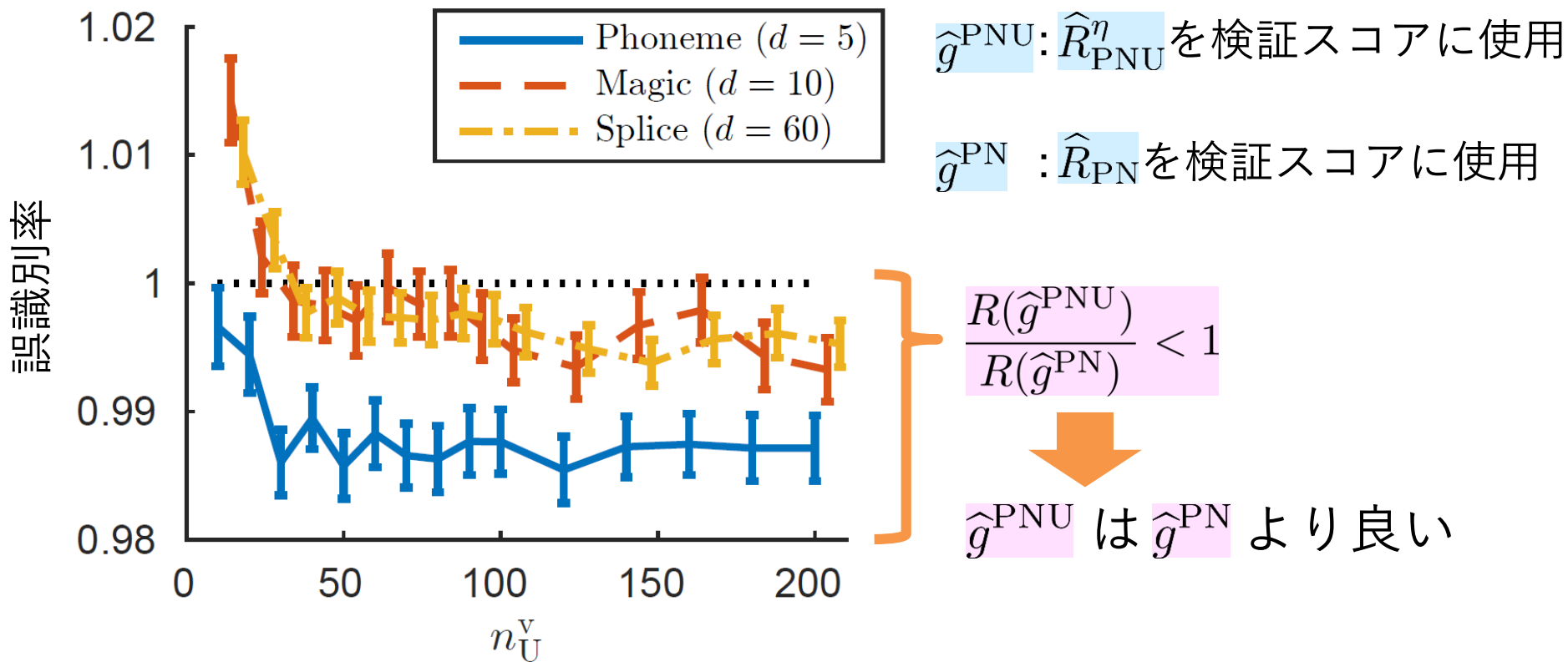
$$(\text{Ref. } R_{\text{PNPU}}^\gamma(g) := (1 - \gamma)R_{\text{PN}}(g) + \gamma R_{\text{PU}}(g) \quad \gamma \in [0, 1])$$

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

$$C_{w, \phi, \delta} := 2C_w C_\phi + \sqrt{2 \ln(3/\delta)}$$

PNUリスクを用いたハイパーパラメータ選択 18

Q. PNリスク $R_{PN}(g)$ と PNUリスク $R_{PNU}^\eta(g)$,
どちらを検証スコアに用いると良いか？



A. PNUリスク $R_{PNU}^\eta(g)$

分散低減効果

分布に対する強い仮定なしに，以下を証明：

$$\text{Var}[\widehat{R}_{\text{PNPU}}^\gamma(g)] < \text{Var}[\widehat{R}_{\text{PN}}(g)] \text{ ただし } \gamma \in (0, 2\gamma_{\text{PNPU}})$$

$$\text{Var}[\widehat{R}_{\text{PNNU}}^\gamma(g)] < \text{Var}[\widehat{R}_{\text{PN}}(g)] \text{ ただし } \gamma \in (0, 2\gamma_{\text{PNNU}})$$

$n_U \rightarrow \infty$ のとき

$$\gamma_{\text{PNPU}} := \underset{\gamma}{\operatorname{argmin}} \text{Var}[\widehat{R}_{\text{PNPU}}^\gamma(g)]$$

$$\gamma_{\text{PNNU}} := \underset{\gamma}{\operatorname{argmin}} \text{Var}[\widehat{R}_{\text{PNNU}}^\gamma(g)]$$

PNUリスクの方がPNリスクよりも分散が小さい (安定)

- 安定なPNUリスクがハイパーパラメータ選択時の
検証スコアに有用 (交差確認法)

$$R_{\text{PNU}}^\eta(g) := \begin{cases} R_{\text{PNPU}}^\eta(g) & (\eta \geq 0) \\ R_{\text{PNNU}}^{-\eta}(g) & (\eta < 0) \end{cases} \quad \eta \in [-1, 1]$$

発表の流れ

1. 背景
2. 問題設定と従来法
3. 提案法
4. 解析
5. **実験** 
6. まとめ

ガウスカーネルモデルを利用:

$$g(\mathbf{x}) = \sum_{i=1}^n w_i \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right)$$

比較手法:

- エントロピー正則化 (**ER**) (Grandvalet & Bengio, NIPS, 2004)
- ラプラシアンSVM (**LapSVM**) (Belkin et al., JMLR, 2006)
- 二乗損失相互情報量正則化 (**SMIR**) (Niu et al., ICML, 2013)
- 弱ラベル付きSVM (**WellSVM**) (Li et al., JMLR, 2013)
- 安全半教師付きSVM (**S4VM**) (Li & Zhou, PAMI, 2015)

誤識別率 (低いほど良い)

平均と標準誤差

$n_U = 300$

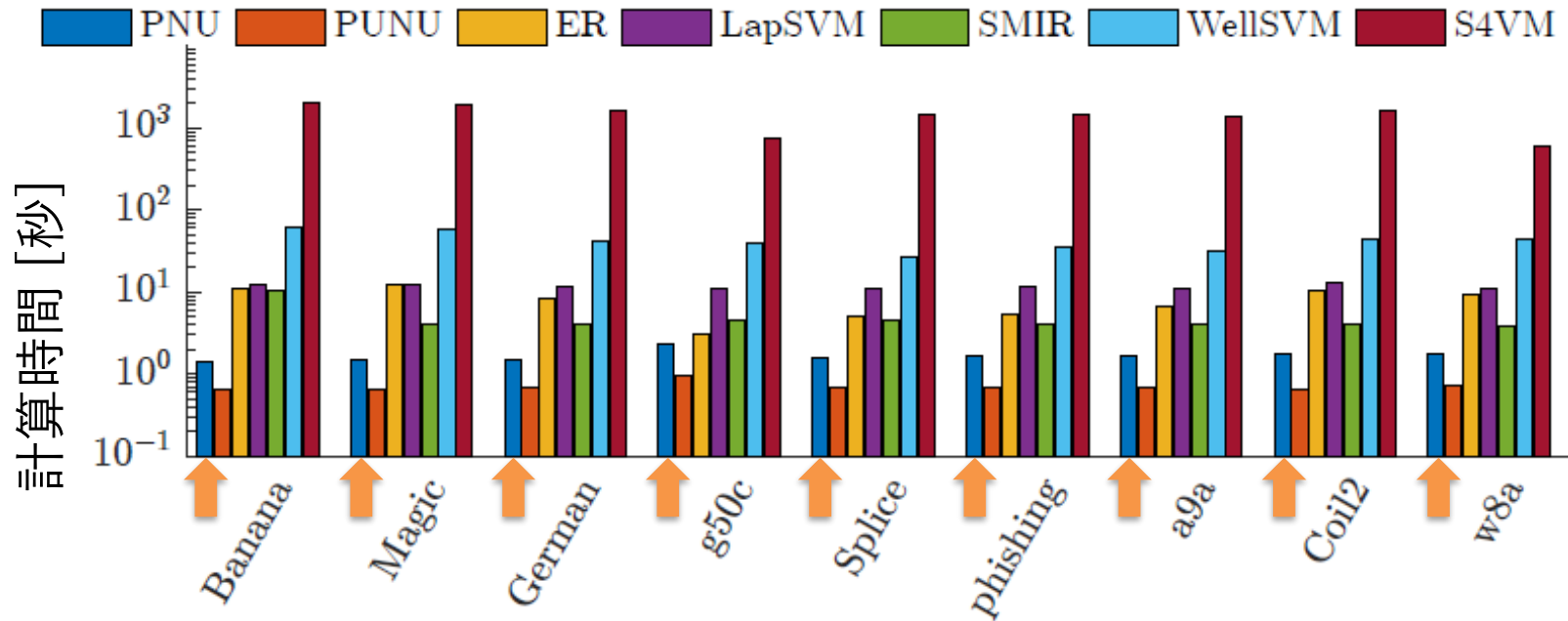
Data set	n_L	PNU	PUNU	ER	LapSVM	SMIR	WellSVM	S4VM
Banana	10	30.1 (1.0)	32.1 (1.1)	35.8 (1.0)	36.9 (1.0)	37.7 (1.1)	41.8 (0.6)	45.3 (1.0)
	$d = 2$ 50	19.0 (0.6)	26.4 (1.2)	20.6 (0.7)	21.3 (0.7)	21.1 (1.0)	42.6 (0.5)	38.7 (0.9)
Magic	10	31.7 (0.8)	34.1 (0.9)	34.2 (1.1)	37.9 (1.3)	36.0 (1.2)	30.1 (0.8)	33.3 (0.9)
	$d = 10$ 50	29.9 (0.8)	33.4 (0.9)	30.9 (0.5)	31.0 (0.9)	30.8 (0.9)	28.8 (0.8)	29.2 (0.4)
German	10	40.8 (0.9)	42.4 (0.7)	43.6 (0.9)	45.9 (0.7)	46.2 (0.8)	42.4 (0.8)	42.0 (0.7)
	$d = 20$ 50	36.2 (0.8)	39.0 (0.8)	38.9 (0.6)	40.6 (0.6)	38.4 (1.1)	38.5 (1.0)	34.9 (0.5)
g50c	10	11.4 (0.6)	12.5 (0.6)	23.3 (2.3)	39.8 (1.6)	21.9 (1.3)	6.6 (0.4)	27.0 (1.4)
	$d = 50$ 50	12.5 (1.1)	10.1 (0.6)	8.7 (0.4)	22.5 (1.5)	10.6 (0.6)	7.4 (0.4)	12.1 (0.5)
Splice	10	38.3 (0.8)	39.3 (0.8)	43.9 (0.8)	47.9 (0.5)	41.6 (0.7)	42.0 (1.0)	42.4 (0.6)
	$d = 60$ 50	30.6 (0.8)	34.7 (0.9)	30.9 (0.8)	38.8 (1.0)	30.6 (0.9)	40.9 (0.8)	35.9 (0.7)
phishing	10	24.2 (1.2)	25.8 (1.0)	27.3 (1.6)	37.2 (1.6)	27.6 (1.6)	27.5 (1.4)	31.7 (1.3)
	$d = 68$ 50	15.8 (0.6)	18.3 (0.8)	15.4 (0.5)	21.1 (1.3)	14.7 (0.8)	17.2 (0.7)	16.7 (0.8)
a9a	10	31.4 (0.9)	31.3 (1.0)	34.3 (1.2)	41.0 (1.1)	37.3 (1.3)	33.1 (1.2)	34.3 (1.2)
	$d = 83$ 50	27.9 (0.6)	29.9 (0.8)	28.6 (0.7)	33.3 (1.0)	26.9 (0.7)	28.9 (0.8)	26.2 (0.4)
Coil2	10	38.7 (0.8)	40.1 (0.8)	42.8 (0.7)	43.9 (0.8)	43.2 (0.8)	39.1 (0.9)	44.0 (0.8)
	$d = 241$ 50	23.2 (0.6)	30.5 (0.9)	23.6 (0.9)	22.8 (0.9)	25.1 (0.9)	22.6 (0.8)	25.4 (0.8)
w8a	10	35.9 (0.9)	33.6 (1.0)	41.6 (1.0)	46.6 (0.8)	39.4 (0.9)	42.1 (0.8)	43.0 (0.8)
	$d = 300$ 50	28.1 (0.7)	27.6 (0.6)	27.0 (0.9)	38.7 (0.8)	28.0 (0.9)	33.7 (0.8)	35.2 (1.0)

➤ PNU分類は分類精度が良い

* 色付きのセルは平均誤識別率が最も低い手法と、それと有意差5%のt検定で同等の手法を示す

計算時間

平均計算時間



➤ PNU分類は計算効率が良い

画像分類での実験設定

二つの似た景色を分類する (データセット: Places 205)
(Zhou et al., NIPS, 2014)

➤ タスクの例

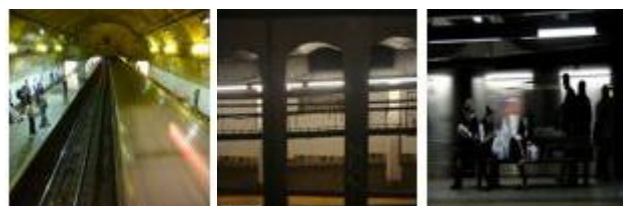
◆ Desert sand



vs. Desert vegetation



◆ Subway station platforms vs. Train station platforms



➤ AlexNetを用いて抽出した特徴ベクトル ($d = 4096$) を使用
入力に関する線形モデルを利用: $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$

誤識別率 (低いほど良い)

平均と標準誤差

 $n_L = 100$

Data set	n_U	θ_P	$\hat{\theta}_P$	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分類は性能が良い

クラス事前確率 θ_P はエネルギー距離最小化に基づいた手法により推定
(Kawakubo et al., IEICE, 2015)

* "N/A"は計算時間が2時間を超えた手法

まとめ

- PU分類に基づく半教師付き分類手法を提案
 - ◆ PNU分類
- 提案したリスク推定量を理論的に解析
 - ◆ 汎化誤差上界
 - ◆ 分散低減効果
- 提案法の有用性を実験的に示した

コードが利用可能:

<http://www.ms.k.u-tokyo.ac.jp/software.html#PNU>