ICLR2019における不完全ラベル学習

Ridge-i inc.
Masanari Kimura (mkimura@ridge-i.com)

About

Education & Career

- 筑波大学卒 (2018)
- Ridge-iエンジニア (2018 ~)
- 産総研特専研究員 (2019 ~)

Twitterやってます

@machinery81

Researches ater joining Ridge-i

- Interpretation of Feature Space using Multi-Channel Attentional Sub-Networks (CVPRW2019)
- Progressive Data Increasing as the Neural Network Initializer (JSAI2019)
- Anomaly Detection Using GANs for Visual Inspection in Noisy Training Data (ACCVW2018)
- Analyzing Centralities of Embedded Nodes (ICDMW2018)

概要

- ICLR2019に採択された不完全ラベル学習のまとめ
- ラベルが不完全な状況での学習という研究領域を知ってもらう

今回紹介する論文たち

- [1] Learning from Positive and Unlabeled Data with a Selection Bias
- [2] On the Minimal Supervision for Training Any Binary Classifier from Only Unlabeled Data
- [3] Multi-Class Classification without Multi-Class Labels

不完全ラベル学習

- 学習に用いられるデータのラベルが欠損しているという問題設定
  - Weakly-Supervised Learningなどとも
- 今回は特に，Kクラス分類の際に，T(<=K)クラスのデータにラベルがついていないケースを考える.
  - e.g, 2値分類でPositiveクラスのデータにしかラベルが無い(PU)

Learning from Positive and Unlabeled Data with a Selection Bias

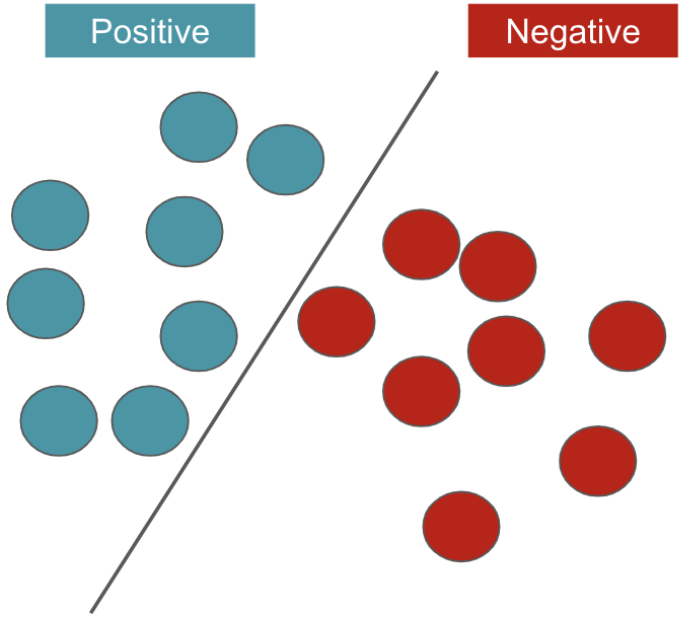
- Masahiro Kato, Takeshi Teshima, Junya Honda

Abstract

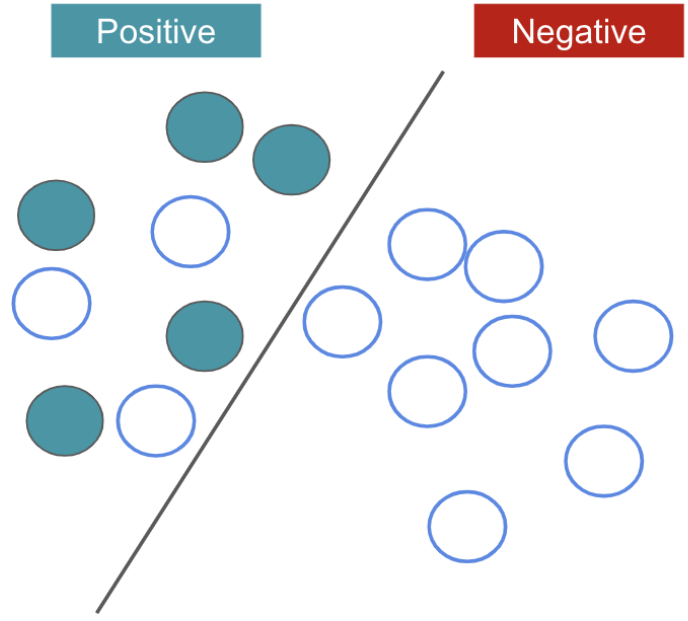
- SCARを仮定せずにpositiveデータとunlabeledデータのみから学習
  - より現実の問題設定に即すようにselection biasを考慮

PU Learning

- positveクラスとのデータとラベル無しデータのみから学習



一般的な教師あり学習



PU学習

Unlabeled data

Selected Completely At Random (SCAR)

- (Assumption) Positiveなラベル付きデータはPositiveなラベル無しデータと同様の分布に属する
  - $\{x_i\}_{i=1}^{n} \sim^{i.i.d.} p(x|y=+1)$

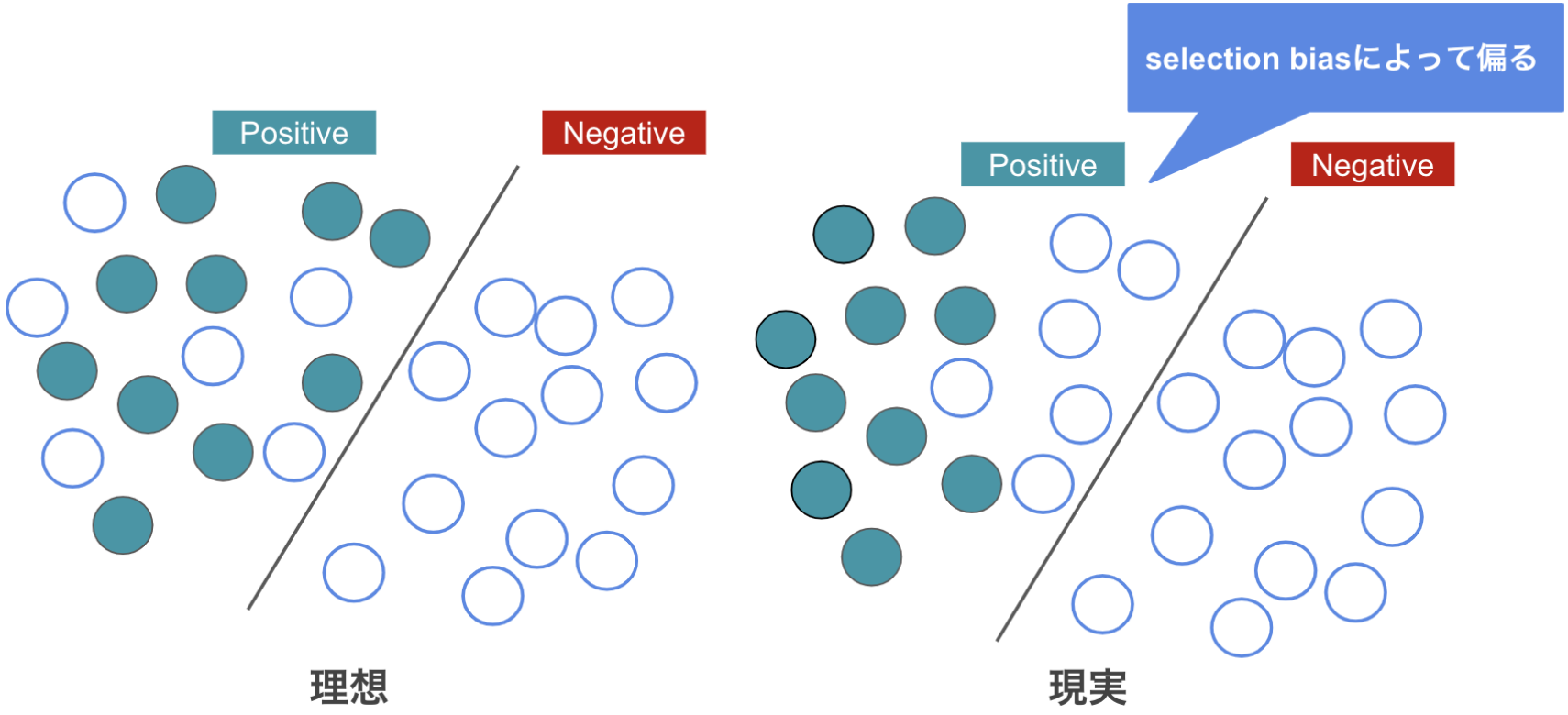
Is SCAR Always True?

Is SCAR Always True?

- 現実問題ではラベリングの際のデータの選択に"バイアス"が掛かる
  - e.g. わかりやすいデータにはラベル付けがされやすい

Selection Bias in the Labeling Process

- ラベリング時のバイアスによって，ラベルのついているpositiveデータとラベルのついていないpositiveデータの分布がずれる
  - 多くの現実の問題設定ではSCARは成り立たない



PU Learning with Selection Bias

本論文の目的：PU Learningのモデルから SCARの仮定を取り去る

- positiveデータ集合$\{x_i\}_{i=1}^{n}$とunlabaledデータ集合$\{x'_i\}_{i=1}^{n'}$
  - $\{x_i\}_{i=1}^{n} \sim^{i.i.d} p(x|y=+1, o=+1),$
  - $\{x'_i\}_{i=1}^{n'} \sim^{i.i.d} p(x),$
- class prior $\pi = p(y=+1)$は既知
  - データ全体のpositiveデータの割合についての事前知識

Identification Strategy

- Elkan & Noto (2008)[4]によって，PU learningに一切の仮定無しに$p(y=+1|x)$を推定することは出来ないことが示されている
  - 一般的にはSCARを仮定
  - $p(x|y=+1,o=+1)=p(x|y=+1,o=0)$

$$p(y=+1|x)=\frac{p(x,y=+1)}{p(x)}=\frac{p(x|y=+1)\pi}{p(x)}=\frac{p(x|y=+1,o=+1)\pi}{p(x)}$$

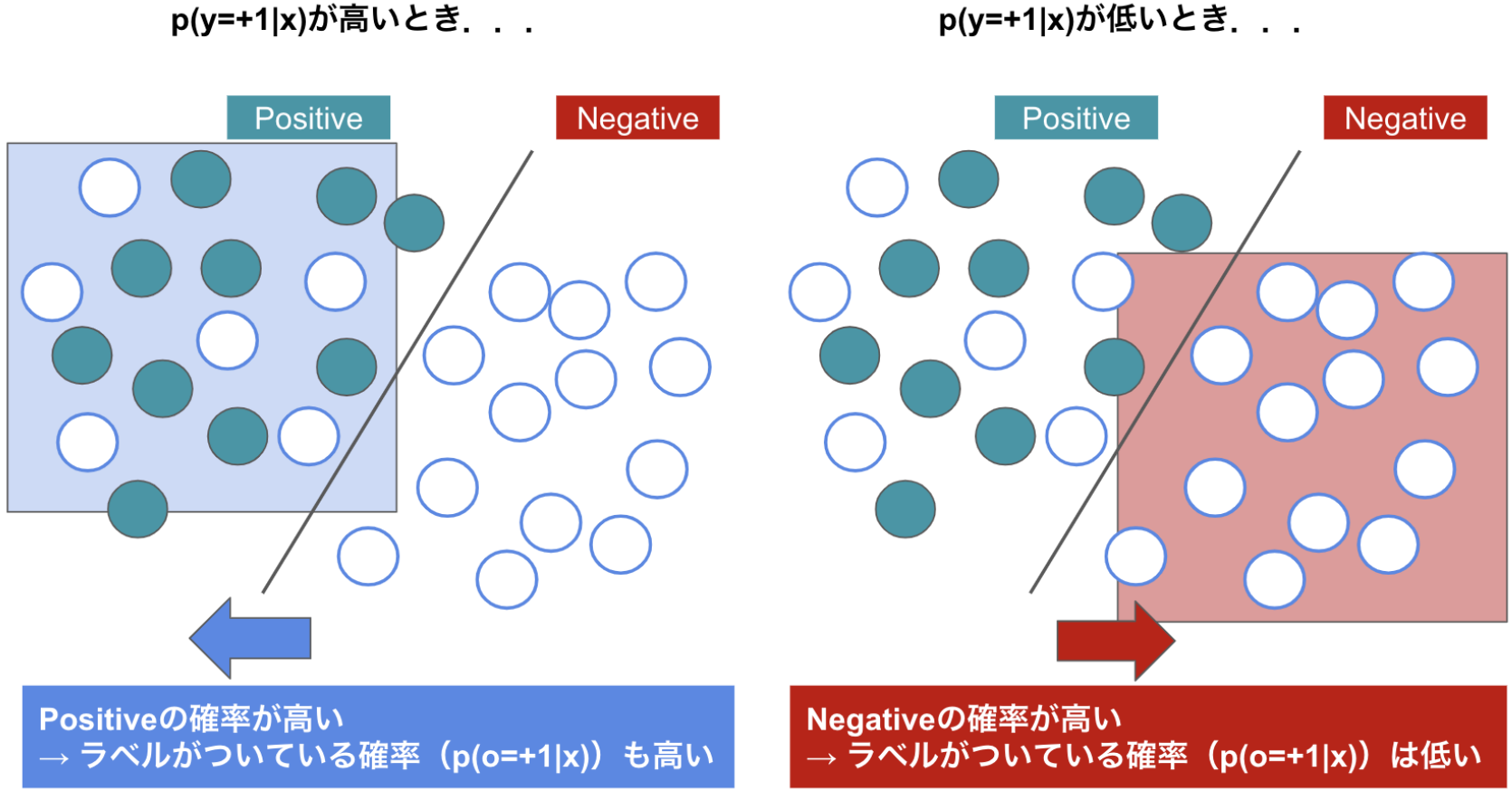
- 3番目の等号にSCARを仮定
  - $p(x|y=+1,o=+1)$は実際のサンプルから推定できる
  - πは過去の事前知識を活用できる

Invariance of Order Assumption

$x_i, x_j \in \chi$について,

$$p(y = +1|x_i) \leq p(y = +1|x_j) \Leftrightarrow p(o = +1|x_i) \leq p(o = +1|x_j)$$

- ラベルはpositive data($y = +1$)のみに付与されることから



Strategy for Partial Identification and Classification

(Theorem) density ratio $r(x) = \frac{p(x|y=+1,o=+1)}{p(x)}$ について,

$$p(y=+1|x_i) \leq p(y=+1|x_j) \Leftrightarrow r(x_i) \leq r(x_j)$$

が成り立つ.

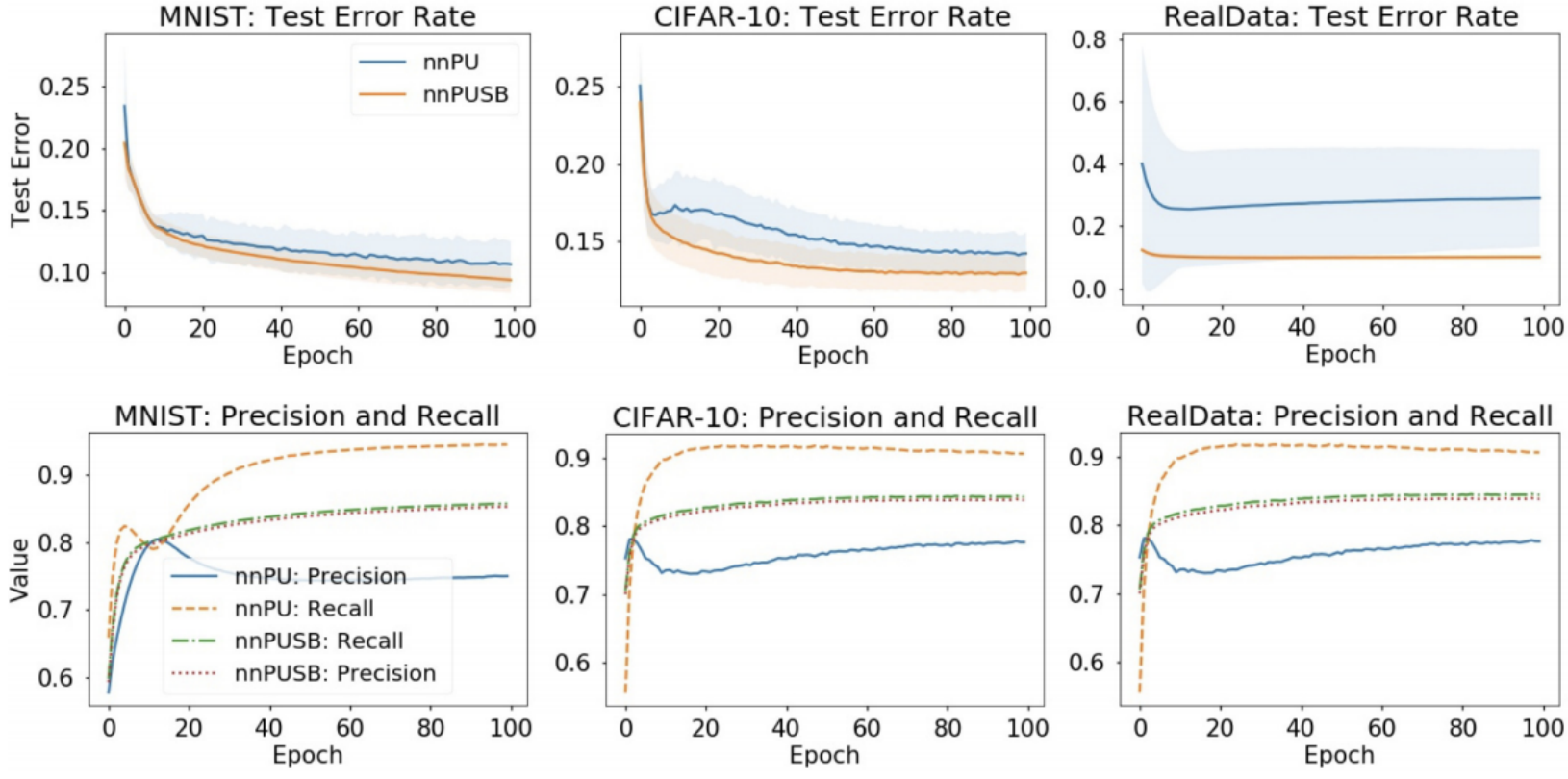
- この$r(x)$を利用して，Binary Classifierがつくれる
  - ある閾値θ_πを置いたとき，$r(x) > \theta_\pi$であればpositive
- 閾値θ_πは，事前知識πを利用して計算
  - $\pi = \int 1[r(x) \geq \theta_\pi] p(x) dx$

Algorithm Overview

1. Input: $p(x|y=+1)$, $p(x)$, class-prior π
2. $p(x|y=+1)$と$p(x)$を使ってdensity ratio $r(x)$を計算
3. $r(x)$を使って閾値θ_πを計算
4. 得られる分類器は$h(x) = sign(r(x) - \theta_\pi)$

Experimental Results

- MNIST, CIFAR-10, RealDataで実験



On the Minimal Supervision for Training Any Binary Classifier from Only Unlabeled Data

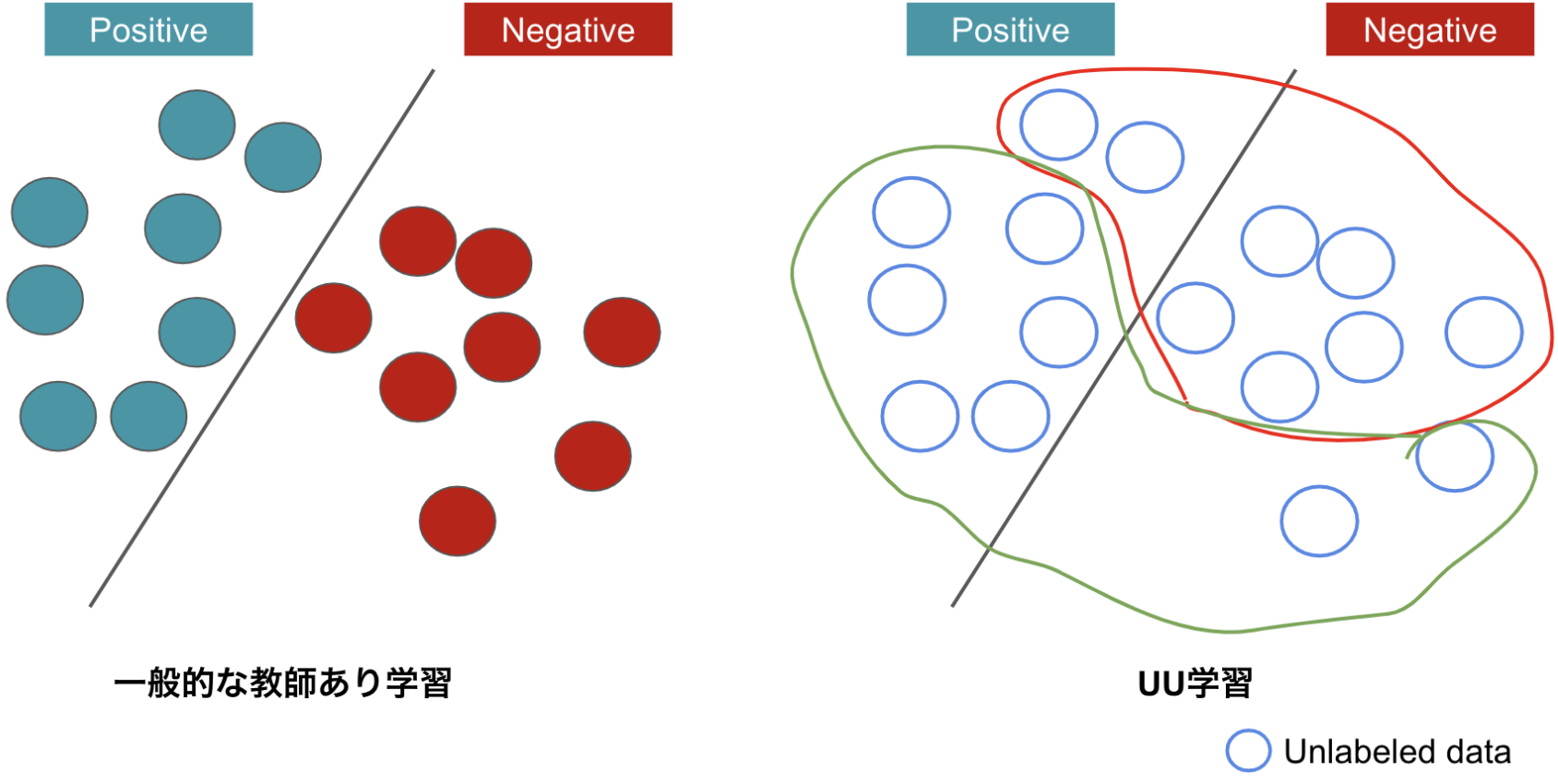
- Nan Lu, Gang Niu, Aditya Krishna Menon, Masashi Sugiyama

Abstract

- 出どころの違う２つのラベル無しデータ集合のみから分類器を学習

UU Learning

- データの分布が違う２つのラベルなしデータセットから学習
- クラスラベルの代わりに"データの出どころ"にラベリングするイメージ
  - クラスタリングではなく弱教師あり学習の区分



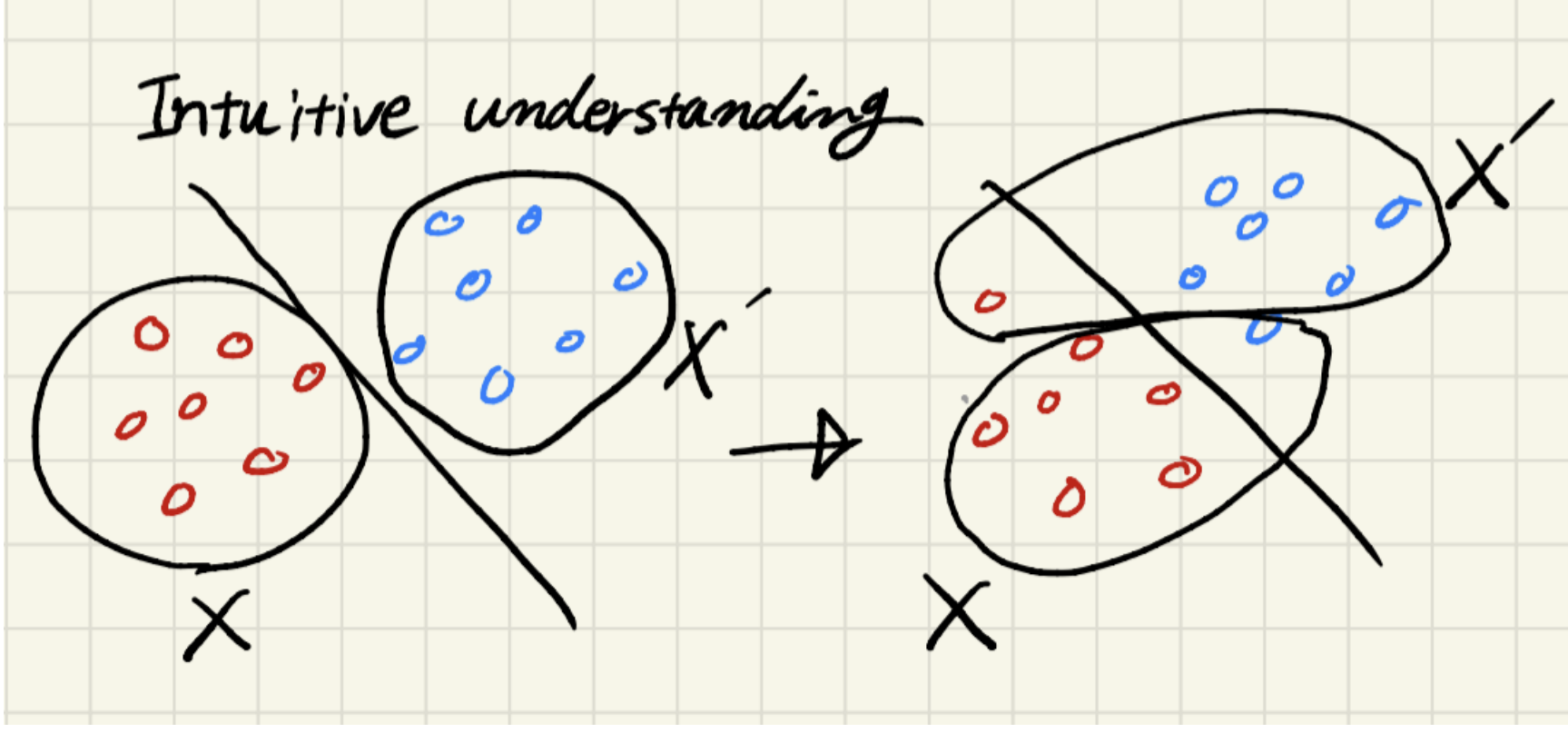
Contribution

- 任意のBinary Classifierが単一のラベル無しデータ集合のみから学習を行うことが不可能であることを証明
- 分布の違う2つのラベル無しデータ集合からであれば学習を行うことが可能になることを証明

Intuitive Understanding

直感的理解

- XとX'の分布が違うことがわかっている
- →分布の違い≒データの含まれる割合の違い
- →Xにとっての多数派（赤）とX'にとっての少数派（青）が同じクラスになるはず



A Brief Review of Empirical Risk Minimization

- ERMは以下リスク$R(g)$を最小化するようにモデルを選択する.

$$\hat{R}(g) = \frac{\pi_p}{n}\sum_{i=1}^{n} l(g(x_i^+)) + \frac{1-\pi_p}{n'}\sum_{j=1}^{n'} l(-g(x_j^-))$$

- ここでx_i^+はpositiveデータ，x_j^-はnegtiveデータ

本論文の目的：ERMのxからpositive/negativeの区別を取り去る

Risk Rewrite for UU-Learning

異なるデータの分布p_{tr}, p'_{tr}に対して，$R(g)$を以下のように書き換える

$$R(g) = \mathbb{E}_{p_{tr}}[\bar{l}_+(g(X))] + \mathbb{E}_{p'_{tr}}[\bar{l}_-(-g(X))]$$

ここで,

- $\bar{l}_+(z) = al(z) + bl(-z)$
- $\bar{l}_-(z) = cl(z) + dl(-z)$

$\bar{l}(\cdot)$は何を意味しているか？

- ノイジーなデータセットに対して，損失関数$l(\cdot)$を補正している
  - label correctionという分野
  - 損失関数に対して適切な係数をかけることでノイジーなデータで学習できることを示している

Label Correction

2値分類問題において，入力zのラベルがノイジーであるとする，e.g.,

- ラベル0の入力zは1/4の確率で本当はクラス1
- ラベル1の入力zは1/5の確率で本当はクラス0

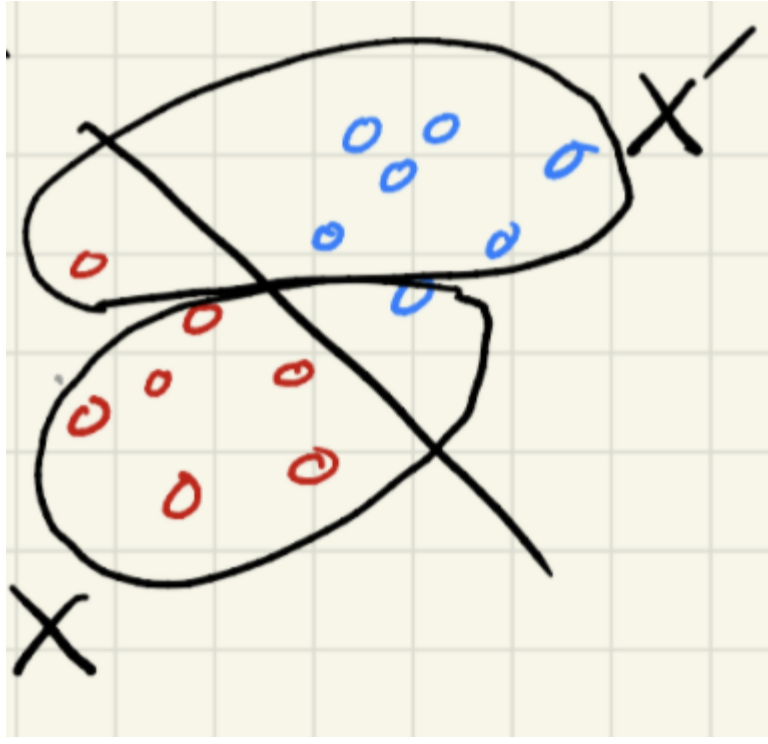
ノイズ発生確率に応じて損失関数に係数をかけると,

$$\bar{l}(z) = 0.75 \times l(z) + 0.8 \times l(-z)$$

Why Label Correction?

ラベル無しデータ集合をノイジーなデータ集合**に置き換えている.**

- データセットXはノイジーなPositiveデータセット
- データセットX'はノイジーなNegativeデータセット



本論文では，$\bar{l}'_+$および$\bar{l}_-$の係数を，

$$a = \frac{(1-\theta)\pi_p}{\theta-\theta'}, b = -\frac{\theta(1-\pi_p)}{\theta-\theta'}, c = \frac{\theta(1-\pi_p)}{\theta-\theta'}, \ d = -\frac{(1-\theta)\pi_p}{\theta-\theta'}$$

とする．これを代入して式変形していくと，

$$\hat{R}_{uu} = \frac{1}{n}\sum_{i=1}^{n} \alpha l(g(x_i)) + \frac{1}{n}\sum_{j=1}^{n'} \alpha' l(-g(x'_j)) - T$$

- $\alpha = (\theta' + \pi_p - 2\theta'\pi_p)/(\theta - \theta')$
- $T = \frac{\theta'(1-\pi_p)+(1-\theta)\pi_p}{\theta-\theta'}$

θおよびθ'はXおよびX'に正例が含まれる確率，π_pはデータ全体に正例が含まれる確率．← 事前知識を利用

UU-Learning with two Unlabeld Data Sets

目的（再活）：ERMのxからpositive/negativeの区別を取り去る

得られた$\hat{R}_{uu}(g)$は,

$$\hat{R}_{uu} = \frac{1}{n}\sum_{i=1}^{n} \alpha l(g(x_i)) + \frac{1}{n}\sum_{j=1}^{n'} \alpha' l(-g(x'_j)) - T$$

ERMの式から明示的なラベルを削除する代わりに，データの分布に関する事前知識を導入することでUU-Learningを達成

Multi-class classification without multi-class labels

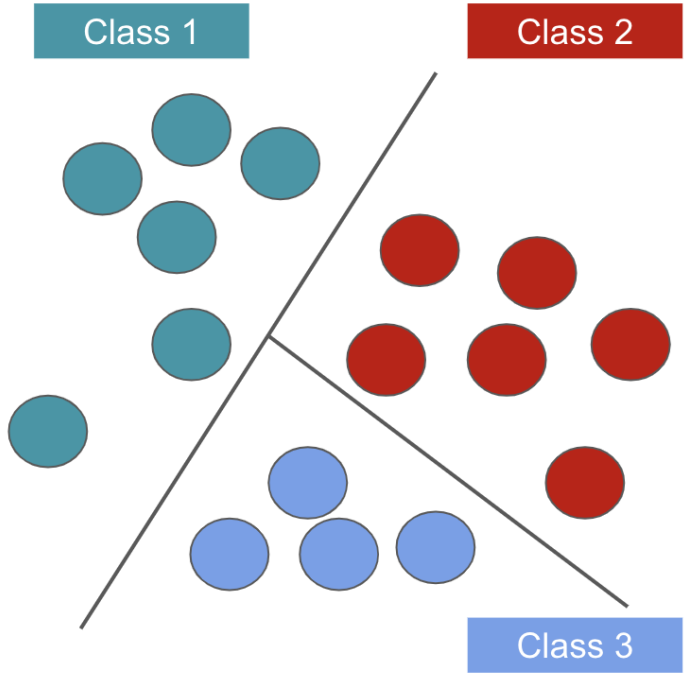
- Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, Zsolt Kira

Abstract

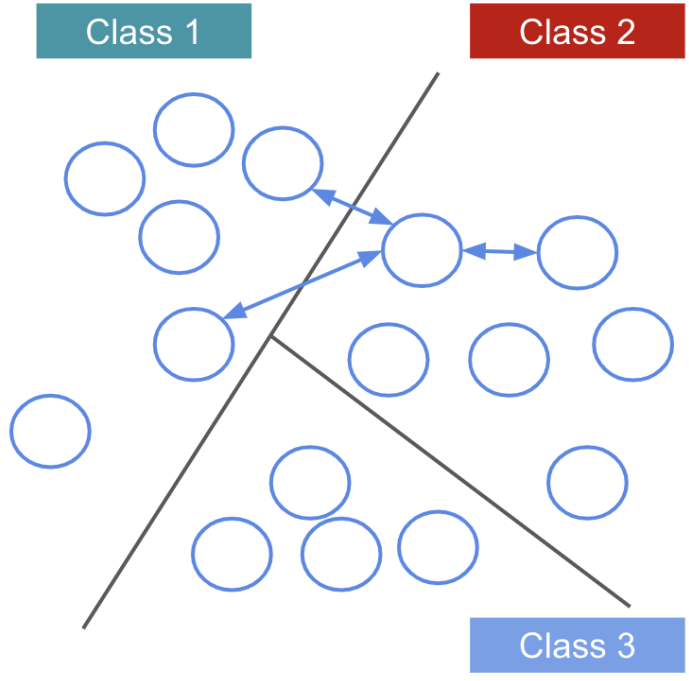
- 明示的なクラスラベルの付与なしに分類器を学習
    - クラスラベルの代わりにサンプル同士の類似度を活用

Pairwise Similarity Learning

- クラスラベルではなくペアが似てるかどうかをラベリング
  - クラス数が膨大な時に効率的にアノテーション可能
  - タスクによって再アノテーションが必要ない
  - クラス数可変のタスクに適用できる



一般的な教師あり学習

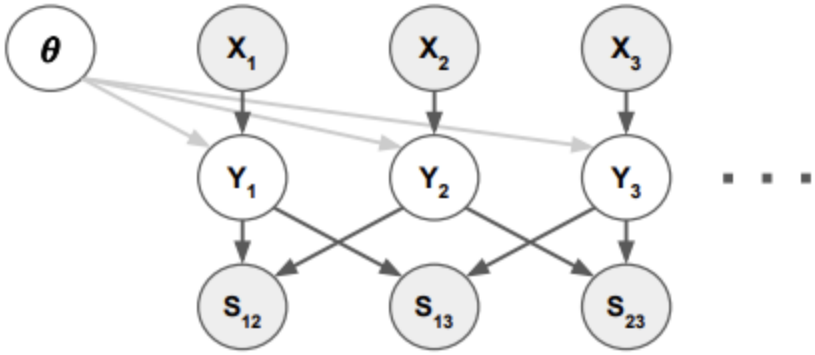


Pairwise Similarity 学習

Pairwise Similarity

Notation of Meta Classification Learning

- 解きたいタスクは以下のグラフィカルモデルで表現できる



- 観測
  - サンプル集合$X = \{X_1, ..., X_n\}$
  - 類似度集合$S = \{S_{ij}\}_{1 \leq i,j \leq n}$
- 隠れ変数
  - クラスラベル集合$Y = \{Y_1, ..., Y_n\}$
  - モデルのパラメータθ

Meta Classification Learning

尤度は,

$$L(\theta; X, Y, S) = P(X, Y, S; \theta) = P(S|Y)P(Y|X; \theta)P(X)$$

本論文の目的：損失関数内からラベル集合Yを取り去る

A Loss Function

最終的に得たい目的関数は,

$$L_{meta} = -\sum_{i,j} s_{ij} \log \hat{s}_{ij} + (1 - s_{ij}) \log(1 - \hat{s}_{ij})$$

- ここで$\hat{s}_{ij}$はx_iとx_jの間の類似度の予測値
- 類似度を近づけるように学習　→　ラベル不要

Pairwise Similarity

- 類似度にはベクトルの内積を使える．例えば，

$$\vec{v_1} = (0.0, 0.2, 0.8, 0.0)$$
$$\vec{v_2} = (0.2, 0.2, 0.3, 0.3)$$
$$\vec{v_3} = (0.1, 0.3, 0.6, 0.0)$$

のとき，

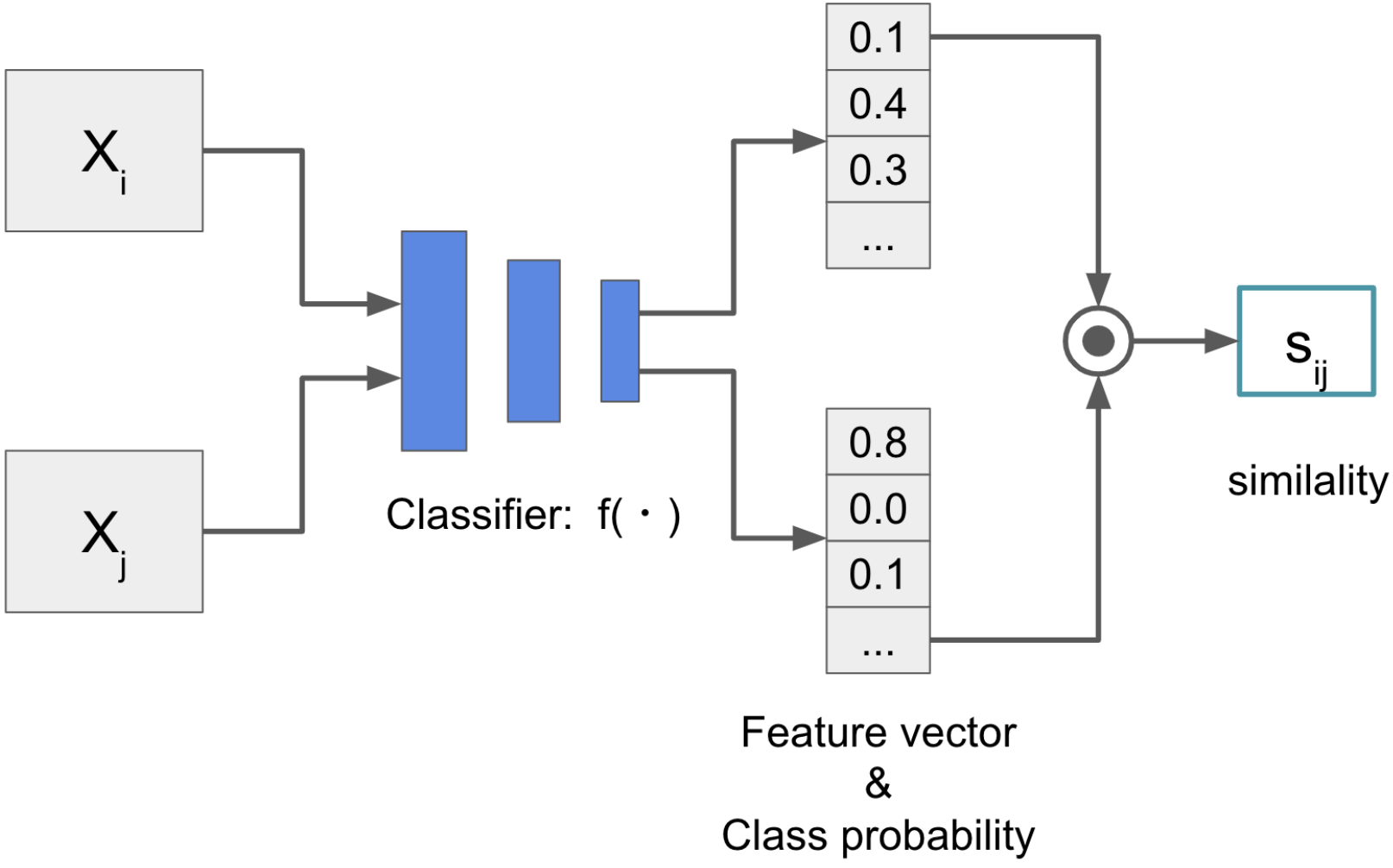
$$s_{12} = 0.28$$
$$s_{13} = 0.56$$
$$s_{23} = 0.26$$

となり，$\vec{v_1}$と$\vec{v_3}$が似ているとみなせる．

Pairwise Similarity for Multi-Label Classifier

- $\hat{s}_{ij}$はx_iとx_jの類似度.
  - x_iとx_jに対応するベクトルを決めたい.
  - 多クラス分類器$f(x_*)$を用意すると都合が良さそう

- 多クラス分類器$f(x_*)$を用意すると都合が良さそう



- （再活）最終的な目的関数

$$L_{meta} = -\sum_{i,j} s_{ij} \log \hat{s}_{ij} + (1 - s_{ij}) \log(1 - \hat{s}_{ij})$$

$$\hat{s}_{ij} = f(x_i; \theta)^T f(x_j; \theta)$$

明示的なラベルYを一切使わずに，目的関数に多クラス分類器$f(\cdot; \theta)$を導入できた

Conclusion

- ラベル不完全な問題設定のICLR2019採択論文を紹介
- 本来解けない問題設定でも仮定を導入することで可解になる
- 見えているタスクだけではなく見えない仮定を意識することが重要
  - 暗黙のうちに好ましくない仮定を置いていないか？
  - ある仮定を置くことで可解なタスクに落ちないか？

References

- [1] Kato, Masahiro, Teshima, Takeshi, and Honda, Junya.Learning from positive and unlabeled data witha selection bias. InInternational Conference onLearning Representations, 2019.
- [2] Lu, Nan, Niu, Gang, Menon, Aditya K., andSugiyama, Masashi. On the minimal supervisionfor training any binary classifier from only unla-beled data. InInternational Conference on LearningRepresentations, 2019.
- [3] Hsu, Yen-Chang, Lv, Zhaoyang, Schlosser, Joel,Odom, Phillip, and Kira, Zsolt. Multi-class clas-sification without multi-class labels.InInter-national Conference on Learning Representations,2019.
- [4] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In ICDM, pp. 213–220, 2008.