Multi-label Ranking from Positive and Unlabeled Data

Atsushi Kanehira and Tatsuya Harada

Background

- Label incompleteness in the multi-label dataset

creating dataset

Label **all objects** in this image

man, wo... pe... e, smart phone, ch... e, window, paper... etc...

annotator

Too much cost !

Background

- Label incompleteness in the multi-label dataset

creating dataset

Label **at least one object** in this image

man, chair, table

annotator

Created dataset

assigned labels

man, chair
table

absent but positive labels

pet-bottle, smart
Phone, window,
etc...

Goal

- Training a classifier from data with incomplete labels

Problem setting

1. assigned labels are definitely positive,
2. absent labels are **Not** necessarily negative,
3. sar es are allowed to take more than one labels

PU (Positive and Unlabeled) classification + **Multi-label classification** = **Multi-label PU classification**

assigned labels	absent but positive labels
man, chair table	pet-bottle, smart Phone, window, etc…

Formulation

- Formulation of multi-label ranking

$$\min L_{\text{true}} = \mathbb{E}_{\mathrm{xy}}[R(f(\mathrm{x}), \mathrm{y})]$$

$$R(\mathrm{f}(\mathrm{x}), \mathrm{y}) = p(f_i < f_j | y_i = 1, y_j = 0)$$ (mis-rank rate)

$\mathrm{x} \in \mathbb{R}^d$: sample, $\mathrm{y} \in \{0, 1\}^m$: true label

where *d* Is feature dimension and *m* is the number of classes

mis-rank rate

$f_{\text{dog}}($  $) < f_{\text{cat}}($  $)$

Formulation

- Formulation of multi-label ranking

$$\boxed{\min L_{\text{true}} = \mathbb{E}_{\mathrm{xy}}[R(f(\mathrm{x}), \mathrm{y})]}$$

$$R(\mathrm{f}(\mathrm{x}), \mathrm{y}) = p(f_i < f_j | y_i = 1, y_j = 0)$$ (mis-rank rate)

$\mathrm{x} \in \mathbb{R}^d$: sample, $\mathrm{y} \in \{0,1\}^m$: true label

where *d* Is feature dimension and *m* is the number of classes

mis-rank rate

$f_{\text{dog}}($  $) < f_{\text{cat}}($  $)$

Formulation

- Formulation of multi-label ranking

$$\min\ L_{\text{true}} = \mathbb{E}_{\mathbf{xy}}[R(f(\mathbf{x}),\mathbf{y})]$$

$$\boxed{R(\mathbf{f}(\mathbf{x}),\mathbf{y}) = p(f_i < f_j | y_i = 1, y_j = 0)\ \text{(mis-rank rate)}}$$

$\mathbf{x} \in \mathbb{R}^d$: sample, $\mathbf{y} \in \{0,1\}^m$: true label

where *d* Is feature dimension and *m* is the number of classes

mis-rank rate

$f_{\text{dog}}($  $) < f_{\text{cat}}($  $)$

Formulation

- Formulation of multi-label **PU** ranking

$$\min \ L_{\text{true}} = \mathbb{E}_{\mathbf{xy}}[R(f(\mathbf{x}), \mathbf{y})]$$

$$R(\mathbf{f}(\mathbf{x}), \mathbf{y}) = p(f_i < f_j | y_i = 1, y_j = 0)$$ (mis-rank rate)

$\mathbf{x} \in \mathbb{R}^d$: sample, ~~$\mathbf{y} \in \{0,1\}^m$: true label~~ unknown

$\boxed{s \in \{0,1\}^m : \text{observed label}}$ known

where *d* is feature dimension and *m* is the number of classes

minimizing ranking loss, with only *observed s*

Relationship between two variables

$s \in \{0,1\}^m$: observed label

$y \in \{0,1\}^m$: true label

s_i=1 (assigned) | s_i=0 (assigned)

y_i=1 (positive) | y_i=0 (negative)

assigned labels

man, chair
table

absent but positive labels

pet-bottle, smart
Phone, window,
etc...

Analysis of multi-label PU ranking

$$\min L_{\mathrm{true}} = \mathbb{E}_{\mathbf{xy}}[R(f(\mathbf{x}), \mathbf{y})]$$

$$R(\mathbf{f}(\mathbf{x}), \mathbf{y}) = p(f_i < f_j | y_i = 1, y_j = 0)$$ (mis-rank rate)

We can't estimate mis-rank rate because it depends on true (unknown) label

- Instead, we can observe

$$R_X(\mathbf{f}(\mathbf{x}), \mathbf{s}) = p(f_i < f_j | s_i = 1, s_j = 0)$$ (pseudo mis-rank rate)

- set loss function as

$$L_{\mathrm{PU}} = \mathbb{E}_{\mathbf{xs}}[c_{ij} R_X(f(\mathbf{x}), \mathbf{s})] = L_{\mathrm{true}} - \mathrm{const}$$

(where $c_{ij} = \dfrac{p(y_i = 1)}{p(s_i = 1, s_j = 0)}$)

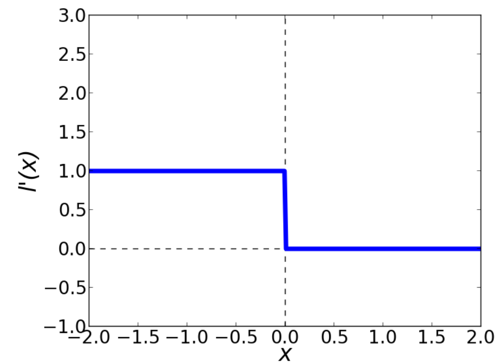
Conclusion ①

Loss function should be weighted properly

Surrogate loss

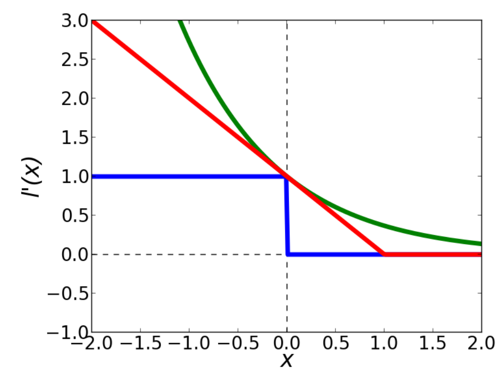
- optimization of $L_{\mathrm{PU}} = \mathbb{E}_{\mathrm{xs}}[c_{ij} R_X(f(\mathrm{x}), \mathrm{s})]$

$$R_X(f(\mathrm{x}), \mathrm{y}) = p(f_i < f_j | s_i = 1, s_j = 0)$$
$$= \mathbb{E}_{\mathrm{x}|s_i=1,s_j=0}[l_{\text{0-1}}(f_i - f_j)]$$



Due to computationally complexity, surrogate loss (e.g. hinge) is usually used.

$$= \mathbb{E}_{\mathrm{x}|s_i=1,s_j=0}[l'_{\mathrm{sur}}(f_i - f_j)]$$



Using surrogate loss,

$$L_{\mathrm{PU}} \approx L'_{\mathrm{PU}}$$

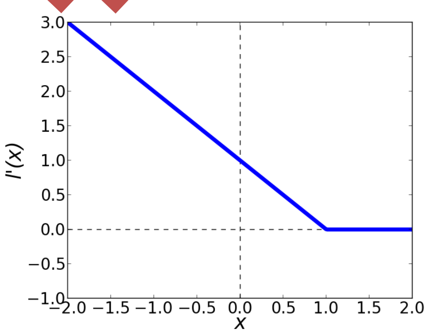
Analysis of multi-label PU ranking

- Bias with surrogate loss

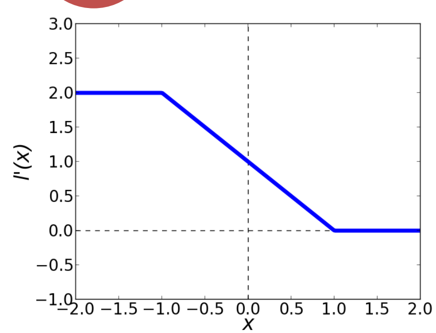
$$L'_{\mathrm{PU}} = L'_{\mathrm{true}}$$
$$\cancel{+\, p(y_i = 1, y_j = 1)\mathbb{E}_{x|y_i=1,y_j=1}[l'(f_i - f_j) + l'(f_j - f_i)]}$$

If we select surrogate loss to meet $l'(f_i - f_j) + l'(f_j - f_i) = \mathrm{const}$
bias term can be cancelled

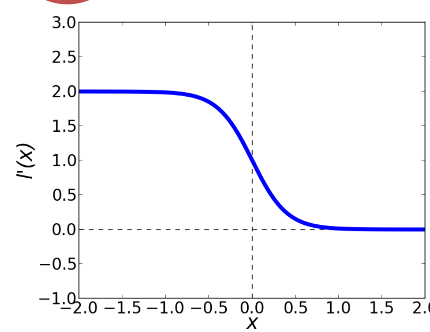
hinge loss



ramp loss



sigmoid loss



Conclusion ②

Symmetric surrogate loss should be used

Proposed Conditions

- We showed
  1. loss function should be weighted properly, and
  2. symmetric surrogate loss should be used.

(similar to "Analysis of learning from Positive and Unlabeled Data" [du Plessis, NIPS 2014])

Experiment (setting)

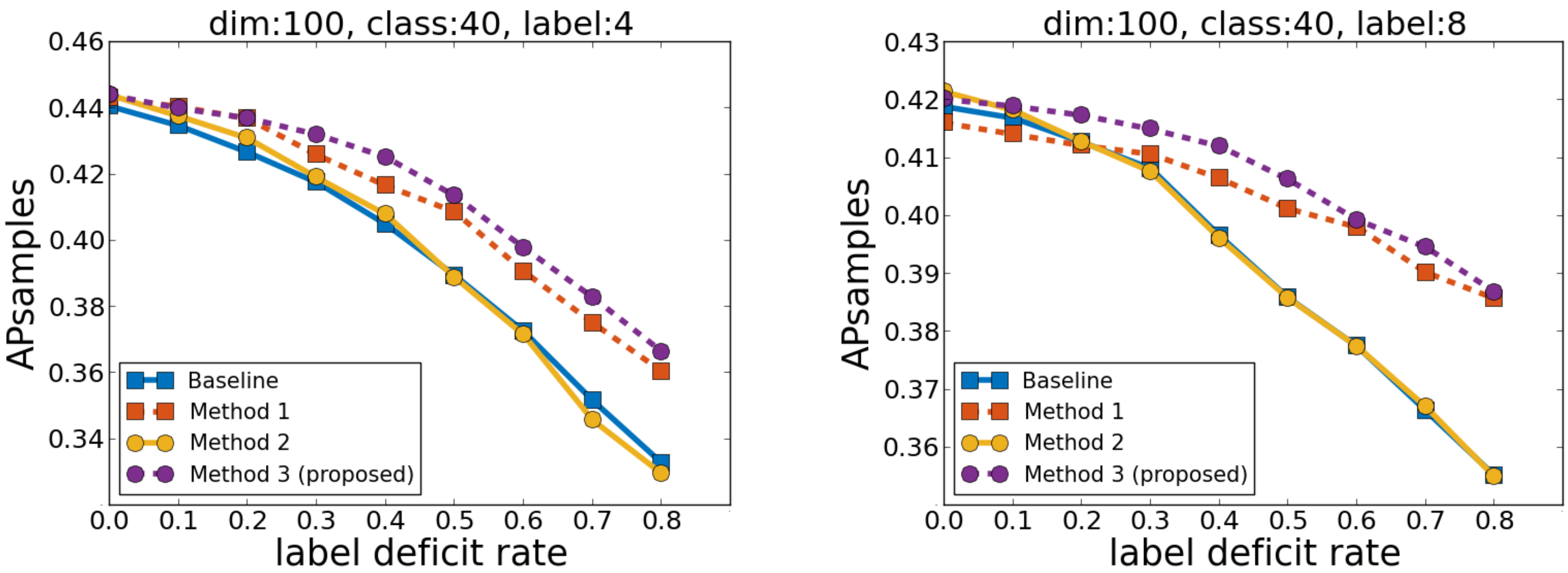
- Experiment with synthetic dataset
  - train ： 8000 samples, test: 2000 samples, class :40
  - 0%～80% label noise in positive labels
  - Optimize linear classifier with SGD

Condition 1. weigh loss function properly
Condition 2. use symmetric surrogate loss

		Surrogate loss	
		Not symmetric (hinge loss)	Symmetric (ramp loss)
weight	no	Baseline	Method ②
	yes	Method ①	Method ③(proposed)

Experimental (results)

- Method with proposed conditions outperform others



Conclusion

- ✓ We showed two conditions, which should be met in multi-label PU ranking
- ✓ We demonstrated the effectiveness of these condition by experimental results