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, woman, pet, table, smart phone, chair, bottle, window, paper, book, can etc...

Too much cost!
Background

- Label incompleteness in the multi-label dataset

Creating dataset

Label at least one object in this image

man, chair, table

Created dataset

assigned labels: man, chair, table

absent but positive labels: pet-bottle, smart phone, window, etc...

annotator
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. samples are allowed to take more than one labels

PU (Positive and Unlabeled) classification

Multi-label classification

Multi-label PU classification

assigned labels

man, chair, table

absent but positive labels

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

Formulation of multi-label ranking

\[
\min L_{\text{true}} = \mathbb{E}_{x,y}[R(f(x), y)]
\]
\[
R(f(x), y) = p(f_i < f_j | y_i = 1, y_j = 0)
\]

where \(d\) is feature dimension and \(m\) is the number of classes

mis-rank rate

\[f_{\text{dog}}(\quad)<f_{\text{cat}}(\quad)\]
Formulation of multi-label ranking

\[
\min L_{\text{true}} = \mathbb{E}_{xy}[R(f(x), y)] \\
R(f(x), y) = p(f_i < f_j | y_i = 1, y_j = 0) \quad \text{(mis-rank rate)}
\]

where \( d \) is feature dimension and \( m \) is the number of classes

mis-rank rate

\[
f_{\text{dog}}(\quad) < f_{\text{cat}}(\quad)
\]
Formulation

Formulation of multi-label ranking

\[
\min L_{\text{true}} = \mathbb{E}_{x,y}[R(f(x), y)] \\
R(f(x), y) = p(f_i < f_j | y_i = 1, y_j = 0) \quad \text{(mis-rank rate)}
\]

\(x \in \mathbb{R}^d: \text{sample, } y \in \{0,1\}^m: \text{true label}\)

where \(d\) is feature dimension and \(m\) is the number of classes

mis-rank rate

\[
f_{\text{dog}}(\quad) < f_{\text{cat}}(\quad)
\]
Formulation of multi-label PU ranking

\[ \min L_{\text{true}} = \mathbb{E}_{xy}[R(f(x), y)] \]
\[ R(f(x), y) = p(f_i < f_j | y_i = 1, y_j = 0) \] (mis-rank rate)
\[ x \in \mathbb{R}^d : \text{sample, } y \in \{0, 1\}^m : \text{true label unknown} \]
\[ s \in \{0, 1\}^m : \text{observed label known} \]

where \( d \) is feature dimension and \( m \) is the number of classes

minimizing ranking loss, with only observed \( s \)
Relationship between two variables

\[ s \in \{0, 1\}^m : \text{observed label} \]

\[ y \in \{0, 1\}^m : \text{true label} \]

assigned labels
- man, chair, table

absent but positive labels
- pet-bottle, smart Phone, window, etc...
Analysis of multi-label PU ranking

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

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

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

➢ Instead, we can observe

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

➢ set loss function as

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

\[
= L_{\text{true}} - \text{const}
\]

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

Conclusion ①

Loss function should be weighted properly
Surrogate loss

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

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

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

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

Using surrogate loss,

\( L_{PU} \approx L'_{PU} \)
Analysis of multi-label PU ranking

Bias with surrogate loss

\[ L'_\text{PU} = L'_\text{true} + p(y_i = 1, y_j = 1)\Pr_{x|y_i=1,y_j=1}[l'(f_i - f_j) + l'(f_j - f_i)] \]

If we select surrogate loss to meet \( l'(f_i - f_j) + l'(f_j - f_i) = \text{const} \), 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

<table>
<thead>
<tr>
<th>Surrogate loss</th>
<th>weight</th>
<th>Baseline</th>
<th>Method ②</th>
<th>Method ①</th>
<th>Method ③(proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not symmetric</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>Method ①</td>
<td></td>
<td>Method ③</td>
<td></td>
</tr>
</tbody>
</table>
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