On the Consistency of Multi-Label Learning

Wei Gao and Zhi-Hua Zhou
{gaow,zhouzh}@lamda.nju.edu.cn

LAMDA Group
National Key Laboratory for Novel Software Technology
Nanjing University, China

COLT, July 2011
Outline

✓ Background

✓ Our Work

✓ Conclusion and Future Work
Multi-Label Learning

An image has multiple labels

Ocean
Beach
Plants
Sky
Cloud

…”
Multi-label learning: predict a set of labels to an instance

Traditional learning: each instance has a single label
Multi-label Learning Setup

- Predefined label set $\mathcal{L} = \{1, 2, \ldots, Q\}$
- Input space $\mathcal{X}$, output space $\mathcal{Y} = 2^\mathcal{L} = \{-1, +1\}^Q$
- Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

Given: Sample of instances $S = \{(X_i, Y_i)_{i=1}^m\} \sim \mathcal{D}^m$

Goal: Learn a prediction $h: \mathcal{X} \rightarrow \mathcal{Y}$ to assign a set of labels

<table>
<thead>
<tr>
<th>Loss functions or criterions</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking loss</td>
<td>precision</td>
</tr>
<tr>
<td>hamming loss</td>
<td>accuracy</td>
</tr>
<tr>
<td>one-error</td>
<td>average precision</td>
</tr>
<tr>
<td>coverage</td>
<td>...</td>
</tr>
</tbody>
</table>
## Surrogate Loss

<table>
<thead>
<tr>
<th>Loss functions or criterions</th>
<th>recall</th>
<th>precision</th>
<th>accuracy</th>
<th>average precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ranking loss</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>hamming loss</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one-error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Non-convex and discontinuous

- **Boosting algorithm** (Schapire and Singer, 2000)
- **Neural network algorithm BP-MIL** (Zhang and Zhou, 2006)
- **SVM-style algorithms** (Elisseeff and Weston, 2002, Taskar et al., 2004, Hariharan et al., 2010)

All these methods try to optimize some surrogate losses, such as exponential loss and hinge loss.
Binary Classification

This surrogate method has been used for other learning problems

Binary classification

Given: Instance $X$, output $Y \in \{+1, -1\}$
Learn: Classification function $f$

true loss $I[Yf(X) \leq 0]$

surrogate loss $\Psi(Yf(X))$

hinge loss
exponential loss
logistic loss

Hard

Tractable
**Definition 1:** The surrogate loss $\Psi$ is said to be **consistent** with loss $L$ if it holds that, for every function $f$,

$$f \in \arg\min E[\Psi(f(X), Y)] \quad \text{then} \quad f \in \arg\min E[L(f(X), Y)]$$
Previous Consistent Work

- **Two-class learning** (Zhang, 2004b; Steinwart, 2005; Bartlett et al. 2006)
  - Most methods are consistent, e.g. Boosting, SVM, Logistic Boost, etc.

- **Multi-class learning** (Zhang, 2004a; Tewari and Bartlett 2007)
  - SVM-style algorithms are inconsistent
  - Remains open for other algorithms

- **Multi-label learning**
  - **Remains totally open -- Our Work**

- **Learning to rank** (Cossock and Zhang, 2008, Xia et al., 2008, Duchi et al., 2010)
Outline

✓ Background

✓ Our Work

✓ Conclusion and Future Work
Conditional Risk

Conditional risk and conditional surrogate risk:

\[ l(p, f) = \sum_{Y \in \mathcal{Y}} p_Y L(f(X), Y) = \sum_{Y \in \mathcal{Y}} \Pr(Y|X) L(f(X), Y) \]

\[ W(p, f) = \sum_{Y \in \mathcal{Y}} p_Y \Psi(f(X), Y) = \sum_{Y \in \mathcal{Y}} \Pr(Y|X) \Psi(f(X), Y) \]

Set of Bayes predictors: \[ A(p) = \arg \min_{f'} l(p, f') \]

Example

For an instance \( X \), outputs \( Y_1, Y_2 \), \( \Pr(Y_1|X) = 0.5 \) and \( \Pr(Y_2|X) = 0.5 \), then

\[ l(p, f) = 0.5L(f(X), Y_1) + 0.5L(f(X), Y_2) \]

\[ W(p, f) = 0.5\Psi(f(X), Y_1) + 0.5\Psi(f(X), Y_2) \]
Consistency Theorem

**Theorem 1:** Surrogate loss $\Psi$ is consistent with $L$ if and only if

$$\inf_f \left\{ \sum_{Y \in \mathcal{Y}} \Pr(Y|X) \Psi(f(X), Y) \right\} < \inf_f \left\{ \sum_{Y \in \mathcal{Y}} \Pr(Y|X) \Psi(f(X), Y) \mid f \notin A(p) \right\}$$

which is equivalent to $\inf_f \{W(p, f)\} < \inf_f \{W(p, f) \mid f \notin A(p)\}$.

**Intuition**

$$\arg\min_f \sum_{Y \in \mathcal{Y}} p_Y \Psi(f(X), Y)$$

$$\arg\min_f \sum_{Y \in \mathcal{Y}} p_Y L(f(X), Y)$$
Ranking Loss

For a vector \( f = (f_1, f_2, \ldots, f_Q) \), the ranking loss is

\[
L_{\text{rankloss}}(f, (X, Y)) = \sum_{y_i=-1, y_j=+1} a_Y I[f_i(X) \geq f_j(X)] \\
= \sum_{y_i < y_j} a_Y I[f_i(X) \geq f_j(X)]
\]

Example

\( L = \{1, 2, 3, 4\} \quad f_i := f_i(X) \)

\( (X, Y) \quad Y = \{1, 3\} \quad Y = (+1, -1, +1, -1) \)

\[
L_{\text{rankloss}}(f, (X, Y)) = a_Y (I[f_1 \leq f_2] + I[f_1 \leq f_4] \\
+ I[f_3 \leq f_2] + I[f_3 \leq f_4])
\]
Surrogate Loss

- Ranking loss
  \[ L_{\text{rankloss}} (f, (X, Y)) = \sum_{y_i < y_j} a_Y I[f_i(X) \geq f_j(X)] \]

- Surrogate loss
  \[ \Psi(f, (X, Y)) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X)) \]

Many multi-label algorithms fall into this formulation
(Schapire and Singer, 2000; Dekel et al., 2004; Zhang and Zhou, 2006)

\[ \phi(x) = e^{-x} \]

(Elisseeff and Weston, 2002)
\[ \phi(x) = (1 - x)_+ \]
Inconsistency for Ranking Loss

**Theorem 2:** For **every** convex function $\phi$, the surrogate loss

$$\Psi(f, (X, Y)) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X))$$

is **inconsistent** with ranking loss

$$L_{\text{rankloss}}(f, (X, Y)) = \sum_{y_i < y_j} a_Y I[f_i(X) \geq f_j(X)].$$

**Proof.**

$$W(p, f) = \sum_{Y \in \mathcal{Y}} p_Y \Psi(f(X), Y)$$

$$= 0.5\phi(f_2 - f_1) + 0.5\phi(f_1 - f_2) \implies f_1 = f_2$$

$$l(p, f) = \sum_{Y \in \mathcal{Y}} p_Y L(f(X), Y)$$

$$= 0.5I[f_1 \geq f_2] + 0.5I[f_2 \geq f_1] \implies f_1 \neq f_2$$

$p_{Y_1} = p_{Y_2} = 0.5$

$a_{Y_1} = a_{Y_2} = 1$
What is the Problem?

\[ L_{\text{rankloss}}(f, (X, Y)) = \sum_{y_i < y_j} a_Y(I[f_i(X) > f_j(X)] + I[f_i(X) = f_j(X)]) \]

\[ l(p, f) = \sum_{1 \leq i < j \leq Q} I[f_i > f_j] \Delta_{i,j} + I[f_i < f_j] \Delta_{j,i} + I[f_i = f_j](\Delta_{i,j} + \Delta_{j,i}) \]

Minimizing \( l \) gives

\[ A(p) = \{ f: f_i > f_j \text{ if } \Delta_{i,j} < \Delta_{j,i}; f_i < f_j \text{ if } \Delta_{i,j} > \Delta_{j,i}; f_i \neq f_j \text{ if } \Delta_{i,j} = \Delta_{j,i} \} \]

Conditional surrogate risk

\[ W(p, f) = \sum_{1 \leq i < j \leq Q} \phi(f_j - f_i) \Delta_{i,j} + \phi(f_i - f_j) \Delta_{j,i} \]

when \( \Delta_{i,j} = \Delta_{j,i} \) \( \min W(p, f) \Rightarrow f_i = f_j \)
Partial Ranking Loss

**Partial ranking loss** is given by

\[
L_{p\text{-}rankloss}(f, (X, Y)) = \sum_{y_i < y_j} a_Y \left( I[f_i(X) > f_j(X)] + 0.5 I[f_i(X) = f_j(X)] \right)
\]

\[
L_{\text{rankloss}}(f, (X, Y)) = \sum_{y_i < y_j} a_Y (I[f_i(X) > f_j(X)] + I[f_i(X) = f_j(X)])
\]

Set of Bayes predictors for partial ranking loss

\[
A(p) = \{ f : f_i > f_j \text{ if } \Delta_{i,j} < \Delta_{j,i} ; f_i < f_j \text{ if } \Delta_{i,j} > \Delta_{j,i} \}
\]

Conditional surrogate risk

\[
W(p, f) = \sum_{1 < i, j < Q} \phi(f_j - f_i) \Delta_{i,j} + \phi(f_i - f_j) \Delta_{j,i}
\]

when \( \Delta_{i,j} = \Delta_{j,i} \), \( \min W(p, f) \Rightarrow f_i = f_j \)
Consistency for Partial Ranking Loss

**Theorem 3:** If $\phi$ is differential, non-increasing, and $\phi(x) + \phi(-x) = 2\phi(0)$, then the surrogate loss

$$\Psi(f, (X, Y)) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X))$$

is **consistent** with partial ranking loss

$$L_{p\text{-rankloss}}(f, (X, Y)) = \sum_{y_i < y_j} a_Y \left( I[f_i(X) > f_j(X)] + \frac{1}{2} I[f_i(X) = f_j(X)] \right)$$

**Examples**

$$\phi(x) = -\arctan(x) \quad \phi(x) = \frac{1 - e^{2x}}{1 + e^{2x}}$$

$$\Psi(f(X), Y) = \sum_{y_i < y_j} -a_Y (f_j(X) - f_i(X)) + \tau \sum_{i=1}^{Q} f_i^2(X) \text{(convex)}$$
Inconsistency for Partial Ranking Loss

**Theorem 4:** If $\phi$ is **convex**, differential, non-increasing and **non-linear**, then the surrogate loss

$$\Psi(f, (X, Y)) = \sum_{y_i < y_j} a_Y \phi(f_j(X) - f_i(X))$$

is **inconsistent** with partial ranking loss

$$L_{p\text{-rankloss}}(f, (X, Y)) = \sum_{y_i < y_j} a_Y \left( I[f_i(X) > f_j(X)] + \frac{1}{2} I[f_i(X) = f_j(X)] \right)$$

**Examples**

**Exponential loss:** $\phi(x) = e^{-x}$ (Schapire and Singer, 2000; Zhang and Zhou, 2006)

**Logistic loss:** $\phi(x) = \ln(1 + e^{-x})$
Hamming Loss

For a vector $f$ and a prediction $F$, the hamming loss is

$$L_{\text{hamloss}}(F(f(X)), Y) = \frac{1}{Q} \sum_{i=1}^{Q} I[\hat{y}_i \neq y_i]$$

where $\hat{Y} = F(f(X)) = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_Q)$

Surrogate loss

$$\Psi(f(X), Y) = \max_{\hat{Y} \neq Y} \phi(\delta(\hat{Y}, Y) + f_{\hat{Y}}(X) - f_Y(X))$$

where $\phi(x) = \max(0, x)$ and $\delta(Y, \hat{Y}) = \sum_{i=1}^{Q} I[y_i \neq \hat{y}_i]$

(Taskar et al. 2004; Hariharan et al. 2010) work with this formula
Definition 2: In multi-label classification, if for every $X \in \mathcal{X}$, there exists a label $Y \in \mathcal{Y}$, such that $\Pr(Y|X) > 0.5$, then the task is deterministic, and non-deterministic otherwise.

Theorem 4: For multi-label deterministic classification, the surrogate loss

$$
\Psi(f(X), Y) = \max_{\hat{Y} \neq Y} \phi(\delta(\hat{Y}, Y) + f_{\hat{Y}}(X) - f_Y(X))
$$

$$
\phi(x) = \max(0, x) \quad \delta(Y, \hat{Y}) = \sum_{i=1}^{Q} I[y_i \neq \hat{y}_i]
$$

is inconsistent w.r.t. hamming loss.

(Taskar et al. 2004; Hariharan et al. 2010) are inconsistent.
Consistency for Deterministic Case

Definition 2: In multi-label classification, if for every \( X \in \mathcal{X} \), there exists a label \( Y \in \mathcal{Y} \), such that \( \Pr(Y|X) > 0.5 \), then the task is deterministic, and non-deterministic otherwise.

Theorem 4: For multi-label deterministic classification, the surrogate loss

\[
\Psi(f(X), Y) = \max_{\hat{Y} \neq Y} \phi(\delta(\hat{Y}, Y) + f_Y(X) - f_Y(X))
\]

\[
\phi(x) = \max(0, x) \quad \delta(Y, \hat{Y}) = \sum_{i=1}^{Q} I[y_i \neq \hat{y}_i]
\]

is inconsistent w.r.t. hamming loss; but consistent if \( \delta(Y, \hat{Y}) = I[Y \neq \hat{Y}] \)
Conclusion

We present a study on consistency of surrogate loss functions in multi-label learning.

Main Contributions:

- A general theorem for consistency
- Study two popular loss functions
  - Ranking loss:
    - all convex surrogate loss functions -> inconsistent
    - partial ranking loss -> with consistent surrogate loss
  - Hamming loss:
    - current surrogate loss -> inconsistent even for deterministic learning setting
    - a new surrogate loss -> consistent
Future Work

 ✓ Optimizing partial ranking loss
 ✓ Studying the consistency of other multi-label losses
 ✓ Considering label correlation

Thanks for your attention