Dual Set Multi-Label Learning

Chong Liu ¹, Peng Zhao ¹, Sheng-Jun Huang ², Yuan Jiang ¹, and Zhi-Hua Zhou ¹

¹ LAMDA Group, Nanjing University, China
² Nanjing University of Aeronautics and Astronautics, China
Outline

• Introduction
• Potential Solutions and Deficiencies
• Our Approach
• Theoretical Results
• Experiments
• Conclusion
Introduction

• An example of traditional multi-label learning

<table>
<thead>
<tr>
<th>Sky</th>
<th>Lake</th>
<th>Road</th>
<th>Mountain</th>
<th>Bird</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Introduction

• An example different from traditional multi-label learning

<table>
<thead>
<tr>
<th>Chinese Character</th>
<th>Calligraphers</th>
<th>Fonts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Xizhi Wang</td>
<td>Running Font</td>
</tr>
<tr>
<td></td>
<td>Shaoji He</td>
<td>Regular Font</td>
</tr>
<tr>
<td></td>
<td>Longji Li</td>
<td>Clerical Font</td>
</tr>
<tr>
<td></td>
<td>Mengfu Zhao</td>
<td>Semi-Running Font</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instance</th>
<th>Calligrapher Set</th>
<th>Font Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Character</th>
<th>Wang</th>
<th>He</th>
<th>Li</th>
<th>Zhao</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2nd</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3rd</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>4th</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Semi-running</th>
<th>Running</th>
<th>Clerical</th>
<th>Regular</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>4th</td>
<td>✓</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Introduction

• Similar cases are popular among our lives, such as

<table>
<thead>
<tr>
<th>Movie Classification</th>
<th>Car Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production Company Set</td>
<td>Production Company Set</td>
</tr>
<tr>
<td>20th Century Fox</td>
<td>Audi</td>
</tr>
<tr>
<td>Warner Bros. Pictures</td>
<td>BMW</td>
</tr>
<tr>
<td>Columbia Pictures</td>
<td>Mercedes-Benz</td>
</tr>
<tr>
<td>Paramount Pictures</td>
<td>Opel</td>
</tr>
<tr>
<td>Universal Pictures</td>
<td>Porsche</td>
</tr>
<tr>
<td>Walt Disney Pictures</td>
<td>Volkswagen</td>
</tr>
<tr>
<td>Genre Set</td>
<td>Type Set</td>
</tr>
<tr>
<td>Action</td>
<td>Economy</td>
</tr>
<tr>
<td>Adventure</td>
<td>Family</td>
</tr>
<tr>
<td>Comedy</td>
<td>Sedan</td>
</tr>
<tr>
<td>Horror</td>
<td>Luxury vehicle</td>
</tr>
<tr>
<td>Science Fiction</td>
<td>Sports</td>
</tr>
<tr>
<td>War</td>
<td>Commercial</td>
</tr>
</tbody>
</table>
Problem Formulation

• Definition

**Definition 1.** (Dual Set Multi-Label Learning) Given the training set $\mathcal{D}$, the task is to learn a mapping function from the input space to the output space,

$$h : \mathcal{X} \rightarrow \mathcal{Y}^a \times \mathcal{Y}^b.$$  

For an unseen instance $\mathbf{x} \in \mathcal{X}$, the mapping function $h(\cdot)$ predicts $h(\mathbf{x}) \subseteq \mathcal{Y}^a \times \mathcal{Y}^b$ as the dual labels for $\mathbf{x}$.

![Diagram of traditional multi-label learning and dual set multi-label learning](image)
Problem Formulation

• Key challenge: exploiting label relationships
  • Intra-set: the exclusive relationship within the same set
  • Inter-set: the pairwise label set relationship
Outline

• Introduction

• Potential Solutions and Deficiencies
  • Our Approach
  • Theoretical Results
  • Experiments

• Conclusion
Potential Solutions

• Independent Decomposition
  • Decomposing the original problem into two classification problems

• Co-occurrence Based Decomposition
  • Decomposing the original problem into a new multi-class problem

• Label Stacking
  • Transforming the original problem into sequential problems
Potential Solutions

- Independent Decomposition
  - Decomposing the original problem into two classification problems

**Deficiency:**
Inter-set relationship is neglected.
Potential Solutions

• Co-occurrence Based Decomposition
  • Decomposing the original problem into a new multi-class problem

• How do we assign new labels by label co-occurrence?
Potential Solutions

• Co-occurrence Based Decomposition
  • An example showing how to assign new labels

<table>
<thead>
<tr>
<th>Instance</th>
<th>A-1</th>
<th>A-2</th>
<th>A-3</th>
<th>A-4</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>5&lt;sup&gt;th&lt;/sup&gt;</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td>7&lt;sup&gt;th&lt;/sup&gt;</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td>8&lt;sup&gt;th&lt;/sup&gt;</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

New multi-class label

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Deficiency:
It is unable to handle new label co-occurrence cases.
Potential Solutions

• Label Stacking
  • Transforming the original problem into two sequential problems

- Label Stacking

- Transforming the original problem into two sequential problems

- How do we train classifier A and B?
Potential Solutions

- Label Stacking
  - An example showing how to train classifier A and B

**Deficiency:**
Only one label set helps the other one.
Outline

• Introduction

• Potential Solutions and Deficiencies

• Our Approach

• Theoretical Results

• Experiments

• Conclusion
Our Approach

• Key Problem
  • How to find a better way to exploit intra-set and inter-set label relationship simultaneously?

• Key ideas
  • Multi-class classifiers are used to exploit intra-set label relationship.
  • Model-reuse mechanism and distribution adjusting mechanism are used to make label sets help each other, all of which exploit inter-set label relationship.

• Boosting framework is used to carry out these ideas.
Our Approach

- The DSML algorithm
- How does it work?

**Algorithm 1** The DSML algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initialize sample weights</td>
</tr>
<tr>
<td>3-4</td>
<td>Resample</td>
</tr>
<tr>
<td>5</td>
<td>Train base learners with <strong>model-reuse mechanism</strong></td>
</tr>
<tr>
<td>6</td>
<td>Calculate error rates</td>
</tr>
<tr>
<td>7-9</td>
<td>Check error rates</td>
</tr>
<tr>
<td>10-12</td>
<td>Update sample weights with <strong>distribution adjusting mechanism</strong></td>
</tr>
</tbody>
</table>

### Step 1
Initialize sample weights:
- \( w_{1,i}^a = w_{1,i}^b = 1/m \)

### Step 2
For \( t = 1 \) to \( T \) do:
1. Sample the training set \( \mathcal{D} = \{(x_i, y_i^a, y_i^b)| 1 \leq i \leq m\} \), base learning algorithm \( A \), number of rounds \( T \), weight tuning parameter \( B \).
2. Sample the training set \( \mathcal{D} = \{(x_i, y_i^a, y_i^b)| 1 \leq i \leq m\} \), base learning algorithm \( A \), number of rounds \( T \), weight tuning parameter \( B \).
3. Training three models \( h_t^{raw} \), \( h_t^a \) and \( h_t^b \) with model-reuse mechanism by Eq. (1), (2) and (3).
4. Calculating error rate \( e_t^a \) and \( e_t^b \) by Eq. (4) and (5).
5. If \( e_t^a > \frac{(L_1 - 1)}{L_1} \) or \( e_t^b > \frac{(L_2 - 1)}{L_2} \) then:
   - Break
6. Updating model weight \( \alpha_t^a \) and \( \alpha_t^b \) by Eq. (6) and (7).
7. Updating sample distribution \( w_{t+1}^a \) and \( w_{t+1}^b \) by \( \alpha_t^a \), \( \alpha_t^b \) and \( B \) with distribution adjusting mechanism according to Eq. (8) and (9).
8. Performing normalization to \( w_{t+1}^a \) and \( w_{t+1}^b \).
Our Approach

- The DSML algorithm
  - Training base learners with **model-reuse mechanism**

**Algorithm 1 The DSML algorithm**

| Input: | Training set $D = \{(x_i, y_i^a, y_i^b) | 1 \leq i \leq m\}$, base learning algorithm $A$, number of rounds $T$, weight tuning parameter $B$ |
|--------|-----------------------------------------------------------------|
| Training process: | 1. $w_{1,i}^{a} = w_{1,i}^{b} = 1/m$;  |
| | 2. for $t = 1$ to $T$ do  |
| | 3. $(X_t^a, y_t^a) \leftarrow Sample(D, w_t^a)$  |
| | 4. $(X_t^b, y_t^b) \leftarrow Sample(D, w_t^b)$  |
| | 5. Training three models $h_t^{a,w}, h_t^{a}, h_t^{b}$ with model-reuse mechanism by Eq. (1), (2) and (3)  |
| | 6. Calculating error rate $e_t^a$ and $e_t^b$ by Eq. (4) and (5)  |
| | 7. if $e_t^a > (L_1 - 1)/L_1$ or $e_t^b > (L_2 - 1)/L_2$ then  |
| | 8. Break  |
| | 9. end if  |
| | 10. Updating model weight $\alpha_t^a$ and $\alpha_t^b$ by Eq. (6) and (7)  |
| | 11. Updating sample distribution $w_{t+1}^a$ and $w_{t+1}^b$ by $\alpha_t^a$, $\alpha_t^b$ and $B$ with distribution adjusting mechanism according to Eq. (8) and (9)  |
| | 12. Performing normalization to $w_{t+1}^a$ and $w_{t+1}^b$  |
| Output: | Predict labels for dual set: $f^a(x)$ and $f^b(x)$ by Eq. (10) and (11) |
Our Approach

• The DSML algorithm
  • Calculating error rate and updating model weight

**Algorithm 1** The DSML algorithm

**Input:** Training set \( \mathcal{D} = \{(x_i, y^a_i, y^b_i) | 1 \leq i \leq m\} \), base learning algorithm \( \mathcal{A} \), number of rounds \( T \), weight tuning parameter \( B \)

**Training process:**
1. \( w^a_{1,t} = w^b_{1,t} = 1/m; \)
2. for \( t = 1 \) to \( T \) do
3. \( (X^a_t, y^a_t) \leftarrow \text{Sample}(\mathcal{D}, w^a_t) \)
4. \( (X^b_t, y^b_t) \leftarrow \text{Sample}(\mathcal{D}, w^b_t) \)
5. Training three models \( h^{aw}_t \), \( h^a_t \) and \( h^b_t \) with model-reuse mechanism by Eq. (1), (2) and (3)
6. Calculating error rate \( \epsilon^a_t \) and \( \epsilon^b_t \) by Eq. (4) and (5)
7. if \( \epsilon^a_t > (L_1 - 1)/L_1 \) or \( \epsilon^b_t > (L_2 - 1)/L_2 \) then
8. Break
9. end if
10. Updating model weight \( \alpha^a_t \) and \( \alpha^b_t \) by Eq. (6) and (7)
11. Updating sample distribution \( w^a_{t+1} \) and \( w^b_{t+1} \) by \( \alpha^a_t \), \( \alpha^b_t \) and \( B \) with distribution adjusting mechanism according to Eq. (8) and (9)
12. Performing normalization to \( w^a_{t+1} \) and \( w^b_{t+1} \)
13. end for

**Output:** Predict labels for dual set: \( \hat{f}^a(x) \) and \( \hat{f}^b(x) \) by Eq. (10) and (11)

\[
\epsilon^a_t = \sum_{i=1}^{m} \left[ h^a_t([X^a_s, \hat{Y}^b_i]_i) \neq (y^a_s)_i \right]
\]
\[
\epsilon^b_t = \sum_{i=1}^{m} \left[ h^b_t([X^b_s, \hat{Y}^a_i]_i) \neq (y^b_s)_i \right]
\]
\[
\alpha^a_t = \frac{1}{L_1} \left[ \log \frac{1 - \epsilon^a_t}{\epsilon^a_t} + \log(L_1 - 1) \right]
\]
\[
\alpha^b_t = \frac{1}{L_2} \left[ \log \frac{1 - \epsilon^b_t}{\epsilon^b_t} + \log(L_2 - 1) \right]
\]
Our Approach

• The DSML algorithm
  • Updating sample weight with distribution adjusting mechanism

```
Algorithm 1 The DSML algorithm

Input: Training set \( \mathcal{D} = \{(x_i, y_i^a, y_i^b) | 1 \leq i \leq m\} \), base learning algorithm \( \mathcal{A} \), number of rounds \( T \), weight tuning parameter \( B \)

Training process:
1: \( w_{1,i}^a = w_{1,i}^b = 1/m; \)
2: for \( t = 1 \) to \( T \) do
3: \( (X_i^a, y_i^a) \leftarrow \text{Sample}(\mathcal{D}, w_t^a) \)
4: \( (X_i^b, y_i^b) \leftarrow \text{Sample}(\mathcal{D}, w_t^b) \)
5: Training three models \( h_{t}^{raw}, h_{t}^{a} \) and \( h_{t}^{b} \) with model-reuse mechanism by Eq. (1), (2) and (3)
6: Calculating error rate \( e_t^a \) and \( e_t^b \) by Eq. (4) and (5)
7: if \( e_t^a > (L_1 - 1)/L_1 \) or \( e_t^b > (L_2 - 1)/L_2 \) then
8: Break
9: end if
10: Updating model weight \( \alpha_t^a \) and \( \alpha_t^b \) by Eq. (6) and (7)
11: Updating sample distribution \( w_{t+1}^a \) and \( w_{t+1}^b \) by \( \alpha_t^a, \alpha_t^b \) and \( B \) with distribution adjusting mechanism according to Eq. (8) and (9)
12: Performing normalization to \( w_{t+1}^a \) and \( w_{t+1}^b \)
13: end for

Output: Predict labels for dual set: \( f^a(x) \) and \( f^b(x) \) by Eq. (10) and (11)
```

\[ w_{t+1,i}^a = w_{t,i}^a \exp(\alpha_t^a \cdot [y_i^a \neq \hat{y}_i^a]) \cdot B[y_i^b \neq \hat{y}_i^b] \]

\[ w_{t+1,i}^b = w_{t,i}^b \exp(\alpha_t^b \cdot [y_i^b \neq \hat{y}_i^b]) \cdot B[y_i^a \neq \hat{y}_i^a] \]
Outline

• Introduction

• Potential Solutions and Deficiencies

• Our Approach

• Theoretical Results

• Experiments

• Conclusion
Theoretical Results

• Superiority of learning by splitting the label set

**Theorem 1.** For dual-set multi-label learning problems, $h^a$ and $h^b$ are classifiers trained on the instance space $\mathcal{X}$ and label space $\mathcal{Y}^a$, $\mathcal{Y}^b$ respectively. $h$ is a classifier trained directly from $\mathcal{X} \times [\mathcal{Y}^a \times \mathcal{Y}^b]$, namely,

$$h : \mathbf{x} \rightarrow \arg \max_{y^a, y^b \in [\mathcal{Y}^a \times \mathcal{Y}^b]} h(\mathbf{x}, y),$$

where $y = [y^a, y^b]$, then margin of learning from dual label set is larger than that of directly learning from all labels:

$$\min\{\bar{\rho}_{h^a}(\mathbf{x}, y^a), \bar{\rho}_{h^b}(\mathbf{x}, y^b)\} \geq \bar{\rho}_h(\mathbf{x}, y),$$

where

- $\bar{\rho}_{h^a}(\mathbf{x}, y^a)$: margin of multi-class learning
- $\bar{\rho}_{h^b}(\mathbf{x}, y^b)$: margin of learning directly from all labels

**Remark:**

It shows the effectiveness of learning by splitting the label set into two disjoint label sets, which implies that we should explicitly considering the dual label sets.
Theoretical Results

- Generalization bound of learning by splitting the label set

**Theorem 2.** Let $H = \{(x, y^a, y^b) \in \mathcal{X} \times [\mathcal{Y}^a \times \mathcal{Y}^b] \rightarrow w^T \phi(x) | \sum_{\ell=1}^{L_1+L_2} \|w\|_{\mathcal{H}}^2 \leq \Lambda^2 \}$ be a hypothesis set with $y^a = 1, \ldots, L_1, y^b = 1, \ldots, L_2$, where $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is a feature mapping induced by some positive definite kernel $\kappa$. Assume that $S \subset \{x : \kappa(x, x) \leq \tau^2 \}$, and fix $\rho > 0$, then for any $\delta > 0$, with probability at least $1 - \delta$, the following generalization bound holds for all $h^{spl} = [h^a, h^b] \in H$:

$$R(h^{spl}) \leq \hat{R}_\rho(h^{spl}) + \frac{2r \Lambda}{\rho} \sqrt{\frac{\max\{L_1, L_2\}}{m}} + 3 \sqrt{\frac{\log(2/\delta)}{m}} + O(1/\sqrt{m})$$

**Remark:**
The convergence rate of the generalization error is standard as $O(1/\sqrt{m})$. And the error bound exhibits a radical dependence on the maximal number of labels in dual label sets.
Outline

• Introduction
• Potential Solutions and Deficiencies
• Our Approach
• Theoretical Results
• Experiments
• Conclusion
Experiments

• Datasets
  • We collected or adapted three real-world dataset. Now take Calligrapher-Font dataset for example
    • We collected 23195 calligraphic images
    • We transformed each of them into 512-dimensional feature vector
    • There are 14 calligraphers and 5 kinds of fonts

• Statistics of three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of instances</th>
<th>No. of dimensions</th>
<th>Size of label set A</th>
<th>Size of label set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calligrapher-Font</td>
<td>23195</td>
<td>512</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Brand-Type</td>
<td>2247</td>
<td>4096</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Frequency-Gender</td>
<td>3157</td>
<td>19</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Experiments

• Evaluation Measures
  • Accuracy of the label set A
  • Accuracy of the label set B
  • Overall accuracy

Definition 4. Let $\mathcal{Z} = \{z_i, y_i^a, y_i^b|1 \leq i \leq n\}$ denote the testing set where $n$ is the total number of testing instances and let $h^a, h^b$ be the underlying classifiers learned from the training process associated with two label sets respectively. Three accuracies are defined to evaluate the performance,

\[
\text{Accuracy}_a = \frac{1}{n} \sum_{i=1}^{n} [h^a(z_i) = y_i^a],
\]

\[
\text{Accuracy}_b = \frac{1}{n} \sum_{i=1}^{n} [h^b(z_i) = y_i^b],
\]

\[
\text{Accuracy}_{all} = \frac{1}{n} \sum_{i=1}^{n} [h^a(z_i) = y_i^a] \cdot [h^b(z_i) = y_i^b].
\]
Experiments

- Comparing DSML with other algorithms
  - Multi-class RBF neural networks are used as base learner for DSML and potential solutions.
  - The outputs of classical multi-label learning approaches are modified to fit dual set multi-label learning.
  - 5-fold cross-validation performance of these algorithms (mean±std.)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cal.-Font</td>
<td>Accy_a</td>
<td>0.6562 ± .0059</td>
<td>.5967 ± .0082</td>
<td>N/A</td>
<td>0.6019 ± .0088</td>
<td>0.6337 ± .0075</td>
<td>0.6372 ± .0045</td>
<td>0.1493 ± .0051</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Accy_b</td>
<td>0.7223 ± .0079</td>
<td>.6751 ± .0040</td>
<td>N/A</td>
<td>0.6801 ± .0078</td>
<td>0.7101 ± .0030</td>
<td>0.7100 ± .0087</td>
<td>0.4104 ± .0067</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Accy_all</td>
<td>0.5672 ± .0087</td>
<td>.4836 ± .0099</td>
<td>0.5609 ± .0050</td>
<td>0.4889 ± .0094</td>
<td>0.5570 ± .0048</td>
<td>0.5396 ± .0066</td>
<td>0.0764 ± .0077</td>
<td>N/A</td>
</tr>
<tr>
<td>Brand-Type</td>
<td>Accy_a</td>
<td>0.5723 ± .0226</td>
<td>.5661 ± .0129</td>
<td>N/A</td>
<td>0.5968 ± .0254</td>
<td>0.4722 ± .0160</td>
<td>0.5207 ± .0223</td>
<td>0.1206 ± .0182</td>
<td>.5238 ± .0352</td>
</tr>
<tr>
<td></td>
<td>Accy_b</td>
<td>0.7730 ± .0249</td>
<td>.7677 ± .0092</td>
<td>N/A</td>
<td>0.7637 ± .0225</td>
<td>0.7245 ± .0115</td>
<td>0.7405 ± .0126</td>
<td>0.3000 ± .0509</td>
<td>.7517 ± .0137</td>
</tr>
<tr>
<td></td>
<td>Accy_all</td>
<td>0.4949 ± .0227</td>
<td>.4744 ± .0105</td>
<td>0.4784 ± .0294</td>
<td>0.4735 ± .0302</td>
<td>0.3912 ± .0078</td>
<td>0.4201 ± .0160</td>
<td>0.0538 ± .0053</td>
<td>.4183 ± .0345</td>
</tr>
<tr>
<td>Freq.-Gndr.</td>
<td>Accy_a</td>
<td>0.8521 ± .0091</td>
<td>.8321 ± .0212</td>
<td>N/A</td>
<td>0.8375 ± .0170</td>
<td>0.5879 ± .0091</td>
<td>0.7570 ± .0144</td>
<td>0.4004 ± .1464</td>
<td>.0326 ± .0135</td>
</tr>
<tr>
<td></td>
<td>Accy_b</td>
<td>0.9547 ± .0061</td>
<td>.9579 ± .0067</td>
<td>N/A</td>
<td>0.9550 ± .0051</td>
<td>0.6953 ± .0196</td>
<td>0.9661 ± .0047</td>
<td>0.5014 ± .0271</td>
<td>0.5382 ± .0643</td>
</tr>
<tr>
<td></td>
<td>Accy_all</td>
<td>0.8220 ± .0082</td>
<td>.8039 ± .0214</td>
<td>0.8068 ± .0187</td>
<td>0.8096 ± .0183</td>
<td>0.4587 ± .0161</td>
<td>0.7387 ± .0134</td>
<td>0.1704 ± .0847</td>
<td>0.0127 ± .0116</td>
</tr>
</tbody>
</table>

- DSML is better than others
Experiments

• Study on **model-reuse mechanism**

![Diagram showing steps of model-reuse mechanism]

• 5-fold cross-validation performance of DSML on the *Brand-Type* dataset (mean)
  • Boosting round increases to 10
  • Distribution adjusting parameter is set to 1.00 and 1.10
Experiments

- Study on model-reuse mechanism
  - 5-fold cross-validation Performance of DSML on the Brand-Type dataset (mean)
    - Boosting round increases to 10
    - Distribution adjusting parameter is set to 1.00 and 1.10

- It validates the effectiveness of model-reuse mechanism
  - Similar phenomena can be observed in other datasets
Experiments

• Study on distribution adjusting mechanism
  • $B$ is the distribution adjusting parameter
    \[
    w_{t+1,i}^a = w_{t,i}^a \exp(\alpha_t^a \cdot [y_i^a \neq \hat{y}_i^a]) B[y_i^a \neq \hat{y}_i^a]
    \]
    \[
    w_{t+1,i}^b = w_{t,i}^b \exp(\alpha_t^b \cdot [y_i^b \neq \hat{y}_i^b]) B[y_i^b \neq \hat{y}_i^b]
    \]
  • When $B = 1.00$, algorithms perform without distribution adjusting mechanism
  • 5-fold cross-validation performance of DSML algorithm (mean±std.)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Measure</th>
<th>1.00</th>
<th>1.01</th>
<th>1.02</th>
<th>1.03</th>
<th>1.05</th>
<th>1.10</th>
<th>1.15</th>
<th>1.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cal.-Font</td>
<td>Accy.a</td>
<td>.6536 ± .0054</td>
<td>.6576 ± .0064</td>
<td>.6567 ± .0051</td>
<td>.6557 ± .0067</td>
<td>.6562 ± .0059</td>
<td>.6541 ± .0033</td>
<td>.6546 ± .0076</td>
<td>.6528 ± .0060</td>
</tr>
<tr>
<td></td>
<td>Accy.b</td>
<td>.7225 ± .0060</td>
<td>.7244 ± .0062</td>
<td>.7249 ± .0043</td>
<td>.7263 ± .0046</td>
<td>.7223 ± .0079</td>
<td>.7246 ± .0041</td>
<td>.7210 ± .0037</td>
<td>.7230 ± .0054</td>
</tr>
<tr>
<td></td>
<td>Accy.all</td>
<td>.5656 ± .0078</td>
<td>.5697 ± .0062</td>
<td>.5674 ± .0043</td>
<td>.5690 ± .0058</td>
<td>.5672 ± .0087</td>
<td>.5698 ± .0043</td>
<td>.5659 ± .0078</td>
<td>.5660 ± .0045</td>
</tr>
<tr>
<td>Brand-Type</td>
<td>Accy.a</td>
<td>.5710 ± .0296</td>
<td>.5657 ± .0259</td>
<td>.5706 ± .0303</td>
<td>.5706 ± .0206</td>
<td>.5723 ± .0226</td>
<td>.5710 ± .0185</td>
<td>.5710 ± .0201</td>
<td>.5603 ± .0343</td>
</tr>
<tr>
<td></td>
<td>Accy.b</td>
<td>.7784 ± .0142</td>
<td>.7668 ± .0185</td>
<td>.7659 ± .0193</td>
<td>.7650 ± .0212</td>
<td>.7730 ± .0249</td>
<td>.7641 ± .0107</td>
<td>.7788 ± .0182</td>
<td>.7699 ± .0182</td>
</tr>
<tr>
<td></td>
<td>Accy.all</td>
<td>.4905 ± .0324</td>
<td>.4847 ± .0227</td>
<td>.4856 ± .0257</td>
<td>.4882 ± .0231</td>
<td>.4949 ± .0227</td>
<td>.4824 ± .0073</td>
<td>.4922 ± .0228</td>
<td>.4833 ± .0340</td>
</tr>
<tr>
<td>Freq.-Gndr.</td>
<td>Accy.a</td>
<td>.8413 ± .0110</td>
<td>.8432 ± .0107</td>
<td>.8432 ± .0177</td>
<td>.8413 ± .0140</td>
<td>.8521 ± .0091</td>
<td>.8435 ± .0137</td>
<td>.8473 ± .0162</td>
<td>.8476 ± .0119</td>
</tr>
<tr>
<td></td>
<td>Accy.b</td>
<td>.9541 ± .0071</td>
<td>.9531 ± .0041</td>
<td>.9512 ± .0073</td>
<td>.9554 ± .0074</td>
<td>.9547 ± .0061</td>
<td>.9515 ± .0040</td>
<td>.9557 ± .0054</td>
<td>.9560 ± .0038</td>
</tr>
<tr>
<td></td>
<td>Accy.all</td>
<td>.8131 ± .0060</td>
<td>.8134 ± .0118</td>
<td>.8134 ± .0158</td>
<td>.8119 ± .0166</td>
<td>.8220 ± .0082</td>
<td>.8128 ± .0151</td>
<td>.8172 ± .0153</td>
<td>.8175 ± .0155</td>
</tr>
</tbody>
</table>

• Best result appears when $B = 1.05$
Outline

• Introduction
• Potential Solutions and Deficiencies
• Our Approach
• Theoretical Results
• Experiments
• Conclusion
Conclusion

• Dual Set Multi-Label Learning is proposed as a novel learning framework.

• A boosting-like DSML approach is designed to address this kind of problem which outperforms other compared algorithms.

• Theoretical and empirical analyses are presented to show it is better to learn with dual label sets than to learn directly from all labels.
Thank you for listening.

Q & A