Dual Set Multi-Label Learning

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

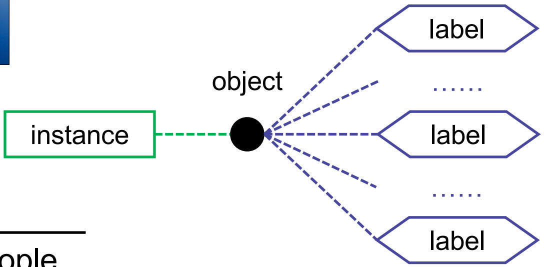
[1] LAMDA Group, Nanjing University, China
[2] Nanjing University of Aeronautics and Astronautics, China

Outline

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

Introduction

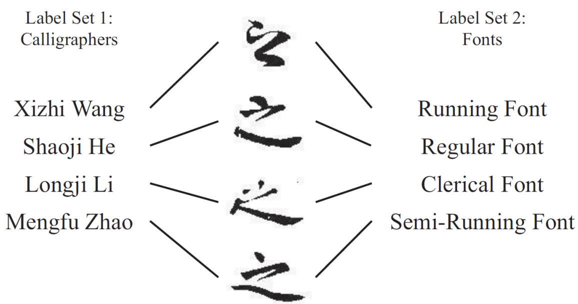
- An example of traditional multi-label learning

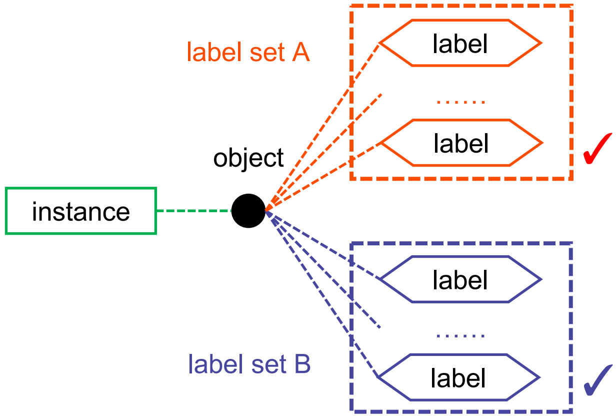


Sky	Lake	Road	Mountain	Bird	People
√	√	×	√	×	×

Introduction

- An example different from traditional multi-label learning





calligrapher set

font set

Chinese Character	Wang	He	Li	Zhao	Semi-running	Running	Clerical	Regular
1st	√	×	×	×	×	√	×	×
2nd	×	√	×	×	×	×	×	√
3rd	×	×	√	×	×	×	√	×
4th	×	×	×	√	√	×	×	×

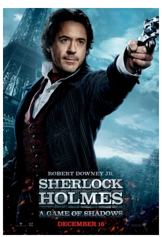
Introduction

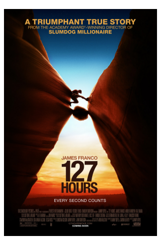
LAMDA
Learning And Mining from DatA

- Similar cases are popular among our lives, such as

movie classification

Production Company Set	Genre Set
20th Century Fox	Action
Warner Bros. Pictures	Adventure
Columbia Pictures	Comedy
Paramount Pictures	Horror
Universal Pictures	Science Fiction
Walt Disney Pictures	War





car classification

Production Company Set	Type Set
Audi	Economy
BMW	Family
Mercedes-Benz	Sedan
Opel	Luxury vehicle
Porsche	Sports
Volkswagen	Commercial

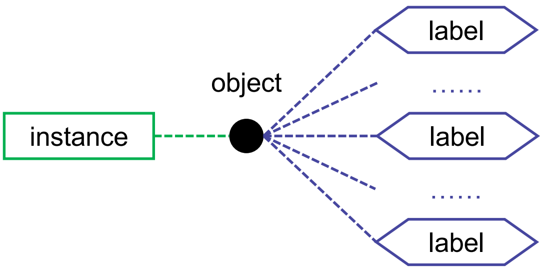
Problem Formulation

- Definition

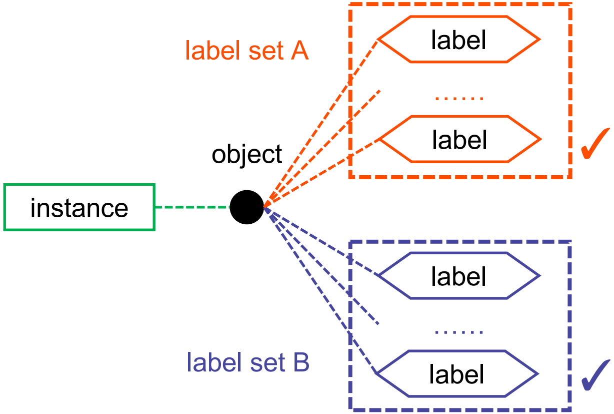
Definition 1. *(Dual Set Multi-Label Learning) Given the training set $\mathcal{D}$, the task is to learn a mapping function from the input space to the output space,*

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

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



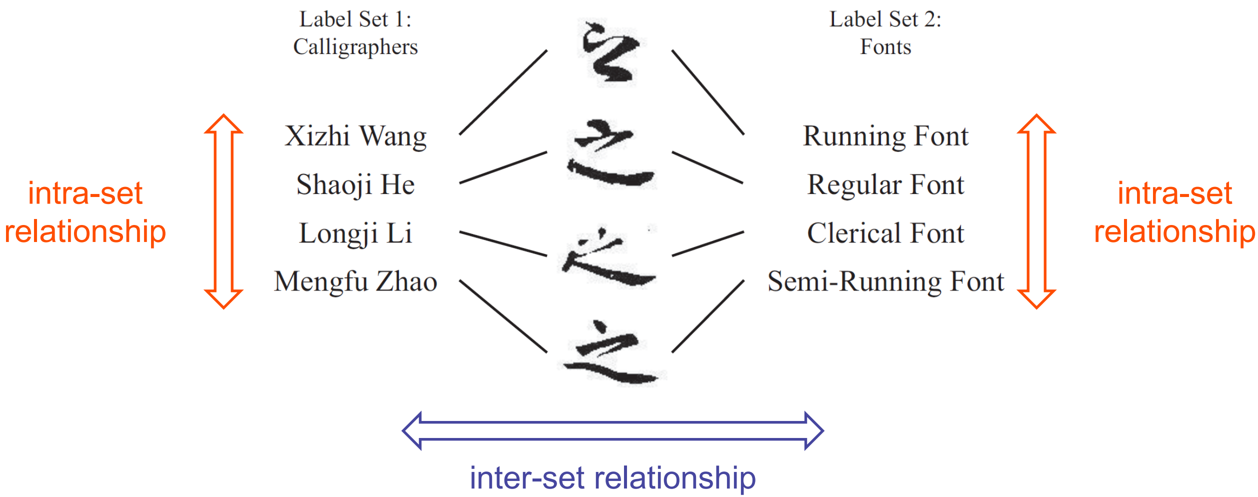
traditional multi-label learning



dual set multi-label learning

Problem Formulation

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



Outline

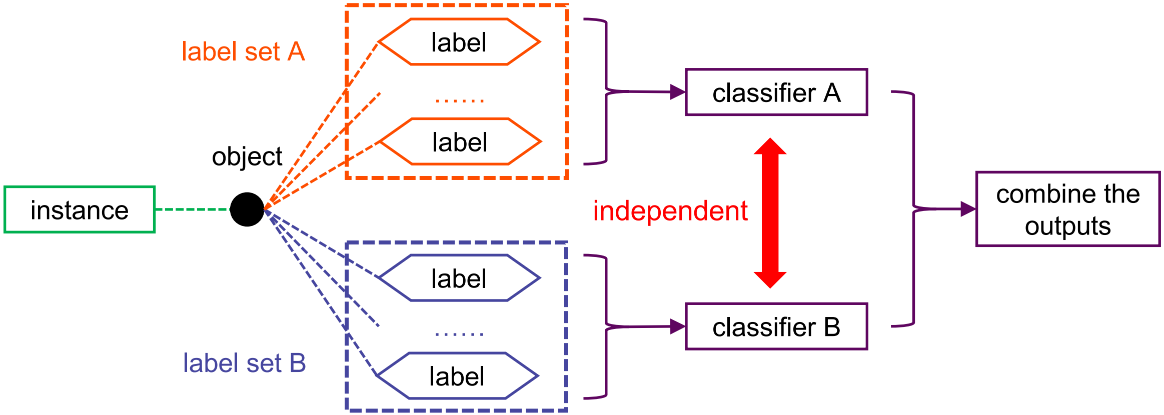
- Introduction
- Potential Solutions and Deficiencies
- Our Approach
- Theoretical Results
- Experiments
- Conclusion

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

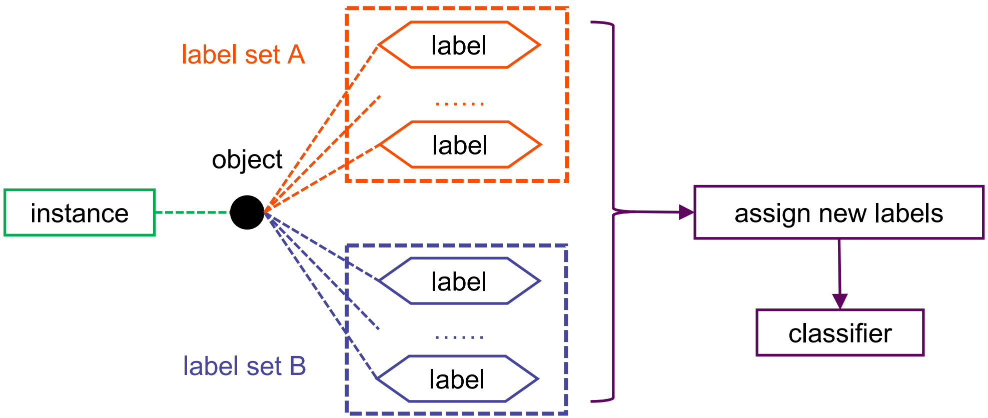


Deficiency:
Inter-set relationship is neglected.

Potential Solutions

Learning And Mining from DatA

- 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

Instance	A-1	A-2	A-3	A-4	B-1	B-2	B-3	B-4
	label set A				label set B			
1st	√	×	×	×	×	√	×	×
2nd	×	√	×	×	×	×	×	√
3rd	×	×	√	×	×	×	√	×
4th	×	×	×	√	√	×	×	×
5th	×	×	√	×	×	×	√	×
6th	×	√	×	×	×	×	×	√
7th	×	×	√	×	×	×	√	×
8th	√	×	×	×	×	√	×	×

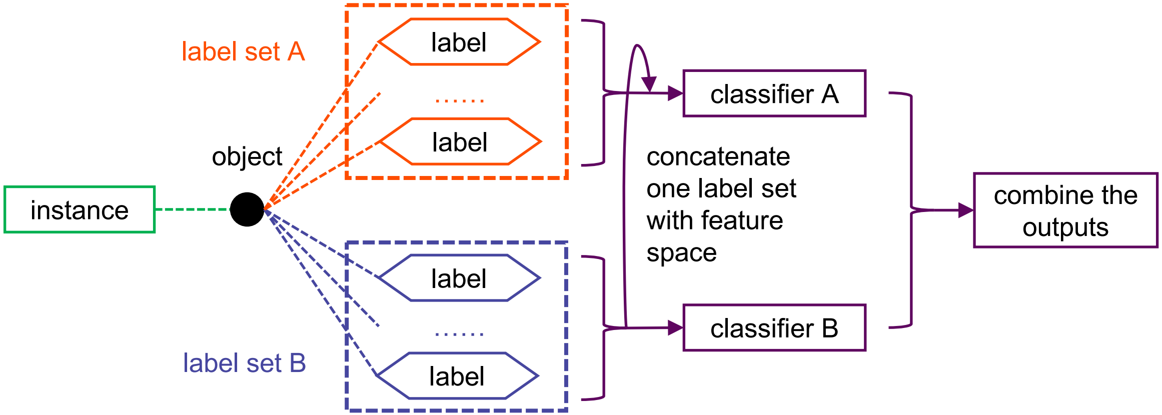
New multi-class label
1
2
3
4
3
2
3
1

Deficiency:

It is unable to handle new label co-occurrence cases.

Potential Solutions

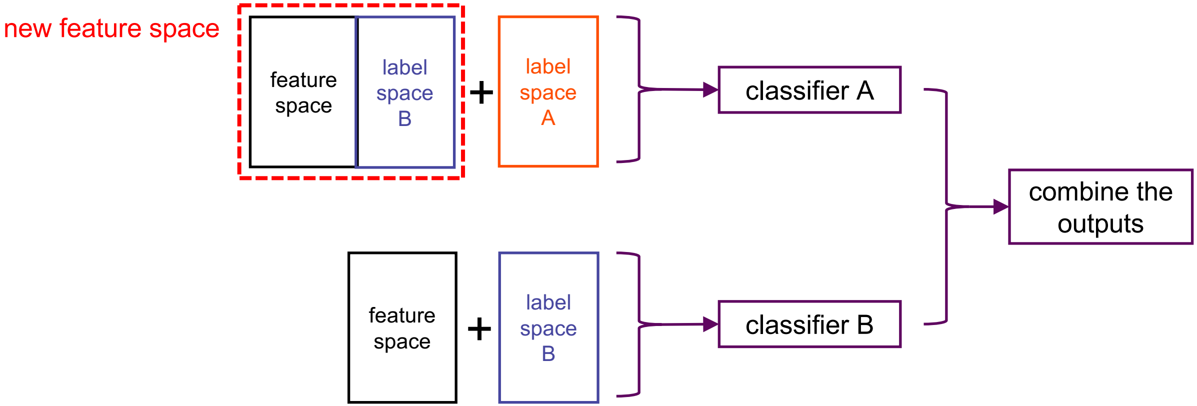
- Label Stacking
  - Transforming the original problem into two sequential problems



- *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

Input: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i^a, y_i^b) | 1 \leq i \leq m\}$, 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: $\quad (X_s^a, y_s^a) \leftarrow Sample(\mathcal{D}, w_t^a)$
4: $\quad (X_s^b, y_s^b) \leftarrow Sample(\mathcal{D}, w_t^b)$
5: $\quad$ Training three models h_t^{raw}, h_t^a and h_t^b with model-reuse mechanism by Eq. (1), (2) and (3)
6: $\quad$ Calculating error rate ϵ_t^a and ϵ_t^b by Eq. (4) and (5)
7: $\quad$ **if** $\epsilon_t^a > (L_1 - 1)/L_1$ or $\epsilon_t^b > (L_2 - 1)/L_2$ **then**
8: $\qquad$ Break
9: $\quad$ **end if**
10: $\quad$ Updating model weight α_t^a and α_t^b by Eq. (6) and (7)
11: $\quad$ Updating sample distribution w_{t+1}^a and w_{t+1}^b by α_t^a, α_t^b and B with distribution adjusting mechanism according to Eq. (8) and (9)
12: $\quad$ Performing normalization to w_{t+1}^a and w_{t+1}^b
13: **end for**

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

Step	Task
1	Initialize sample weights
3-4	Resample
5	Train base learners with model-reuse mechanism
6	Calculate error rates
7-9	Check error rates
10-12	Update sample weights with distribution adjusting mechanism

Our Approach

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

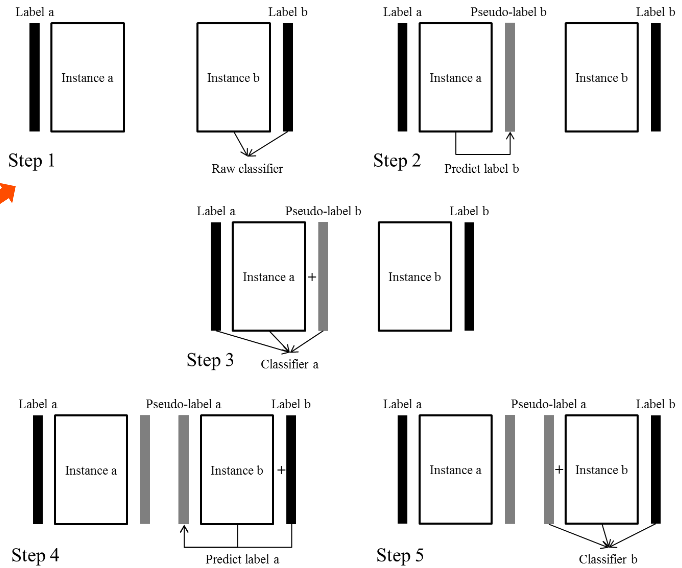
Algorithm 1 The DSML algorithm

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

Training process:

$w_{1,i}^a = w_{1,i}^b = 1/m$;
for $t = 1$ to T **do**
  $(X_s^a, y_s^a) \leftarrow Sample(\mathcal{D}, w_t^a)$
  $(X_s^b, y_s^b) \leftarrow Sample(\mathcal{D}, w_t^b)$
  Training three models h_t^{raw}, h_t^a and h_t^b with model-reuse mechanism by Eq. (1), (2) and (3)
  Calculating error rate ϵ_t^a and ϵ_t^b by Eq. (4) and (5)
  if $\epsilon_t^a > (L_1 - 1)/L_1$ or $\epsilon_t^b > (L_2 - 1)/L_2$ **then**
    Break
  end if
  Updating model weight α_t^a and α_t^b by Eq. (6) and (7)
  Updating sample distribution w_{t+1}^a and w_{t+1}^b by α_t^a, α_t^b and B with distribution adjusting mechanism according to Eq. (8) and (9)
  Performing normalization to w_{t+1}^a and w_{t+1}^b
end for

Output: Predict labels for dual set: $f^a(\boldsymbol{x})$ and $f^b(\boldsymbol{x})$ 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 $\mathcal{D} = \{(\boldsymbol{x}_i, y_i^a, y_i^b) | 1 \leq i \leq m\}$, base learning algorithm $\mathcal{A}$, number of rounds T, weight tuning parameter B

Training process:

```
w^a_{1,i} = w^b_{1,i} = 1/m;
for t = 1 to T do
    (X^a_s, y^a_s) ← Sample(D, w^a_t)
    (X^b_s, y^b_s) ← Sample(D, w^b_t)
    Training three models h^raw_t, h^a_t and h^b_t with model-reuse mechanism by Eq. (1), (2) and (3)
    Calculating error rate ε^a_t and ε^b_t by Eq. (4) and (5)
    if ε^a_t > (L_1 − 1)/L_1 or ε^b_t > (L_2 − 1)/L_2 then
        Break
    end if
    Updating model weight α^a_t and α^b_t by Eq. (6) and (7)
    Updating sample distribution w^a_{t+1} and w^b_{t+1} by α^a_t, α^b_t and B with distribution adjusting mechanism according to Eq. (8) and (9)
    Performing normalization to w^a_{t+1} and w^b_{t+1}
end for
```

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

$$\epsilon_t^a = \sum_{i=1}^{m} [\![h_t^a([X_s^a, \hat{Y}^b]_i) \neq (y_s^a)_i]\!]$$

$$\epsilon_t^b = \sum_{i=1}^{m} [\![h_t^b([X_s^b, \hat{Y}^a]_i) \neq (y_s^b)_i]\!]$$

$$\alpha_t^a = \frac{1}{L_1} \left[\log \frac{1 - \epsilon_t^a}{\epsilon_t^a} + \log(L_1 - 1) \right]$$

$$\alpha_t^b = \frac{1}{L_2} \left[\log \frac{1 - \epsilon_t^b}{\epsilon_t^b} + \log(L_2 - 1) \right]$$

Our Approach

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

Algorithm 1 The DSML algorithm

Input: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i^a, y_i^b) | 1 \leq i \leq m\}$, 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: $\quad (X_s^a, y_s^a) \leftarrow Sample(\mathcal{D}, w_t^a)$
4: $\quad (X_s^b, y_s^b) \leftarrow Sample(\mathcal{D}, w_t^b)$
5: $\quad$ Training three models h_t^{raw}, h_t^a and h_t^b with model-reuse mechanism by Eq. (1), (2) and (3)
6: $\quad$ Calculating error rate ϵ_t^a and ϵ_t^b by Eq. (4) and (5)
7: $\quad$ **if** $\epsilon_t^a > (L_1 - 1)/L_1$ or $\epsilon_t^b > (L_2 - 1)/L_2$ **then**
8: $\qquad$ Break
9: $\quad$ **end if**
10: $\quad$ Updating model weight α_t^a and α_t^b by Eq. (6) and (7)
11: $\quad$ Updating sample distribution w_{t+1}^a and w_{t+1}^b by α_t^a, α_t^b and B with distribution adjusting mechanism according to Eq. (8) and (9)
12: $\quad$ Performing normalization to w_{t+1}^a and w_{t+1}^b
13: **end for**

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

B is the distribution adjusting parameter

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

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

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,*

$$h : \boldsymbol{x} \rightarrow \underset{y^a, y^b \in [\mathcal{Y}^a \times \mathcal{Y}^b]}{\arg\max} \; h(\boldsymbol{x}, y),$$

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

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

margin of multi-class learning

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

Learning And Mining from DatA

- Generalization bound of learning by splitting the label set

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

$$\underbrace{R(h^{spl})}_{\text{generalization error}} \leq \underbrace{\hat{R}_\rho(h^{spl})}_{\text{empirical error}} + \underbrace{\frac{2r\Lambda}{\rho}\sqrt{\frac{\overbrace{\max\{L_1, L_2\}}^{O(\sqrt{L})}}{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}}_{O(1/\sqrt{m})}.$$

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

Learning And Mining from DatA

- 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

Dataset	No. of instances	No. of dimensions	Size of label set A	Size of label set B
Calligrapher-Font	23195	512	14	5
Brand-Type	2247	4096	7	3
Frequency-Gender	3157	19	5	2

Experiments

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

Definition 4. *Let $\mathcal{Z} = \{\boldsymbol{z}_i, y_i^a, y_i^b | 1 \leq i \leq n\}$ 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,*

$$Accuracy_a = \frac{1}{n}\sum_{i=1}^{n}[\![h^a(\boldsymbol{z}_i) = y_i^a]\!],$$

$$Accuracy_b = \frac{1}{n}\sum_{i=1}^{n}[\![h^b(\boldsymbol{z}_i) = y_i^b]\!],$$

$$Accuracy_{all} = \frac{1}{n}\sum_{i=1}^{n}[\![h^a(\boldsymbol{z}_i) = y_i^a]\!] \cdot [\![h^b(\boldsymbol{z}_i) = y_i^b]\!].$$

Experiments

Learning And Mining from DatA

- 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.)

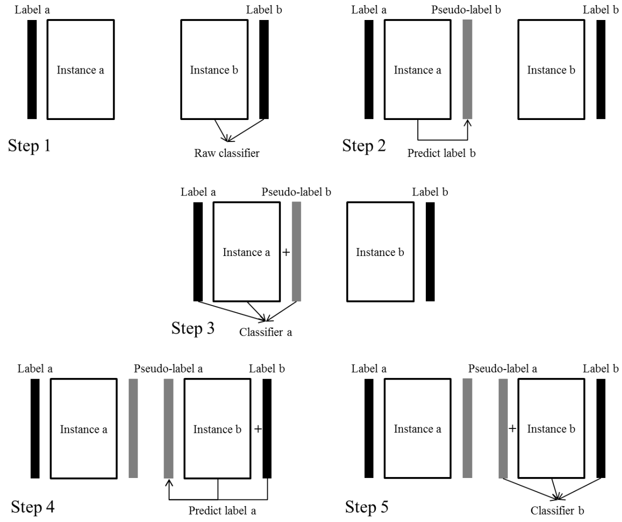
Dataset	Measure	Algorithms							
		DSML	Ind. Dec.	Co-Occ. Dec.	Label Stacking	ML-KNN	ML-RBF	BP-MLL	RankSVM
Cal.-Font	$Accy_{\cdot a}$	**.6562 ± .0059**	.5967 ± .0082	N/A	.6019 ± .0088	.6337 ± .0075	.6372 ± .0045	.1493 ± .0051	N/A
	$Accy_{\cdot b}$	**.7223 ± .0079**	.6751 ± .0040	N/A	.6801 ± .0078	.7101 ± .0030	.7100 ± .0087	.4104 ± .0670	N/A
	$Accy_{\cdot all}$	**.5672 ± .0087**	.4836 ± .0099	.5609 ± .0050	.4889 ± .0094	.5570 ± .0048	.5396 ± .0066	.0764 ± .0077	N/A
Brand-Type	$Accy_{\cdot a}$	.5723 ± .0226	.5661 ± .0129	N/A	**.5968 ± .0254**	.4722 ± .0160	.5207 ± .0223	.1206 ± .0182	.5238 ± .0352
	$Accy_{\cdot b}$	**.7730 ± .0249**	.7677 ± .0092	N/A	.7637 ± .0225	.7245 ± .0115	.7405 ± .0126	.3000 ± .0509	.7517 ± .0137
	$Accy_{\cdot all}$	**.4949 ± .0227**	.4744 ± .0105	.4784 ± .0294	.4735 ± .0302	.3912 ± .0078	.4201 ± .0160	.0538 ± .0053	.4183 ± .0345
Freq.-Gndr.	$Accy_{\cdot a}$	**.8521 ± .0091**	.8321 ± .0212	N/A	.8375 ± .0170	.5879 ± .0091	.7570 ± .0144	.4004 ± .1464	.0326 ± .0135
	$Accy_{\cdot b}$	.9547 ± .0061	.9579 ± .0067	N/A	.9550 ± .0051	.6953 ± .0196	**.9661 ± .0047**	.5014 ± .0271	.5382 ± .0643
	$Accy_{\cdot all}$	**.8220 ± .0082**	.8039 ± .0214	.8068 ± .0187	.8096 ± .0183	.4587 ± .0161	.7387 ± .0134	.1704 ± .0847	.0127 ± .0116

our approach | potential solutions | classical multi-label approaches

- DSML is better than others

Experiments

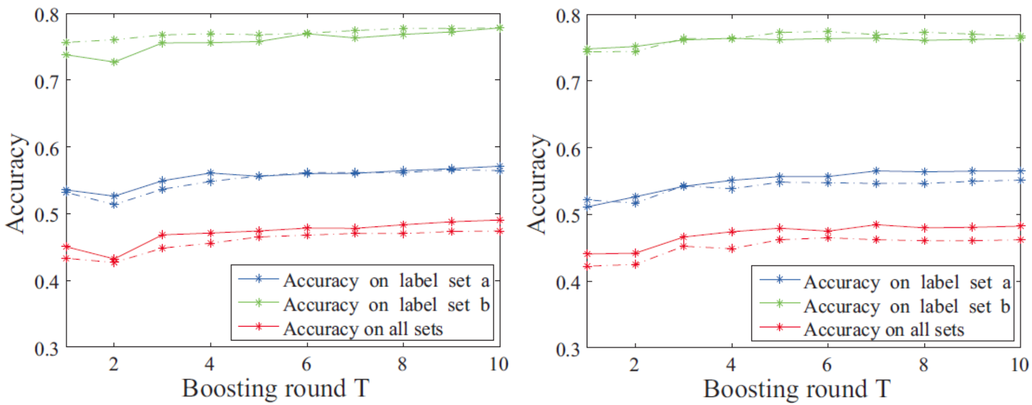
- Study on model-reuse mechanism



- 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

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



with model-reuse

without model-reuse

- 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

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

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

    - When *B* = 1.00, algorithms perform without distribution adjusting mechanism
  - 5-fold cross-validation performance of DSML algorithm (mean±std.)

Dataset	Measure	Distribution Adjusting Parameter B							
		1.00	1.01	1.02	1.03	1.05	1.10	1.15	1.20
Cal.-Font	$Accy._a$	.6536 ± .0054	**.6576 ± .0064**	.6567 ± .0051	.6557 ± .0067	.6562 ± .0059	.6541 ± .0033	.6546 ± .0076	.6528 ± .0060
	$Accy._b$	.7225 ± .0060	.7244 ± .0062	.7249 ± .0043	**.7263 ± .0046**	.7223 ± .0079	.7246 ± .0041	.7210 ± .0037	.7230 ± .0054
	$Accy._{all}$	.5656 ± .0078	.5697 ± .0062	.5674 ± .0043	.5690 ± .0058	.5672 ± .0087	**.5698 ± .0043**	.5659 ± .0078	.5660 ± .0045
Brand-Type	$Accy._a$	.5710 ± .0296	.5657 ± .0259	.5706 ± .0303	.5706 ± .0206	**.5723 ± .0226**	.5648 ± .0185	.5710 ± .0201	.5603 ± .0343
	$Accy._b$	.7784 ± .0142	.7668 ± .0185	.7659 ± .0193	.7650 ± .0212	.7730 ± .0249	.7641 ± .0107	**.7788 ± .0182**	.7699 ± .0182
	$Accy._{all}$	.4905 ± .0324	.4847 ± .0227	.4856 ± .0257	.4882 ± .0231	**.4949 ± .0227**	.4824 ± .0073	.4922 ± .0228	.4833 ± .0340
Freq.-Gndr.	$Accy._a$	.8413 ± .0110	.8432 ± .0107	.8432 ± .0177	.8413 ± .0140	**.8521 ± .0091**	.8435 ± .0137	.8473 ± .0162	.8476 ± .0119
	$Accy._b$	.9541 ± .0071	.9531 ± .0041	.9512 ± .0073	.9554 ± .0074	.9547 ± .0061	.9515 ± .0040	.9557 ± .0054	**.9560 ± .0038**
	$Accy._{all}$	.8131 ± .0060	.8134 ± .0118	.8134 ± .0158	.8119 ± .0166	**.8220 ± .0082**	.8128 ± .0151	.8172 ± .0153	.8175 ± .0155

- 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